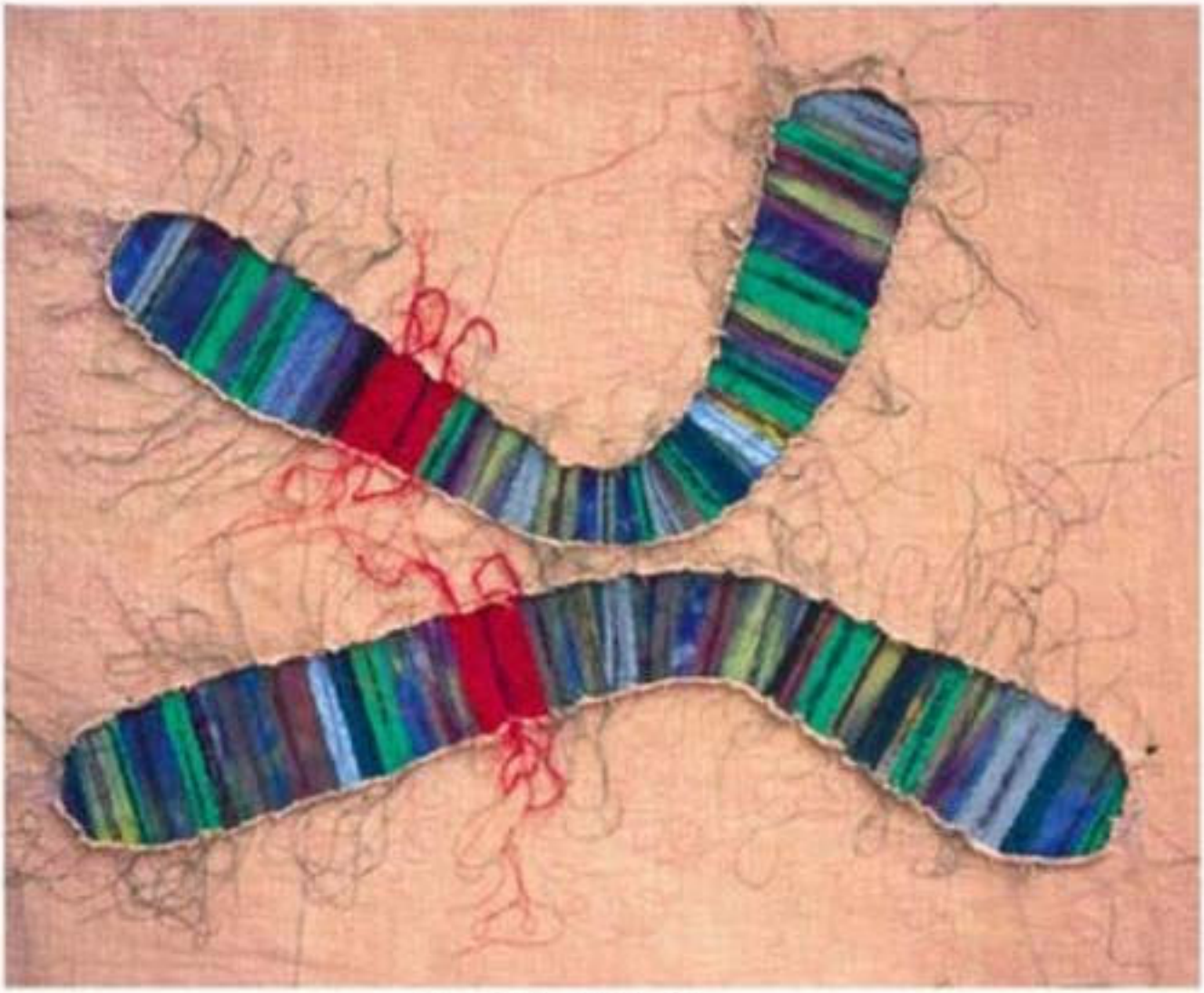
Edited by Christoph W. Sensen

WILEY-VCH

Handbook of Genome Research

Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical & Legal Issues

Volume 1



Handbook of Genome Research

Edited by
Christoph W. Sensen

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

Further Titles of Interest

T. Lengauer, R. Mannhold, H. Kubinyi, H. Timmermann (Eds.)

Bioinformatics

From Genomes to Drugs
2 Volumes

2001, ISBN 3-527-29988-2

M.J. Dunn, L.B. Jorde, P.F.R. Little, S. Subramaniam (Eds.)

Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics

8 Volume Set

2005, ISBN 0-470-84974-6

A.D. Baxevanis, B.F.F. Ouellette (Eds.)

Bioinformatics

A Practical Guide to the Analysis of Genes and Proteins

Third Edition

2005, ISBN 0-471-47878-4

H.-J. Rehm, G. Reed, A. Pühler, P. Stadler, C.W. Sensen (Eds.)

Biotechnology

Vol. 5b Genomics and Bioinformatics

2001, ISBN 0-527-28328-5

C.W. Sensen (Ed.)

Essentials of Genomics and Bioinformatics

2002, ISBN 3-527-30541-6

C. Saccone, G. Pesole

Handbook of Comparative Genomics

Principles and Methodology

2003, ISBN 0-471-39128-X

G. Kahl

The Dictionary of Gene Technology

Genomics, Transcriptomics, Proteomics

Third edition

2004, ISBN 3-527-30765-6

J.W. Dale, M. von Schantz

From Genes to Genomes

Concepts and Applications of DNA Technology

2002, ISBN 0-471-49783-5

R.D. Schmid, R. Hammelehle

Pocket Guide to Biotechnology and Genetic Engineering

2003, ISBN 3-527-30895-4

J. Licinio, M.-L. Wong (Eds.)

Pharmacogenomics

The Search for Individualized Therapies

2002, ISBN 3-527-30380-4

Handbook of Genome Research

Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues

Edited by
Christoph W. Sensen

WILEY-VCH Verlag GmbH & Co. KGaA

Edited by

Prof. Dr. Christoph W. Sensen
University of Calgary
Faculty of Medicine,
Biochemistry & Molecular Biology
3330 Hospital Drive N.W.
Calgary, Alberta T2N 4NI
Canada

Cover illustration:

Margot van Lindenberg: "Obsessed", Fabric, 2002

Fascination with the immense human diversity and immersion in four distinctly different cultures inspired artist Margot van Lindenberg to explore identity embedded in the human genome. In her art she makes reference to various aspects of genetics from microscopic images to ethical issues of bio-engineering. She develops these ideas through thread and cloth constructions, shadow projections and performance work. Margot, who currently lives in Calgary, Alberta, Canada, holds a BFA from the Alberta College of Art & Design in Calgary.

Artist Statement

Obsessed is an image of the DNA molecule, with strips of colours representing genes. The work refers to the experience of finding particular genes and the obsession that occupies those involved. It can be read either positive or negative, used to establish identity or refer to the insertion of foreign genes as in bio-engineering. The text speaks of a message, a code: a hidden knowledge as it is intentionally illegible. One can become obsessed with attempts to decipher this information.

The process of construction is part of the conceptual development of the work. Dyed and found cotton and silk were given texts, then stitched underneath ramie, which was cut away to reveal the underlying coding. The threadwork refers to the delicate structure of DNA and the raw stages of research and discovery in the field of molecular genetics.

Library of Congress Card No. applied for

British Library Cataloguing-in-Publication Data:
A catalogue record for this book is available from the British Library.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

Printed in the Federal Republic of Germany
Printed on acid-free paper

Typesetting Detzner Fotosatz, Speyer
Printing betz-druck GmbH, Darmstadt
Binding Litges & Dopf Buchbinderei GmbH, Heppenheim

ISBN-13: 978-3-527-31348-8
ISBN-10: 3-527-31348-6

Preface

Life-sciences research, especially in biology and medicine, has undergone dramatic changes in the last fifteen years. Completion of the sequencing of the first microbe genome in 1995 was followed by a flurry of activity. Today we have several hundred complete genomes to hand, including that of humans, and many more to follow. Although genome sequencing has become almost a commodity, the very optimistic initial expectations of this work, including the belief that much could be learned simply by looking at the "blueprint" of life, have largely faded into the background.

It has become evident that knowledge about the genomic organization of life forms must be complemented by understanding of gene-expression patterns and very detailed information about the protein complement of the organisms, and that it will take many years before major inroads can be made into a complete understanding of life. This has led to the development of a variety of "omics" efforts, including genomics, proteomics, metabolomics, and metabonomics. It is a typical sign of the times that about four years ago even a journal called "Omics" emerged.

An introduction to the ever-expanding technology of the subject is a major part of this book, which includes detailed description of the technology used to characterize genomic organization, gene expression patterns, protein complements, and the post-translational modification of proteins. The major model organisms and the work done to gain new insights into their biology are another central focus of the book. Several chapters are also devoted to introducing the bioinformatics tools and analytical strategies which are an integral part of any large-scale experiment.

As public awareness of relatively recent advances in life-science research increases, intense discussion has arisen on how to deal with this new research field. This discussion, which involves many groups in society, is also reflected in this book, with several chapters dedicated to the social consequences of research and development which utilizes the new approaches or the data derived from large-scale experiments. It should be clear that nobody can just ignore this topic, because it has already had direct and indirect effects on everyone's day-to-day life.

The new wave of large-scale research might be of huge benefit to humanity in the future, although in most cases we are still years away from this becoming reality. The promises and dangers of this field must be carefully weighed at each step, and this book tries to make a contribution by introducing the relevant topics that are being discussed not only by scientific experts but by Society's leaders also.

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

We would like to thank Dr Andrea Pillmann and the staff of Wiley–VCH in Weinheim, Germany, for the patience they have shown during the preparation of this book. Without their many helpful suggestions it would have been impossible to publish this book.

Christoph W. Sensen
Calgary, May 2005

Contents

Volume 1

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

List of Contributors

Lothar Altschmied
Institute of Plant Genetics and Crop Plant Research (IPK)
Corrensstr. 3
06466 Gatersleben
Germany

Mélanie Arbour
MicroArray Laboratory
National Research Council of Canada
Biotechnology Research Institute
6100 Royalmount Avenue
Montreal
Quebec, H4P 2R2
Canada

Mayi Arcellana-Panlilio
Southern Alberta Microarray Facility
University of Calgary
HM 393b
3330 Hospital Drive, N.W.
Calgary
Alberta, T2N 4N1
Canada

Gary D. Bader
Computational Biology Center
Memorial Sloan-Kettering Cancer Center
Box 460
New York, 10021
USA

François Benoit
MicroArray Laboratory
National Research Council of Canada
Biotechnology Research Institute
6100 Royalmount Avenue
Montreal
Quebec, H4P 2R2
Canada

Ernst M. Bergmann
Alberta Synchrotron Institute
University of Alberta
Edmonton
Alberta, T6G 2E1
Canada

Richard Bourgault
Department of Biological Sciences
University of Calgary
2500 University Drive N.W.
Calgary
Alberta, T2N 1N4
Canada

Detlev Buttgereit
Fachbereich Biologie
Entwicklungsbiologie
Philipps-Universität Marburg
Karl-von-Frisch-Straße 8b
35043 Marburg
Germany

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

Jian Chen
Bristol Myers Squibb Pharmaceutical Research Institute
311 Pennington-Rocky Hill Road
Pennington
New Jersey, 08534
USA

Ming Chen
Department of Bioinformatics / Medical Informatics
Faculty of Technology
University of Bielefeld
33501 Bielefeld
Germany

Antoine Danchin
Institut Pasteur
Unité de Génétique des Génomes Bactériens
Département Structure et Dynamique des Génomes
28 rue du Docteur Roux
75724 PARIS Cedex 15
France

Daniel B. Davison
Bristol Myers Squibb
Pharmaceutical Research Institute
311 Pennington-Rocky Hill Road
Pennington
New Jersey, 08534
USA

Norman J. Dovichi
Department of Chemistry
University of Washington
Seattle
Washington, 98195-1700
USA

Edna Einsiedel
University of Calgary
2500 University Drive N.W., SS318
Calgary
Alberta, T2N 1N4
Canada

Peter J. Facchini
Department of Biological Sciences
University of Calgary
2500 University Drive N.W.
Calgary
Alberta, T2N 1N4
Canada

Abul Fazal
Department of Chemistry
University of Washington
Seattle
Washington, 98195-1700
USA

Horst Feldmann
Adolf-Butenandt-Institut für Physiologische Chemie der Ludwig-Maximilians-Universität
Schillerstraße 44
80336 München
Germany

Greg Finak
Department of Biochemistry
McGill University
3775 University St
Montreal
Quebeck, H3A 2B4
Canada

Andreas Freier
Department of Bioinformatics / Medical Informatics
Faculty of Technology
University of Bielefeld
33501 Bielefeld
Germany

His Excellency Dr. Gebhard Fürst
Bischof von Rottenburg-Stuttgart
Postfach 9
72101 Rottenburg a. N.
Germany

Paul Gordon
University of Calgary
Department of Biochemistry and
Molecular Biology
3330 Hospital Drive N.W.
Calgary
Alberta, T2N 4N1
Canada

Roger C. Green
Faculty of Medicine
Memorial University of Newfoundland
St. Johns
Newfoundland, A1B3Y1
Canada

Michael Hallett
Department of Biochemistry
3775 University St
McGill University
Montreal, H3A 2B4
Canada

Ralf Hofestädt
Department of Bioinformatics /
Medical Informatics
Faculty of Technology
University of Bielefeld
33501, Bielefeld
Germany

Christopher W.V. Hogue
Dept. Biochemistry
University of Toronto and the
Samuel Lunenfeld Research Institute
Mt. Sinai Hospital
600 University Avenue
Toronto, ON M5G 1X5
Canada

Hervé Hogues
MicroArray Laboratory
National Research Council of Canada
Biotechnology Research Institute
6100 Royalmount Avenue
Montreal
Quebec, H4P 2R2
Canada

Patrick G. Humphrey
LI-COR Inc.
4308 Progressive Ave.
P.O. Box 4000
Lincoln
Nebraska, 68504
USA

Gregor Jansen
Department of Biochemistry
McGill University
3655 Promenade Sir William Osler
Montreal
Quebec, H3G 1Y6
Canada

Doris Jording
Fakulät für Biologie
Lehrstuhl für Genetik
Universität Bielefeld
33594 Bielefeld
Germany

Jörn Kalinowski
Fakulät für Biologie
Lehrstuhl für Genetik
Universität Bielefeld
33594 Bielefeld
Germany

Hans-Peter Klenk
e.gene Biotechnologie GmbH
Pöckinger Fußweg 7a
82340 Feldafing
Germany

Bartha Maria Knoppers
University of Montreal
3101, Chemin de la Tour
Montreal
Quebeck, H3C 3J7
Canada

James Kraly
Department of Chemistry
University of Washington
Seattle
Washington, 98195-1700
USA

Manfred Kröger
Institut für Mikro- und Molekularbiologie
Justus-Liebig-Universität
Heinrich-Buff-Ring 26-32
35392 Giessen
Germany

Sheena Lambert
Department of Biochemistry and
Molecular Biology
University of Calgary
3330 Hospital Drive N.W.
Calgary
Alberta, T2N 4N1
Canada

David Michels
Department of Chemistry
University of Washington
Seattle
Washington, 98195-1700
USA

Lyle R. Middendorf
LI-COR Inc.
4308 Progressive Ave.
P.O. Box 4000
Lincoln
Nebraska, 68504
USA

Narasimhachari Narayanan
VisEn Medical, Inc.
12B Cabot Road
Woburn
Massachusetts, 01801
USA

Morag Park
Department of Biochemistry
McGill University
3775 University St.
Montreal
Quebec, H3A 2B4
Canada

François Pepin
Department of Biochemistry
McGill University
3775 University St.
Montreal
Quebec, H3A 2B4
Canada

Stephanie Pollock
Department of Biochemistry
McGill University
3655 Promenade Sir William Osler
Montreal
Quebec, H3G 1Y6
Canada

Alfred Pühler
Fakulät für Biologie
Lehrstuhl für Genetik
Universität Bielefeld
33594 Bielefeld
Germany

Renate Renkawitz-Pohl
Fachbereich Biologie,
Entwicklungsbiologie
Philipps-Universität Marburg
Karl-von-Frisch-Straße 8b
35043 Marburg
Germany

Peter Rice
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge, CB10 1SD
UK

Tracey Rigby
MicroArray Laboratory
National Research Council of Canada
Biotechnology Research Institute
6100 Royalmount Avenue
Montreal
Quebec, H4P 2R2
Canada

Stephen C. Roemer
Fisher Scientific
Chemical Division
One Reagent Lane
Fairlawn
New Jersey, 07410
USA

Clémentine Sallée
University of Montreal
3101, chemin de la tour
Montreal
Quebeck, H3C 3J7
Canada

Christa Schleper
Department of Biology
University of Bergen
Jahnebakken 5
Box 7800
5020 Bergen
Norway

David C. Schriemer
Dept. of Biochemistry and
Molecular Biology
University of Calgary
3330 Hospital Drive N.W.
Calgary
Alberta, T"N 4N1
Canada

Agnieszka Sekowska
Institut Pasteur
Unité de Génétique des
Génomes Bactériens
Département Structure et
Dynamique des Génomes
28 rue du Docteur Roux
75724 Paris Cedex 15
France

Christoph W. Sensen
Faculty of Medicine
Sun Center of Excellence for
Visual Genome Research
University of Calgary
3330 Hospital Drive NW
Calgary
Alberta, T2N 4N1
Canada

Lorraine Sheremeta
Health Law Institute at the
University of Alberta
University of Alberta
402 Law Centre
Edmonton
Alberta, T6G 2H5
Canada

Julie Stromer
University of Calgary
Department of Biochemistry and
Molecular Biology
3330 Hospital Drive NW
Calgary
Alberta, T2N 4N1
Canada

Daniel C. Tessier
IatroQuest Corporation
1000 Chemin du Golf
Verdun
Quebec, H3E 1H4
Canada

David Y. Thomas
Department of Biochemistry
McGill University
3655 Promenade Sir William Osler
Montreal
Quebec, H3G 1Y6
Canada

Theerayut Toojinda
Rice Gene Discovery
National Center for Genetic Engineering and Biotechnology
Kasetsart University
Kamphangsaen
Nakorn Pathom, 73140
Thailand

Somvong Tragoonrung
Rice Gene Discovery
National Center for Genetic Engineering and Biotechnology
Kasetsart University
Kamphangsaen
Nakorn Pathom, 73140
Thailand

Alexander H. Treusch
University of Bergen
Department of Biology
Jahnebakken 5
Box 7800
5020 Bergen
Norway

Andrei Turinsky
University of Calgary
Department of Biochemistry and Molecular Biology
3330 Hospital Drive NW
Calgary
Alberta, T2N 4N1
Canada

Apichart Vanavichit
Center of Excellence for Rice Molecular Breeding and Product Development
National Center for Agricutural Biotechnology
Kasetsart University
Kamphangsaen
Nakorn Pathom, 73140
Thailand

Hans J. Vogel
Department of Biological Sciences
University of Calgary
Calgary
Alberta, T2N 1N4
Canada

Aalim M. Weljie
Chenomx Inc.
#800, 10050 - 112 St.
Edmonton
Alberta, T5K 2J1
Canada

David S. Wishart
Departments of Biological Sciences and Computing Science
University of Alberta
Edmonton
Alberta, T6G 2E8
Canada

Shu Jian Wu
Bristol Myers Squibb Pharmaceutical
Research Institute
311 Pennington-Rocky Hill Road
Pennington
New Jersey, 08534
USA

Emily Xu
Department of Biochemistry and
Molecular Biology
University of Calgary
3330 Hospital Drive N.W.
Calgary
Alberta, T2N 4N1
Canada

Katherine G. Zulak
Department of Biological Sciences
University of Calgary
2500 University Drive N.W.
Calgary
Alberta, T2N 1N4
Canada

Part I
Key Organisms

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

1
Genome Projects on Model Organisms

Alfred Pühler, Doris Jording, Jörn Kalinowski, Detlev Buttgereit, Renate Renkawitz-Pohl, Lothar Altschmied, Antoin Danchin, Agnieszka Sekowska, Horst Feldmann, Hans-Peter Klenk, and Manfred Kröger

1.1 Introduction

Genome research enables the establishment of the complete genetic information of organisms. The first complete genome sequences established were those of prokaryotic and eukaryotic microorganisms, followed by those of plants and animals (see, for example, the TIGR web page at http://www.tigr.org/). The organisms selected for genome research were mostly those which were already important in scientific analysis and thus can be regarded as model organisms. In general, organisms are defined as model organisms when a large amount of scientific knowledge has been accumulated in the past. For this chapter on genome projects of model organisms, several experts in genome research have been asked to give an overview of specific genome projects and to report on the respective organism from their specific point of view. The organisms selected include prokaryotic and eukaryotic microorganisms, and plants and animals.

We have chosen the prokaryotes *Escherichia coli*, *Bacillus subtilis*, and *Archaeoglobus fulgidus* as representative model organisms. The *E. coli* genome project is described by M. KRÖGER (Giessen, Germany). He gives an historical outline of the intensive research on microbiology and genetics of this organism, which cumulated in the *E. coli* genome project. Many of the technological tools currently available have been developed during the course of the *E. coli* genome project. *E. coli* is without doubt the best-analyzed microorganism of all. The knowledge of the complete sequence of *E. coli* has confirmed its reputation as the leading model organism of Gram$^-$ eubacteria.

A. DANCHIN and A. SEKOWSKA (Paris, France) report on the genome project of the environmentally and biotechnologically relevant Gram$^+$ eubacterium *B. subtilis*. The contribution focuses on the results and analysis of the sequencing effort and gives several examples of specific and sometimes unexpected findings of this project. Special emphasis is given to genomic data which

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

support the understanding of general features such as translation and specific traits relevant for living in its general habitat or its usefulness for industrial processes.

A. fulgidus is the subject of the contribution by H.-P. Klenk (Feldafing, Germany). Although this genome project was started before the genetic properties of the organism had been extensively studied, its unique lifestyle as a hyperthermophilic and sulfate-reducing organism makes it a model for a large number of environmentally important microorganisms and species with high biotechnological potential. The structure and results of the genome project are described in the contribution.

The yeast *Saccharomyces cerevisiae* has been selected as a representative eukaryotic microorganism. The yeast project is presented by H. Feldmann (Munich, Germany). *S. cerevisiae* has a long tradition in biotechnology and a long-term research history as a eukaryotic model organism *per se*. It was the first eukaryote to be completely sequenced and has led the way to sequencing other eukaryotic genomes. The wealth of the yeast's sequence information as useful reference for plant, animal, or human sequence comparisons is outlined in the contribution.

Among the plants, the small crucifer *Arabidopsis thaliana* was identified as the classical model plant, because of simple cultivation and short generation time. Its genome was originally considered to be the smallest in the plant kingdom and was therefore selected for the first plant genome project, which is described here by L. Altschmied (Gatersleben, Germany). The sequence of *A. thaliana* helped to identify that part of the genetic information unique to plants. In the meantime, other plant genome sequencing projects were started, many of which focus on specific problems of crop cultivation and nutrition.

The roundworm *Caenorhabditis elegans* and the fruitfly *Drosophila melanogaster* have been selected as animal models, because of their specific model character for higher animals and also for humans. The genome project of *C. elegans* is summarized by D. Jording (Bielefeld, Germany). The contribution describes how the worm - despite its simple appearance - became an interesting model organism for features such as neuronal growth, apoptosis, or signaling pathways. This genome project has also provided several bioinformatic tools which are widely used for other genome projects.

The genome project concerning the fruitfly *D. melanogaster* is described by D. Buttgereit and R. Renkawitz-Pohl (Marburg, Germany). *D. melanogaster* is currently the best-analyzed multicellular organism and can serve as a model system for features such as the development of limbs, the nervous system, circadian rhythms and even for complex human diseases. The contribution gives examples of the genetic homology and similarities between *Drosophila* and the human, and outlines perspectives for studying features of human diseases using the fly as a model.

1.2 Genome Projects of Selected Prokaryotic Model Organisms

1.2.1 The Gram$^-$ Enterobacterium *Escherichia coli*

1.2.1.1 The Organism

The development of the most recent field of molecular genetics is directly connected with one of the best described model organisms, the eubacterium *Escherichia coli*. There is no textbook in biochemistry, genetics, or microbiology which does not contain extensive sec-

tions describing numerous basic observations first noted in *E. coli* cells, or the respective bacteriophages, or using *E. coli* enzymes as a tool. Consequently, several monographs solely devoted to *E. coli* have been published. Although it seems impossible to name or count the number of scientists involved in the characterization of *E. coli*, Tab. 1.1 is an attempt to name some of the most deserving people in chronological order.

The scientific career of *E. coli* (Fig. 1.1) started in 1885 when the German pediatrician T. Escherich described isolation of the first strain from the feces of new-born babies. As late as 1958 this discovery was recognized internationally by use of his name

Table 1.1. Chronology of the most important primary detection and method applications with *E. coli*.

1886	"bacterium coli commune" by T. Escherich
1922	Lysogeny and prophages by d'Herelle
1940	Growth kinetics for a bacteriophage by M. Delbrück (Nobel prize 1969)
1943	Statistical interpretation of phage growth curve (game theorie) by S. Luria (Nobel prize 1969)
1947	Konjugation by E. Tatum and J. Lederberg (Nobel prize 1958) Repair of UV-damage by A. Kelner and R. Dulbecco (Nobel prize for tumor virology)
1954	DNA as the carrier of genetic information, proven by use of radioisotopes by M. Chase and A. Hershey (Nobel prize 1969)
1959	Phage immunity as the first example of gene regulation by A. Lwoff (Nobel prize 1965) Transduction of *gal*-genes (first isolated gene) by E. and J. Lederberg Host-controlled modification of phage DNA by G. Bertani and J.J. Weigle
1959	DNA-polymerase I by A. Kornberg (Nobel prize 1959) Polynucleotide-phosphorylase (RNA synthesis) by M. Grunberg-Manago and S. Ochoa (Nobel prize 1959)
1960	Semiconservative duplication of DNA by M. Meselson and F. Stahl
1961	Operon theory and induced fit by F. Jacob and J. Monod (Nobel prize 1965)
1964	Restriction enzymes by W. Arber (Nobel prize 1978)
1965	Physical genetic map with 99 genes by A.L. Taylor and M.S. Thoman Strain collection by B. Bachmann
1968	DNA-ligase by several groups contemporaneously
1976	DNA-hybrids by P. Lobban and D. Kaiser
1977	Recombinant DNA from *E. coli* and SV40 by P. Berg (Nobel prize 1980) Patent on genetic engineering by H. Boyer and S. Cohen
1978	Sequencing techniques using *lac* operator by W. Gilbert and *E. coli* polymerase by F. Sanger (Nobel prize 1980)
1979	Promoter sequence by H. Schaller Attenuation by C. Yanowsky General ribosome structure by H.G. Wittmann
1979	Rat insulin expressed in *E. coli* by H. Goodmann Synthetic gene expressed by K. Itakura and H. Boyer
1980	Site directed mutagenesis by M. Smith (Nobel prize 1993)
1985	Polymerase chain reaction by K.B. Mullis (Nobel prize 1993)
1988	Restriction map of the complete genome by Y. Kohara and K. Isono
1990	Organism-specific sequence data base by M. Kröger
1995	Total sequence of *Haemophilus influenzae* using an *E. coli* comparison
1999	Systematic sequence finished by a Japanese consortium under leadership of H. Mori
2000	Systematic sequence finished by F. Blattner
2000	Three-dimensional structure of ribosome by four groups contemporaneously

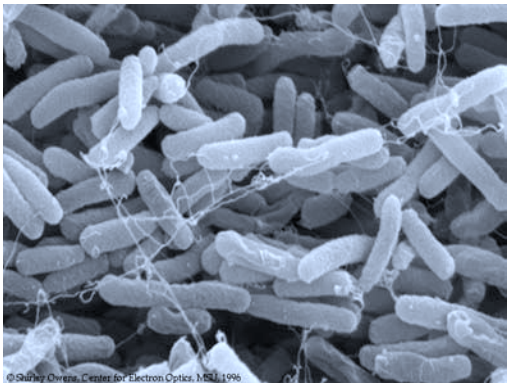


Fig. 1.1 Scanning electron micrograph (SEM) of *Escherichia coli* cells. (Image courtesy of Shirley Owens, Center for Electron Optics, MSU; found at http://commtechab.msu.edu/sites/dlc-me/zoo/ zah0700.html#top#top)

to classify this group of bacterial strains. In 1921 the very first report on virus formation was published for *E. coli*. Today we call the respective observation "lysis by bacteriophages". In 1935 these bacteriophages became the most powerful tool in defining the characteristics of individual genes. Because of their small size, they were found to be ideal tools for statistical calculations performed by the former theoretical physicist M. Delbrück. His very intensive and successful work has attracted many others to this area of research. In addition, Delbrück's extraordinary capability to catalyze the exchange of ideas and methods yielded the legendary Cold Spring Harbor Phage course. Everybody interested in basic genetics has attended this famous summer course or at least came to the respective annual phage meeting. This course, which was an ideal combination of joy and work, became an ideal means of spreading practical methods. For many decades it was the most important exchange forum for results and ideas, and strains and mutants. Soon, the so called "phage family" was formed, which interacted almost like one big laboratory; for example, results were communicated preferentially by means of preprints. Finally, 15 Nobel prize-winners have their roots in this summer-school (Tab. 1.1).

The substrain *E. coli* K12 was first used by E. Tatum as a prototrophic strain. It was chosen more or less by chance from the strain collection of the Stanford Medical School. Because it was especially easy to cultivate and because it is, as an inhabitant of our gut, a nontoxic organism by definition, the strain became very popular. Because of the vast knowledge already acquired and because it did not form fimbriae, *E. coli* K12 was chosen in 1975 at the famous Asilomar conference on biosafety as the only organism on which early cloning experiments were permitted [1]. No wonder that almost all subsequent basic observations in the life sciences were obtained either with or within *E. coli*. What started as the "phage family", however, dramatically split into hundreds of individual groups working in tough competition. As one of the most important outcomes, sequencing of *E. coli* was performed more than once. Because of the separate efforts, the genome finished only as number seven [2–4]. The amount of knowledge acquired, however, is certainly second to none and the way this knowledge was acquired is interesting, both in the history of sequencing methods and bioinformatics, and because of its influence on national and individual pride.

Work on *E. coli* is not finished with completion of the DNA sequence; data will be continuously acquired to fully characterize the genome in terms of genetic function and protein structures [5]. This is very important, because several toxic *E. coli* strains are known. Thus research on *E. coli* has turned from basic science into applied medical research. Consequently, the human toxic strain O157 has been completely sequenced, again more than once (unpublished).

1.2.1.2

Characterization of the Genome and Early Sequencing Efforts

With its history in mind and realizing the impact of the data, it is obvious that an ever growing number of colleagues worldwide worked with or on *E. coli*. Consequently, there was an early need for organization of the data. This led to the first physical genetic map, comprising 99 genes, of any living organism, published in by Taylor and Thoman [6]. This map was improved and was refined for several decades by Bachmann [7] and Berlyn [8]. These researchers still maintain a very useful collection of strains and mutants at Yale University. One thousand and twenty-seven loci had been mapped by 1983 [7]; these were used as the basis of the very first sequence database specific to a single organism [4]. As shown in Fig. 2 of Kröger and Wahl [4], sequencing of *E. coli* started as early as 1967 with one of the first ever characterized tRNA sequences. Immediately after DNA sequencing had been established numerous laboratories started to determine sequences of their personal interest.

1.2.1.3

Structure of the Genome Project

In 1987 Isono's group published a very informative and incredibly exact restriction map of the entire genome [9]. With the help of K. Rudd it was possible to locate sequences quite precisely [8, 10]. But only very few saw any advantage in closing the sometimes very small gaps, and so a worldwide joint sequencing approach could not be established. Two groups, one in Kobe, Japan [3] and one in Madison, Wisconsin [2] started systematic sequencing of the genome in parallel, and another laboratory, at Harvard University, used *E. coli* as a target to develop new sequencing technology. Several meetings, organized especially on *E. coli*, did not result in a unified systematic approach, thus many genes have been sequenced two or three times. Although specific databases have been maintained to bring some order into the increasing chaos, even this type of tool has been developed several times in parallel [4, 10]. Whenever a new contiguous sequence was published, approximately 75 % had already previously been submitted to the international databases by other laboratories. The progress of data acquisition followed a classical e-curve, as shown in Fig. 2 of Kröger and Wahl [4]. Thus in 1992 it was possible to predict the completeness of the sequence for 1997 without knowledge of the enormous technical innovations in between [4].

Both the Japanese consortium and the group of F. Blattner started early; some people say they started too early. They subcloned the DNA first and used manual sequencing and older informatic systems. Sequencing was performed semi-automatically, and many students were employed to read and monitor the X-ray films. When the first genome sequence of *Haemophilus influenzae* appeared in 1995 the science foundations wanted to discontinue support of *E. coli* projects, which received their grant support mainly because of the model character of the sequencing techniques developed.

Three facts and truly international protest convinced the juries to continue financial support. First, in contrast with the other completely sequenced organisms, *E. coli* is an autonomously living organism. Second, when the first complete very small genome sequence was released, even the longest contiguous sequence for *E. coli* was already longer. Third, the other laboratories could only finish their sequences because the *E. coli* sequences were already publicly available. Consequently, the two main competing laboratories were allowed to purchase several of the sequencing machines already developed and use the shotgun approach to complete their efforts. Finally, they finished almost at the same time. H. Mori and his colleagues included already published sequences from other laboratories in their sequence data and sent them to the international databases on December 28th, 1996 [3] and F. Blattner reported an entirely new sequence on January 16th, 1997 [2]. They added the last changes and additions as late as October, 1998. Very sadly, at the end *E. coli* had been sequenced almost three times [4]. Nowadays, however, most people forget about all the other sources and refer to the Blattner sequence.

1.2.1.4

Results from the Genome Project

When the sequences were finally finished, most of the features of the genome were already known. Consequently, people no longer celebrate the *E. coli* sequence as a major breakthrough. At that time everybody knew the genome was almost completely covered with genes, although fewer than half had been genetically characterized. Tab. 1.2 illustrates this and shows the counting differences. Because of this high density of genes, F. Blattner and coworkers defined "gray holes" whenever they found a noncoding region of more than 2 kb [2]. It was found that the termination of replication is almost exactly opposite to the origin of replication. No special differences have been found for either direction of replication. Approximately 40 formerly described genetic features could not be located or supported by the sequence [4, 8]. On the other hand, there are several examples of multiple functions encoded by the same gene. It was found that the multifunctional genes are mostly involved in gene expression and used as a general control factor. M. Riley determined the number of gene duplications, which is also not unexpectedly low when neglecting the ribosomal operons [10].

Everybody is convinced that the real work is starting only now. Several strain differences might be the cause of the deviations between the different sequences available. Thus the numbers of genes and nucleotides differ slightly (Tab. 1.2). Everybody would like to know the function of each of the open reading frames [5], but nobody has received the grant money to work on this important problem. Seemingly, other model organisms are of more public interest; thus it might well be that research on other organisms will now help our understanding of *E. coli*, in just the same way that *E. coli* provided information enabling understanding of them. In contrast with yeast, it is very hard to produce knock-out mutants. Thus, we might have the same situation in the postgenomic era as we had before the genome was finished. Several laboratories will continue to work with *E. coli*, they will constantly characterize one or the other open reading frame, but there will be no mutual effort [5]. A simple and highly efficient method using PCR products to inactivate chromosomal genes was recently developed [11]. This method has greatly facilitated systematic mutagenesis approaches in *E. coli*.

Table 1.2 Some statistical features of the *E. coli* genome.

Total size	***4,639,221 bp***[1)]		
		According to Regulon[4)]	***According to Blattner***[5)]
Transcription units	Proven	528	
	Predicted	2328	
Genes	Total found	4408	4403
	Regulatory	85	
	Essential	200	
	Nonessential[2)]	2363	1897
	Unknown[3)]	1761	2376
	tRNA	84	84
	rRNA	29	29
Promoters	Proven	624	
	Predicted	4643	
Sites		469	
Regulatory interactions	Found	642	
	Predicted	275	
Terminators	Found	96	
RBS		98	
Gene products	Regulatory proteins	85	
	RNA	115	115
	Other peptides	4190	4201

1) Additional 63 bp compared with the original sequence
2) Genes with known or predicted function
3) No other data available other than the existence of an open reading frame with a start sequence and more than 100 codons
4) Data from http://tula.cifn.unam.mx/Computational_Genomics/regulondb/
5) Data from http://www.genome.wisc.edu

1.2.1.5

Follow-up Research in the Postgenomic Era

Today it seems more attractive to work with toxic *E. coli* strains, for example O157, than with *E. coli* K12. This strain has recently been completely sequenced; the data are available via the internet. Comparison of toxic and nontoxic strains will certainly help us to understand the toxic mechanisms. It was, on the other hand, found to be correct to use *E. coli* K12 as the most intensively used strain for biological safety regulations [1]. No additional features changed this. This *E. coli* strain is subject to comprehensive transcriptomics and proteomics studies. For global gene expression profiling different systems like an Affymetrix GeneChip and several oligonucleotide sets for the printing of microarrays are available. These tools have already been extensively

used by researchers during recent years. Proteomics studies resulted in a comprehensive reference map for the *E. coli* K-12 proteome (SWISS-2DPAGE, Two-dimensional polyacrylamide gel electrophoresis database, http://www.expasy.org/ch2d). The "Encyclopedia of *Escherichia coli* K-12 Genes and Metabolism" (EcoCyc) (www.ecocyc.org) is a very useful and constantly growing *E. coli* metabolic pathway database for the scientific community [12].

Surprisingly, colleagues from mathematics or informatics have shown the most interest in the bacterial sequences. They have performed all kinds of statistical analysis and tried to discover evolutionary roots. Here another fear of the public is already formulated – people are afraid of attempts to reconstruct the first living cell. So there are at least some attempts to find the minimum set of genes for the most basic needs of a cell. We have to ask again the very old question: Do we really want to "play God"? If so, *E. coli* could indeed serve as an important milestone.

1.2.2
The Gram$^+$ Spore-forming *Bacillus subtilis*

1.2.2.1
The Organism

Self-taught ideas have a long life – articles about *Bacillus subtilis* (Fig. 1.2) almost invariably begin with words such as: "*B. subtilis*, a soil bacterium ...", nobody taking the elementary care to check on what type of experimental observation this is based. *Bacillus subtilis*, first identified in 1885, is named *ko so kin* in Japanese and *laseczka sienna* in Polish, or "hay bacterium", and this refers to the real biotope of the organism, the surface of grass or low-lying plants [13]. Interestingly, it required its genome to be sequenced to acquire again its right biotope. Of course, plant leaves fall on the soil surface, and one must naturally find *B. subtilis* there, but its normal niche is the surface of leaves, the phylloplane. Hence, if one wishes to use this bacterium in industrial processes, to engineer its genome, or simply to understand the functions coded by its genes, it is of fundamental importance to understand where it normally thrives, and which environmental conditions control its life-cycle and the corresponding gene expression. Among other important ancillary functions, *B. subtilis* has thus to explore, colonize, and exploit local resources, while at the same time it must maintain itself, dealing with congeners and with other organisms: understanding *B. subtilis* requires understanding the general properties of its normal habitat.

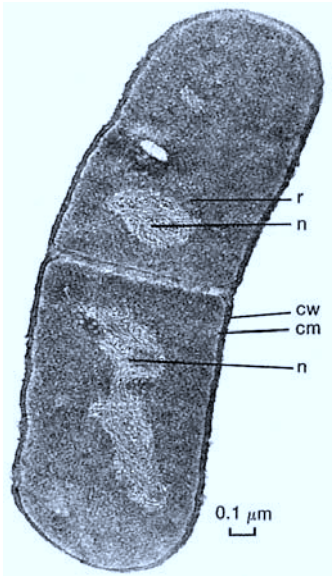


Fig. 1.2 Electron micrograph of a thin section of *Bacillus subtilis*. The dividing cell is surrounded by a relatively dense wall (CW), enclosing the cell membrane (cm). Within the cell, the nucleoplasm (n) is distinguishable by its fibrillar structure from the cytoplasm, densely filled with 70S ribosomes (r).

1.2.2.2

A Lesson from Genome Analysis: The *Bacillus subtilis* Biotope

The genome of *B. subtilis* (strain 168), sequenced by a team in European and Japanese laboratories, is 4,214,630 bp long (http://genolist.pasteur.fr/SubtiList/). Of more than 4100 protein-coding genes, 53 % are represented once. One quarter of the genome corresponds to several gene families which have probably been expanded by gene duplication. The largest family contains 77 known and putative ATP-binding cassette (ABC) permeases, indicating that, despite its large metabolism gene number, *B. subtilis* has to extract a variety of compounds from its environment [14]. In general, the permeating substrates are unchanged during permeation. Group-transfer, in which substrates are modified during transport, plays an important role in *B. subtilis*, however. Its genome codes for a variety of phosphoenolpyruvate-dependent systems (PTS) which transport carbohydrates and regulate general metabolism as a function of the nature of the supplied carbon source. A functionally-related catabolite repression control, mediated by a unique system (not cyclic AMP), exists in this organism [15]. Remarkably, apart from the expected presence of glucose-mediated regulation, it seems that carbon sources related to sucrose play a major role, via a very complicated set of highly regulated pathways, indicating that this plant-associated carbon supply is often encountered by the bacteria. In the same way, *B. subtilis* can grow on many of the carbohydrates synthesized by grass-related plants.

In addition to carbon, oxygen, nitrogen, hydrogen, sulfur, and phosphorus are the core atoms of life. Some knowledge about other metabolism in *B. subtilis* has accumulated, but significantly less than in its *E. coli* counterpart. Knowledge of its genome sequence is, however, rapidly changing the situation, making *B. subtilis* a model of similar general use to *E. coli*. A frameshift mutation is present in an essential gene for surfactin synthesis in strain 168 [16], but it has been found that including a small amount of a detergent into plates enabled these bacteria to swarm and glide extremely efficiently (C.-K. Wun and A. Sekowska, unpublished observations). The first lesson of genome text analysis is thus that *B. subtilis* must be tightly associated with the plant kingdom, with grasses in particular [17]. This should be considered in priority when devising growth media for this bacterium, in particular in industrial processes.

Another aspect of the *B. subtilis* life cycle consistent with a plant-associated life is that it can grow over a wide range of different temperatures, up to 54–55 °C – an interesting feature for large-scale industrial processes. This indicates that its biosynthetic machinery comprises control elements and molecular chaperones that enable this versatility. Gene duplication might enable adaptation to high temperature, with isozymes having low- and high-temperature optima. Because the ecological niche of *B. subtilis* is linked to the plant kingdom, it is subjected to rapid alternating drying and wetting. Accordingly, this organism is very resistant to osmotic stress, and can grow well in media containing 1 M NaCl. Also, the high level of oxygen concentration reached during daytime are met with protection systems – *B. subtilis* seems to have as many as six catalase genes, both of the heme-containing type (*katA*, *katB*, and *katX* in spores) and of the manganese-containing type (*ydbD*, PBX phage-associated *yjqC*, and *cotJC* in spores).

The obvious conclusion from these observations is that the normal *B. subtilis* niche is the surface of leaves [18]. This is consistent with the old observation that

B. subtilis makes up the major population of the bacteria of rotting hay. Furthermore, consistent with the extreme variety of conditions prevailing on plants, *B. subtilis* is an endospore-forming bacterium, making spores highly resistant to the lethal effects of heat, drying, many chemicals, and radiation.

1.2.2.3

To Lead or to Lag: First Laws of Genomics

Analysis of repeated sequences in the *B. subtilis* genome discovered an unexpected feature: strain 168 does not contain insertion sequences. A strict constraint on the spatial distribution of repeats longer than 25 bp was found in the genome, in contrast with the situation in *E. coli*. Correlation of the spatial distribution of repeats and the absence of insertion sequences in the genome suggests that mechanisms aimed at their avoidance and/or elimination have been developed [19]. This observation is particularly relevant for biotechnological processes in which one has multiplied the copy number of genes to improve production. Although there is generally no predictable link between the structure and function of biological objects, the pressure of natural selection has adapted together gene and gene products. Biases in features of predictably unbiased processes is evidence of prior selective pressure. With *B. subtilis* one observes a strong bias in the polarity of transcription with respect to replication: 70 % of the genes are transcribed in the direction of the replication fork movement [14]. Global analysis of oligonucleotides in the genome demonstrated there is a significant bias not only in the base or codon composition of one DNA strand relative to the other, but, quite surprisingly, there is a strong bias at the level of the amino-acid content of the proteins. The proteins coded by the leading strand are valine-rich and those coded by the lagging strand are threonine and isoleucine-rich. This first law of genomics seems to extend to many bacterial genomes [20]. It must result from a strong selection pressure of a yet unknown nature, demonstrating that, contrary to an opinion frequently held, genomes are not, on a global scale, plastic structures. This should be taken into account when expressing foreign proteins in bacteria.

Three principal modes of transfer of genetic material – transformation, conjugation, and transduction – occur naturally in prokaryotes. In *B. subtilis*, transformation is an efficient process (at least in some *B. subtilis* species such as the strain 168) and transduction with the appropriate carrier phages is well understood.

The unique presence in the *B. subtilis* genome of local repeats, suggesting Campbell-like integration of foreign DNA, is consistent with strong involvement of recombination processes in its evolution. Recombination must, furthermore, be involved in mutation correction. In *B. subtilis*, MutS and MutL homologs occur, presumably for the purpose of recognizing mismatched base pairs [21]. No counterpart of MutH activity, which would enable the daughter strand to be distinguished from its parent, has, however, been identified. It is, therefore, not known how the long-patch mismatch repair system corrects mutations in the newly synthesized strand. One can speculate that the nicks caused in the daughter strands by excision of newly misincorporated uracil instead of thymine during replication might provide the appropriate signal. Ongoing fine studies of the distribution of nucleotides in the genome might substantiate this hypothesis.

The recently sequenced genome of the pathogen *Listeria monocytogenes* has many features in common with that of the genome of *B. subtilis* [22]. Preliminary analysis

suggests that the *B. subtilis* genome might be organized around the genes of core metabolic pathways, such as that of sulfur metabolism [23], consistent with a strong correlation between the organization of the genome and the architecture of the cell.

1.2.2.4

Translation: Codon Usage and the Organization of the Cell's Cytoplasm

Exploiting the redundancy of the genetic code, coding sequences show evidence of highly variable biases of codon usage. The genes of *B. subtilis* are split into three classes on the basis of their codon usage bias. One class comprises the bulk of the proteins, another is made up of genes expressed at a high level during exponential growth, and a third class, with A + T-rich codons, corresponds to portions of the genome that have been horizontally exchanged [14].

When mRNA threads are emerging from DNA they become engaged by the lattice of ribosomes, and ratchet from one ribosome to the next, like a thread in a wiredrawing machine [24]. In this process, nascent proteins are synthesized on each ribosome, spread throughout the cytoplasm by the linear diffusion of the mRNA molecule from ribosome to ribosome. If the environmental conditions change suddenly, however, the transcription complex must often break up. Truncated mRNA is likely to be a dangerous molecule because, if translated, it would produce a truncated protein. Such protein fragments are often toxic, because they can disrupt the architecture of multisubunit complexes. A process copes with this kind of accident in *B. subtilis*. When a truncated mRNA molecule reaches its end, the ribosome stops translating, and waits. A specialized RNA, tmRNA, that is folded and processed at its 3′ end like a tRNA and charged with alanine, comes in, inserts its alanine at the C-terminus of the nascent polypeptide, then replaces the mRNA within a ribosome, where it is translated as ASFNQNVALAA. This tail is a protein tag that is then used to direct the truncated tagged protein to a proteolytic complex (ClpA, ClpX), where it is degraded [25].

1.2.2.5

Post-sequencing Functional Genomics: Essential Genes and Expression-profiling Studies

Sequencing a genome is not a goal *per se*. Apart from trying to understand how genes function together it is most important, especially for industrial processes, to know how they interact. As a first step it was interesting to identify the genes essential for life in rich media. The European–Japanese functional genomics consortium endeavored to inactivate all the *B. subtilis* genes one by one [26]. In 2004, the outcome of this work are still the first and only result in which we can list all the essential genes in bacteria. In this genome counting over 4100 genes, 271 seem to be essential for growth in rich medium under laboratory conditions (i.e. without being challenged by competition with other organisms or by changing environmental conditions). Most of these genes can be placed into a few large and predicable functional categories, for example information processing, cell envelope biosynthesis, shape, division, and energy management. The remaining genes, however, fall into categories not expected to be essential, for example some Embden–Meyerhof–Parnas pathway genes and genes involved in purine biosynthesis. This opens the perspective that these enzymes can have novel and unexpected functions in the cell. Interestingly, among the 26 essential genes that belongs to either "other functions" or "unknown genes" categories, seven belong to or carry the signature for (ATP/)GTP-

binding proteins – several now seem to code for tRNA modifications [27], but some could be in charge of coordination of essential processes listed above, as seems to be true for eukaryotes. A remarkable outcome of this project was the discovery that essential genes are grouped along the leading replication strands in the genome [28]. This feature is general and indicates that genes cannot, at least under strong competitive conditions, be shuffled randomly in the chromosome. This has important consequences for genetic manipulation of organisms of biotechnological interest.

Beside identification of *B. subtilis* essential genes, the European–Japanese project produced an almost complete representative collection of mutants of this bacterium. This collection is freely available to the scientific community, in particular for biotechnology-oriented studies. This strategy is a good example of genome-wide approaches that would have been unthinkable two decades ago. The obvious continuation in this line was the use of transcriptome analysis (identification of all transcripts on DNA arrays under a variety of experimental conditions). Several dozen reports appeared in the literature in those years dealing with data obtained by this global approach. With different technical solutions (from commercially available macroarrays with radioactive labeling through custom-made glass microarrays with fluorescent labeling) they offer an almost exhaustive point of view at the level of transcription answering a given question, assuming particular attention has been devoted to controlling all upstream experimental steps (RNA preparation, cDNA synthesis) and to making use of well-chosen statistical analyses. Many reports have been devoted to the study of heat-shock proteins but not much work was devoted to the equally important cold-shock proteins of the bacteria. The two-component system *desKR* was recently identified; this regulates expression of *des* gene coding for desaturase, which participates in cold adaptation through membrane lipid modification. To discover whether the *desKR* system is exclusively devoted to *des* regulation or constitutes a cold-triggered regulatory system of global relevance, macroarray studies seemed to be the method of choice [29]. A major outcome of this study was, it seemed, that the *desKR* system controls *des* gene expression only. Unexpectedly, this work uncovered many novel partners involved in cold shock response, with almost half of these genes annotated as carrying an unknown function. The categories of genes affected by the cold-shock response were, as expected, the cold shock protein genes that were already known, but also heat-shock protein genes and genes involved in translation machinery, amino acid biosynthesis, nucleoid structure, ABC permeases (for acetoin in particular), purine and pyrimidine biosynthesis and glycolysis, the citric acid cycle, and ATP synthesis. Because these last categories are found in most transcriptome studies, it remains to be seen whether they are indeed specific to cold shock. Among genes of unknown function particularly interesting in biotechnology one can mention the *yplP* gene, which codes for a transcriptional regulator that belongs to the NtrC/NifA family and which, when inactivated, causes a cold-specific late-growth phenotype. However, the exact role of YplP protein remains to be understood.

An essential complement to transcriptome studies is exploration of the bacterial proteome, which gives a detailed look at the behavior of the final players. Because what really counts for a cell is the final level of its mature proteins, and because many post-transcriptional modifications can alter the fate of mRNA translation products into mature proteins, transcriptome analysis alone

provides only an approximation of what is going to really happen in the cell. Studying the proteome expressed under specific conditions can help uncover interesting links between different parts of metabolism. This has been explored under conditions important for biotechnology, for example the salt-stress response and iron metabolism [30]. This recent work aiming at establishing the network of proteins affected by osmotic stress has shown that *B. subtilis* cells growing under high-salinity are subjected to iron limitation, as indicated by the increase of expression of several putative iron-uptake systems (*fhuD*, *fhuB*, *feuA*, *ytiY* and *yfmC*) or iron siderophore bacillibactin synthesis and modification genes (*dhbABCE*). The derepression of the *dhb* operon seems to be more a salt-specific effect than a general osmotic effect, because it is not produced by addition of iso-osmotic non-ionic osmolites (sucrose or maltose) to the growth medium. Some high-salinity growth-defect phenotypes could, furthermore, be reversed by supplementation of the medium with excess iron. This work shows that two distinct factors important in fermentors – iron limitation and high-salinity stress – hitherto regarded as separate growth-limiting factors, are indeed not so separate.

1.2.2.6

Industrial Processes

Bacillus subtilis is generally recognized as safe (GRAS). It is much used industrially both for enzyme production and for food-supply fermentation. Riboflavin is derived from genetically modified *B. subtilis* by use of fermentation techniques. For some time high levels of heterologous gene expression in *B. subtilis* was difficult to achieve. In contrast with Gram-negatives, A + T-rich Gram-positive bacteria have optimized transcription and translation signals; although *B. subtilis* has a counterpart of the *rpsA* gene, this organism lacks the function of the corresponding ribosomal S1 protein which enables recognition of the ribosome-binding site upstream of the translation start codons [31]. Traditional techniques (e.g. random mutagenesis followed by screening; *ad hoc* optimization of poorly defined culture media) are important and will continue to be used in the food industry, but biotechnology must now include genomics to target artificial genes that follow the sequence rules of the genome at a precise position, adapted to the genome structure, and to modify intermediary metabolism while complying with the adapted niche of the organism, as revealed by its genome. As a complement to standard genetic engineering and transgenic technology, knowing the genome text has opened a whole new range of possibilities in food-product development, in particular enabling "humanization" of the content of food products (adaptation to the human metabolism, and even adaptation to sick or healthy conditions). These techniques provide an attractive means of producing healthier food ingredients and products that are currently not available or are very expensive. *B. subtilis* will remain a tool of choice in this respect.

1.2.2.7

Open Questions

The complete genome sequence of *B. subtilis* contains information that remains underutilized in the current prediction methods applied to gene functions, most of which are based on similarity searches of individual genes. In particular, it is now clear that the order of the genes in the chromosome is not random, and that some hot spots enable gene insertion without much damage whereas other regions are forbidden. For production of small molecules one must use higher-level information on meta-

bolic pathways to reconstruct a complete functional unit from a set of genes. The reconstruction *in silico* of selected portions of metabolism using the existing biochemical knowledge of similar gene products has been undertaken. Although the core biosynthetic pathways of all twenty amino acids have been completely reconstructed in *B. subtilis*, many satellite or recycling pathways, in particular the synthesis of pyrimidines, have not yet been identified in sulfur and short-carbon-chain acid metabolism. Finally, there remain some 800 genes of completely unknown function in the genome of strain 168, including a few tens of "orphan" genes that have no counterpart in any known genome, and many more in the genome of related species. It remains to be understood whether they play an important role for biotechnological processes.

1.2.3 The Archaeon *Archaeoglobus fulgidus*

1.2.3.1 The Organism

Archaeoglobus fulgidus is a strictly anaerobic, hyperthermophilic, sulfate-reducing archaeon. It is the first sulfate-reducing organism for which the complete genome sequence has been determined and published [32]. Sulfate-reducing organisms are essential to the biosphere, because biological sulfate reduction is part of the global sulfur cycle. The ability to grow by sulfate-reduction is restricted to a few groups of prokaryotes only. The Archaeoglobales are in two ways unique within this group:

1. they are members of the Archaea and therefore unrelated to all other sulfate reducers, which belong to the Bacteria; and
2. the Archaeoglobales are the only hyperthermophiles within the sulfate-reducers, a feature which enables them to occupy extreme environments, for example hydrothermal fields and sub-surface oil fields.

The production of iron sulfide as an end product of high-temperature sulfate reduction by *Archaeoglobus* species contributes to oil-well "souring", which causes corrosion of iron and steel in submarine oil- and gas-processing systems. *A. fulgidus* is also a model for hyperthermopilic organisms and for the Archaea, because it is only the second hyperthermophile whose genome has been completely deciphered (after *Methanococcus jannaschii*), and it is the third species of Archaea (after *M. jannaschii* and *Methanobacterium thermoautotrophicum*) whose genome has been completely sequenced and published.

A. fulgidus DSM4304 (Fig. 1.3) is the type strain of the Archaeoglobales [33]. Its glycoprotein-covered cells are irregular spheres (diameter 2 µm) with four distinct monopolar flagella. It grows not only organoheterotrophically, using a variety of carbon and energy sources, but also lithoautotrophically on hydrogen, thiosulfate, and carbon dioxide. Within the range 60–95 °C it grows best at 83 °C.

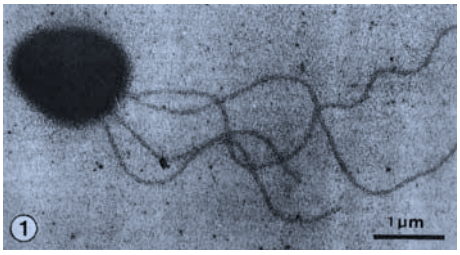


Fig. 1.3 Electron-micrograph of *A. fulgidus* DSM4303 (strain VC-16), kindly provided by K.O. Stetter, University of Regensburg. The bar in the lower right corner represents 1 µm.

Before genome sequencing very little was known about the genomic organization of *A. fulgidus*. The first estimate of its genome size, obtained by use of pulsed field gel electrophoresis, was published after final assembly of the genome sequences had already been achieved. Because extra-chromosomal elements are absent from *A. fulgidus*, it was determined that the genome consists of only one circular chromosome. Although data about genetic or physical mapping of the genome were unknown before the sequencing project, a small-scale approach to physical mapping was performed late in the project for confirmation of the genome assembly. Sequences of only eleven genes from *A. fulgidus* had been published before the sequencing project started; these covered less than 0.7 % of the genome.

1.2.3.2

Structure of the Genome Project

The whole-genome random sequencing procedure was chosen as sequencing strategy for the *A. fulgidus* genome project. This procedure had previously been applied to four microbial genomes sequenced at The Institute for Genomic Research (TIGR): *Haemophilus influenzae*, *Mycoplasma genitalium*, *M. jannaschii*, and *Helicobacter pylori* [34]. Chromosomal DNA for the construction of libraries was prepared from a culture derived from a single cell isolated by means of optical tweezers and provided by K.O. Stetter. Three libraries were used for sequencing – two plasmid libraries (1.42 kbp and 2.94 kbp insert size) for mass sequence production and one large insert λ-library (16.38 kbp insert size) for the genome scaffold. The initial random sequencing phase was performed with these libraries until 6.7-fold sequence coverage was achieved. At this stage the genome was assembled into 152 contigs separated by sequence gaps and five groups of contigs separated by physical gaps. Sequence gaps were closed by a combined approach of editing the ends of sequence traces and by primer walking on plasmid- and λ-clones spanning the gaps. Physical gaps were closed by direct sequencing of PCR-fragments generated by combinatorial PCR reactions. Only 0.33 % of the genome (90 regions) was covered by only one single sequence after the gap-closure phase. These regions were confirmed by additional sequencing reactions to ensure a minimum sequence coverage of two for the whole genome. The final assembly consisted of 29,642 sequencing runs which cover the genome sequence 6.8-fold.

The *A. fulgidus* genome project was financed by the US Department of Energy (DOE) within the Microbial Genome Program. This program financed several of the early microbial genome-sequencing projects performed at a variety of genome centers, for example *M. jannaschii* (TIGR), *M. thermoautotrophicum* (Genome Therapeutics), *Aquifex aeolicus* (Recombinant BioCatalysis, now DIVERSA), *Pyrobaculum aerophilum* (California Institute of Technology), *Pyrococcus furiosus* (University of Utah), and *Deinococcus radiodurans* (Uniformed Services University of the Health Sciences). Like the *M. jannaschii* project which was started one year earlier, the *A. fulgidus* genome was sequenced and analyzed in a collaboration between researchers at TIGR and Carl R. Woese and Gary J. Olsen at the Department of Microbiology at the University of Illinois, Champaign-Urbana. The plasmid libraries were constructed in Urbana, whereas the λ-library was constructed at TIGR. Sequencing and assembly was performed at TIGR using automated ABI sequencers and a TIGR assembler, respectively. Confirmation of the assembly by mapping with large-size restriction fragments was performed in Urbana. Open reading frame (ORF) prediction

and identification of functions, and the data mining and interpretation of the genome content was performed jointly by both teams.

Coding regions in the final genome sequence were identified with a combination of two sets of ORF generated by programs developed by members of the two teams – GeneSmith, by H.O. Smith at TIGR and CRITICA, by G.J. Olsen and J.H. Badger in Urbana. The two sets of ORF identified by GeneSmith and CRITICA were merged into one consensus set containing all members of both initial sets. The amino acid sequences derived from the consensus set were compared with a non-redundant protein database using BLASTX. ORFs shorter than 30 codons were carefully inspected for database hits and eliminated when there was no significant database match. The results of the database comparisons were first inspected and categorized by TIGR's microbial annotation team. This initial annotation database was then further analyzed and refined by a team of experts for all major biological role categories.

The sequencing strategy chosen for the *A. fulgidus* genome project has some advantages compared with alternative strategies applied in genome research:

1. Given the relatively large set of automated sequencers available at TIGR, the whole-genome random sequencing procedure is much faster than any strategy that includes a mapping step before the sequencing phase;
2. Within the DOE Microbial Genome Program the TIGR strategy and the sequencing technology used for the *M. jannaschii* and *A. fulgidus* genome projects proved to be clearly superior in competition with projects based on multiplex sequencing (*M. thermoautotrophicum* and *P. furiosus*), by finishing two genomes in less time than the competing laboratories needed for one genome each; and
3. The interactive annotation with a team of experts for the organism and for each biological category ensured a more sophisticated final annotation than any automated system could achieve at that time.

1.2.3.3

Results from the Genome Project

Although the initial characterization of the genome revealed all its basic features, annotation of biological functions for the ORF will continue to be updated for new functions identified either in *A. fulgidus* or for homologous genes characterized in other organisms. The size of the *A. fulgidus* genome was determined to be 2,178,400 bp, with an average G + C content of 48.5 %. Three regions with low G + C content (<39 %) were identified, two of which encode enzymes for lipopolysaccharide biosynthesis. The two regions with the highest G + C content (>53 %) contain the ribosomal RNA and proteins involved in heme biosynthesis. With the bioinformatics tools available when genome characterization was complete, no origin of replication could be identified. The genome contains only one set of genes for ribosomal RNA. Other RNA encoded in *A. fulgidus* are 46 species of tRNA, five of them with introns 15–62 bp long, no significant tRNA clusters, 7S RNA and RNase P. All together 0.4 % of the genome is covered by genes for stable RNA. Three regions with short (< 49 bp) non-coding repeats (42–60 copies) were identified. All three repeated sequences are similar to short repeated sequences found in *M. jannaschii* [35]. Nine classes of long, coding repeats (> 95 % sequence identity) were identified within the genome, three of them might represent IS elements, and three other repeats encode conserved hypothetical proteins found previously in other genomes. The consensus set of ORF contains 2436 members with an average length of

822 bp, similar to *M. jannaschii* (856 bp), but shorter than in most bacterial genomes (average 949 bp). With 1.1 ORF per kb, the gene density seems to be slightly higher than in other microbial genomes, although the fraction of the genome covered by protein coding genes (92.2 %) is comparable with that for other genomes. The elevated number of ORF per kbp might be artificial, because of a lack of stop codons in high G + C organisms. Predicted start codons are 76 % ATG, 22 % GTG, and 2 % TTG. No inteins were identified in the genome. The isoelectric point of the predicted proteins in *A. fulgidus* is rather low (median p*I* is 6.3); for other prokaryotes distributions peak between 5.5 and 10.5. Putative functions could be assigned to about half of the predicted ORF (47 %) by significant matches in database searches. One quarter (26.7 %) of all ORF are homologous to ORF previously identified in other genomes ("conserved hypotheticals"), whereas the remaining quarter (26.2 %) of the ORF in *A. fulgidus* seem to be unique, without any significant database match. *A. fulgidus* contains an unusually large number of paralogous gene families: 242 families with 719 members (30 % of all ORF). This might explain why the genome is larger than most other archaeal genomes (average approximately 1.7 Mbp). Interestingly, one third of the identified families (85 out of 242) have no single member for which a biological function could be predicted. The largest families contain genes assigned to "energy metabolism", "transporters", and "fatty acid metabolism".

The genome of *A. fulgidus* is neither the first archaeal genome to be sequenced completely nor is it the first genome of a hyperthermophilic organism. The novelties for both features had already been reported together with the genome of *M. jannaschii* [35]. *A. fulgidus* is, however, the first sulfate-reducing organism whose genome was completely deciphered. The next genome of a sulfate reducer followed almost seven years later with that of *Desulfotalea psychrophila*, an organism whose optimum growth temperature is 75 ° lower than that of *A. fulgidus* [36]. Model findings in respect of sulfur and sulfate metabolism were not expected from the genome, because sulfate metabolism had already been heavily studied in *A. fulgidus* before the genome project. The genes for most enzymes involved in sulfate reduction were already published, and new information from the genome confirmed only that the sulfur oxide reduction systems in Archaea and Bacteria were very similar. The single most exciting finding in the genome of *A. fulgidus* was identification of multiple genes for acetyl-CoA synthase and the presence of 57 β-oxidation enzymes. It has been reported that the organism is incapable of growth on acetate [37], and no system for β-oxidation has previously been described in the Archaea. It appears now that *A. fulgidus* can gain energy by degradation of a variety of hydrocarbons and organic acids, because genes for a least five types of ferredoxin-dependent oxidoreductases and at least one lipase were also identified. Interestingly, at about the same time as the unexpected genes for metabolizing enzymes were identified, it was also reported that a close relative of *A. fulgidus* is able to grow on olive oil (K. O. Stetter, personal communication), a feature that would require the presence of the genes just identified in *A. fulgidus*. On the other hand, not all genes necessary for the pathways described in the organism could be identified. Glucose has been described as a carbon source for *A. fulgidus* [38], but neither an uptake-transporter nor a catabolic pathway for glucose could be identified in the genome. There is still a chance that the required genes are hidden in the pool of functionally

uncharacterized ORFs. Other interesting findings in respect of the biology of *A. fulgidus* concern sensory functions and regulation of gene expression. *A. fulgidus* seems to have complex sensory and regulatory networks – a major difference from results reported for *M. jannaschii* – consistent with its extensive energy-producing metabolism and versatile system for carbon utilization. These networks contain more than 55 proteins with presumed regulatory functions and several iron-dependent repressor proteins. At least 15 signal-transducing histidine kinases were identified, but only nine response regulators.

1.2.3.4

Follow-up Research

Almost 30 papers about *A. fulgidus* were published between the initial description of the organism in 1987 and the genome sequence ten years later. In the six years since the genome was finished *A. fulgidus* was mentioned in the title or abstract of more than 165 research articles. Although functional genomics is now a hot topic at many scientific meetings and discussions, *A. fulgidus* seems to be no prime candidate for such studies. So far not a single publication has dealt with proteomics, transcriptome, or serial mutagenesis in this organism. The recently published papers that refer to the *A. fulgidus* genome sequence fall into three almost equally represented categories: comparative genomics, structure of *A. fulgidus* proteins, and characterization of expressed enzymes. One of the most interesting follow up stories is probably that on the flap endonucleases. In October 1998 Hosfield et al. described newly discovered archaeal flap endonucleases (FEN) from *A. fulgidus, M. jannaschii* and *P. furiosus* with a structure-specific mechanism for DNA substrate binding and catalysis resembling human flap endonuclease [39]. In Spring 1999 Lyamichev et al. showed how FEN could be used for polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes [40]. In May 2000, Cooksey et al. described an invader assay based on *A. fulgidus* FEN that enables linear signal amplification for identification of mutatons [41]. This procedure could eventually become important as a non-PCR based procedure for SNP detection. The identification of 86 candidates for small non-messenger RNA by Tang et al. [42] and identification of the unusual organization of the putative replication origin region by Maisnier-Patin et al. [43] also mark noteworthy progress in our knowledge about the model organism *A. fulgidus*.

1.3 Genome Projects of Selected Eukaryotic Model Organisms

1.3.1 The Budding Yeast *Saccharomyces cerevisiae*

1.3.1.1

Yeast as a Model Organism

The budding yeast, *Saccharomyces cerevisiae* (Fig. 1.4), can be regarded as one of the most important fungal organisms used in biotechnological processes. It owes its name to its ability to ferment saccharose, and has served mankind for several thousand years in the making of bread and alcoholic beverages. The introduction of yeast as an experimental system dates back to the 1930s and has since attracted increasing attention. Unlike more complex eukaryotes, yeast cells can be grown on defined media giving the investigator complete control over environmental conditions. The elegance of yeast genetics and the ease of manipulation of yeast have substantially con-

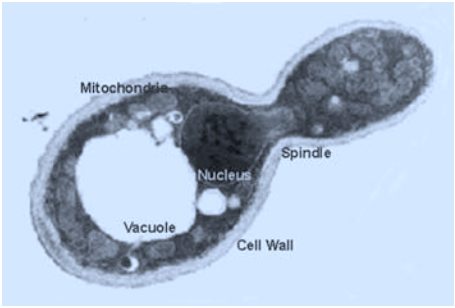


Fig. 1.4 Micrograph of the budding yeast *Saccharomyces cerevisiae* during spore formation. The cell wall, nucleus, vacuole, mitochondria, and spindle are indicated.

tributed to the explosive growth in yeast molecular biology. This success is also a consequence of the notion that the extent to which basic biological processes have been conserved throughout eukaryotic life is remarkable and makes yeast a unique unicellular model organism in which cell architecture and fundamental cellular mechanisms can be successfully investigated. No wonder, then, that yeast had again reached the forefront in experimental molecular biology by being the first eukaryotic organism of which the entire genome sequence became available [44, 45]. The wealth of sequence information obtained in the yeast genome project was found to be extremely useful as a reference against which sequences of human, animal, or plant genes could be compared.

The first genetic map of *S. cerevisiae* was published by Lindegren in 1949 [46]; many revisions and refinements have appeared since. At the outset of the sequencing project approximately 1200 genes had been mapped and detailed biochemical knowledge about a similar number of genes encoding either RNA or protein products had accumulated [47]. The existence of 16 chromosomes ranging in size between 250 and ~2500 kb was firmly established when it became feasible to separate all chromosomes by pulsed-field gel electrophoresis (PFGE). This also provided definition of "electrophoretic karyotypes" of strains by sizing chromosomes [48]. Not only do laboratory strains have different karyotypes, because of chromosome length polymorphisms and chromosomal rearrangements, but so do industrial strains. A defined laboratory strain (αS288C) was therefore chosen for the yeast sequencing project.

1.3.1.2

The Yeast Genome Sequencing Project

The yeast sequencing project was initiated in 1989 within the framework of EU biotechnology programs. It was based on a network approach into which 35 European laboratories initially became involved, and chromosome III – the first eukaryotic chromosome ever to be sequenced – was completed in 1992. In the following years and engaging many more laboratories, sequencing of further complete chromosomes was tackled by the European network. Soon after its beginning, laboratories in other parts of the world joined the project to sequence other chromosomes or parts thereof, ending up in a coordinated international enterprise. Finally, more than 600 scientists in Europe, North America, and Japan became involved in this effort. The sequence of the entire yeast genome was completed in early 1996 and released to public databases in April 1996.

Although the sequencing of chromosome III started from a collection of overlapping plasmid or phage lambda clones, it was expected that cosmid libraries would subsequently have to be used to aid large-scale sequencing [49]. Assuming an average insert length of 35–40 kb, a cosmid library containing 4600 random clones would represent the yeast genome at approximately twelve times the genome equivalent. The

advantages of cloning DNA segments in cosmids were at hand – clones were found to be stable for many years and the small number of clones was advantageous in setting up ordered yeast cosmid libraries or sorting out and mapping chromosome-specific sublibraries. High-resolution physical maps of the chromosomes to be sequenced were constructed by application of classical mapping methods (fingerprints, cross-hybridzation) or by novel methods developed for this program, for example site-specific chromosome fragmentation [50] or a high-resolution cross-hybridization matrix, to facilitate sequencing and assembly of the sequences.

In the European network chromosome-specific clones were distributed to the collaborating laboratories according to a scheme worked out by the DNA coordinators. Each contracting laboratory was free to apply sequencing strategies and techniques of its own provided that the sequences were entirely determined on both strands and unambiguous readings were obtained. Two principle approaches were used to prepare subclones for sequencing:

1. generation of sublibraries by use of a series of appropriate restriction enzymes or from nested deletions of appropriate subfragments made by exonuclease III; and
2. generation of shotgun libraries from whole cosmids or subcloned fragments by random shearing of the DNA.

Sequencing by the Sanger technique was either performed manually, labeling with [^{35}S]dATP being the preferred method of monitoring, or by use of automated devices (on-line detection with fluorescence labeling or direct blotting electrophoresis system) following a variety of established procedures. Similar procedures were applied to the sequencing of the chromosomes contributed by the Sanger laboratory and by laboratories in the USA, Canada, and Japan. The American laboratories largely relied on machine-based sequencing.

Because of their repetitive substructure and the lack of appropriate restriction sites, the yeast chromosome telomeres were a particular problem. Conventional cloning procedures were successful for a few exceptions only. Telomeres were usually physically mapped relative to the terminal-most cosmid inserts using the chromosome fragmentation procedure [50]. The sequences were then determined from specific plasmid clones obtained by "telomere trap cloning", an elegant strategy developed by Louis and Borts [51].

Within the European network, all original sequences were submitted by the collaborating laboratories to the Martinsried Institute of Protein Sequences (MIPS) who acted as an informatics center. They were kept in a data library, assembled into progressively growing contigs, and, in collaboration with the DNA coordinators, the final chromosome sequences were derived. Quality controls were performed by anonymous resequencing of selected regions and suspected or difficult zones (total of 15–20 % per chromosome). Similar procedures were used for sequence assembly and quality control in the other laboratories. During recent years further quality controls were carried and resulted in a nearly absolute accuracy of the total sequence.

The sequences of the chromosomes were subjected to analysis by computer algorithms, identifying ORF and other genetic entities, and monitoring compositional characteristics of the chromosomes (base composition, nucleotide pattern frequencies, GC profiles, ORF distribution profiles, etc.). Because the intron splice site/branchpoint pairs in yeast are highly conserved, they could be detected by using defined search patterns. It was finally found that only 4 % of the yeast genes contain (mostly

short) introns. Centromere and telomere regions, and tRNA genes, sRNA genes, or the retrotransposons, were sought by comparison with previously characterized datasets or appropriate search programs. All putative proteins were annotated by using previously established yeast data and evaluating searches for similarity to entries in the databases or protein signatures detected by using the PROSITE dictionary.

1.3.1.3

Life with Some 6000 Genes

With its 12.8 Mb, the yeast genome is approximately a factor of 250 smaller than the human genome. The complete genome sequence now defines some 6000 ORFs which are likely to encode specific proteins in the yeast cell. A protein-encoding gene is found every 2 kb in the yeast genome, with nearly 70 % of the total sequence consisting of ORF. This leaves only limited space for the intergenic regions which can be thought to harbor the major regulatory elements involved in chromosome maintenance, DNA replication, and transcription. The genes are usually rather evenly distributed among the two strands of the single chromosomes, although arrays longer than eight genes that are transcriptionally oriented in the same direction can be found eventually. With a few exceptions, transcribed genes on complementing strands are not overlapping, and no "genes-in-genes" are observed. Although the intergenic regions between two consecutive ORF are sometimes extremely short, they are normally maintained as separate units and not coupled for transcription. In "head-to-head" gene arrangements the intervals between the divergently transcribed genes might be interpreted to mean that their expression is regulated in a concerted fashion involving the common promoter region. This, however, seems not to be true for most genes and seems to be a principle reserved for a few cases. The sizes of the ORFs vary between 100 to more than 4000 codons; less than 1 % is estimated to be below 100 codons. In addition, the yeast genome contains some 120 ribosomal RNA genes in a large tandem array on chromosome XII, 40 genes encoding small nuclear RNA (sRNA) and 275 tRNA genes (belonging to 43 families) which are scattered throughout the genome. Overall, the yeast genome is remarkably poor in repeated sequences, except the transposable elements (Tys) which account for approximately 2 % of the genome, and, because of to their genetic plasticity, are the major source of polymorphisms between different strains. Finally, the sequences of non-chromosomal elements, for example the 6 kb of the 2μ plasmid DNA, the killer plasmids present in some strains, and the yeast mitochondrial genome (ca.75 kb), must be considered.

On completion of the yeast genome sequence it became possible for the first time to define the proteome of a eukaryotic cell. Detailed information was laid down in inventory databases and most of the proteins could be classified according to function. It was seen that almost 40 % of the proteome consisted of membrane proteins, and that an estimated 8 to 10 % of nuclear genes encode mitochondrial functions. It came as an initial surprise that no function could be attributed to approximately 40 % of the yeast genes. However, even with the exponential growth of entries in protein databases and the refinement of *in silico* analyses, this figure could not be reduced substantially. The same was observed for all other genomes that have since been sequenced. As an explanation, we have to envisage that a considerable portion of every genome is reserved for species- or genus-specific functions.

An interesting observation made for the first time was the occurrence of regional

variations of base composition with similar amplitudes along the chromosomes. Analysis of chromosomes III and XI revealed almost regular periodicity of the GC content, with a succession of GC rich and GC poor segments of ~50 kb each. Another interesting observation was that the compositional periodicity correlated with local gene density. Profiles obtained from similar analyses of chromosomes II and VIII again showed these phenomena, albeit with less pronounced regularity. Similar compositional variation has been found along the arms of other chromosomes, with pericentromeric and subtelomeric regions being AT-rich, though spacing between GC-rich peaks is not always regular. Usually, however, there is a broad correlation between high GC content and high gene density.

Comparison of all yeast sequences revealed there is substantial internal genetic redundancy in the yeast genome, which at the protein level is approximately 40 %. Whereas an estimate of sequence similarity (at both the nucleotide and the amino acid level) is highly predictive, it is still difficult to correlate these values with functional redundancy. Interestingly, the same phenomenon has since been observed in all other genomes sequenced. Gene duplications in yeast are of different type. In many instances, the duplicated sequences are confined to nearly the entire coding region of these genes and do not extend into the intergenic regions. Thus, the corresponding gene products share high similarity in terms of amino acid sequence, and sometimes are even identical, and, therefore, might be functionally redundant. As suggested by sequence differences within the promoter regions or demonstrated experimentally, however, expression varies. It is possible one gene copy is highly expressed whereas another is poorly expressed. Turning on or off expression of a particular copy within a gene family might depend on the differentiated status of the cell (for example mating type, sporulation, etc.). Biochemical studies also revealed that in particular instances "redundant" proteins can substitute each other, thus accounting for the observation that large amounts of single-gene disruption in yeast do not impair growth or cause "abnormal" phenotypes. This does not imply, however, that these "redundant" genes are *a priori* dispensable. Rather they might have arisen from the need to help adapt yeast cells to particular environmental conditions.

Subtelomeric regions in yeast are rich in duplicated genes which are of functional importance to carbohydrate metabolism or cell-wall integrity, but there is also much variety of (single) genes internal to chromosomes that seem to have arisen from duplications. An even more surprising phenomenon became apparent when the sequences of complete chromosomes were compared with each other. This revealed 55 large chromosome segments (up to 170 kb) in which homologous genes are arranged in the same order, with the same relative transcriptional orientations, on two or more chromosomes [52]. The genome has continued to evolve since this ancient duplication occurred – genes have been inserted or deleted, and Ty elements and introns have been lost and gained between two sets of sequences. If optimized for maximum coverage, up to 40 % of the yeast genome is found to be duplicated in clusters, not including Ty elements and subtelomeric regions. No observed clusters overlap and intra- and interchromosomal cluster duplications have similar probabilities.

The availability of the complete yeast genome sequence not only provided further insight into genome organization and evolution in yeast but also offered a reference to search for orthologs in other organisms. Of particular interest were those genes that

are homologs of genes that perform differentiated functions in multicellular organisms or that might be of relevance to malignancy. Comparing the catalog of human sequences available in the databases with the yeast ORF reveals that more than 30 % of yeast genes have homologs among human genes of known function. Approximately 100 yeast genes are significantly similar to human disease genes [53], and some of the latter could even be predicted from comparison with the yeast genes.

1.3.1.4

The Yeast Postgenome Era

It was evident to anyone engaged in the project that determination of the entire sequence of the yeast genome should only be regarded as a prerequisite for functional studies of the many novel genes to be detected. Thus, a European functional analysis network (EUROFAN) was initiated in 1995 and similar activities were started in the international collaborating laboratories in 1996. The general goal was to systematically investigate yeast genes of unknown function by use of the approaches:

1. improved data analysis by computer (*in silico* analysis);
2. systematic gene disruptions and gene overexpression;
3. analysis of phenotypes under different growth conditions, for example temperature, pH, nutrients, stress;
4. systematic transcription analysis by conventional methods; gene expression under different conditions;
5. *in situ* cellular localization and movement of proteins by the use of tagged proteins (GFP-fusions);
6. analysis of gene expression under different conditions by 2D gel-electrophoresis of proteins; and
7. complementation tests with genes from other organisms.

In this context, a most compelling approach is the genome-wide analysis of gene expression profiles by chip technology. High-density micro-arrays of all yeast ORF were the first to be successfully used in studying various aspects of a transcriptome [54, 55]. Comprehensive and regularly updated information can be found in the yeast Genome Database (http://www.yeastgenome.org).

Now the entire sequence of a laboratory strain of *S. cerevisiae* is available, the complete sequences of other yeasts of industrial or medical importance are within our reach. Such knowledge would considerably accelerate the development of productive strains needed in other areas (e.g. *Kluyveromyces*, *Yarrowia*) or the search for novel anti-fungal drugs. It might even be unnecessary to finish the entire genome if a yeast or fungal genome has substantial synteny with that of *S. cerevisiae*. A special program devoted to this problem, analysis of hemiascomycetes yeast genomes by tag-sequence studies for the approach of speciation mechanisms and preparation of tools for functional genomics, has recently been finalized by a French consortium [56].

In all, the yeast genome project has demonstrated that an enterprise like this can successfully be conducted in "small steps" and by teamwork. Clearly, the wealth of fresh and biologically relevant information collected from yeast sequences and from functional analyses has had an impact on other large-scale sequencing projects.

1.3.2

The Plant *Arabidopsis thaliana*

1.3.2.1

The Organism

Arabidopsis thaliana (Fig. 1.5) is a small cruciferous plant of the mustard family first described by the German physician Johan-

Fig. 1.5 The model plant *Arabidopsis thaliana*. (a) Adult plant, approx. height 20 cm [59]; (b) flower, approx. height 4 mm [60]; (c) chromosome plate showing the five chromosomes of *Arabidopsis* [57].

nes Thal in 1577 in his Flora Book of the Harz mountains, Germany, and later named after him. In 1907 *A. thaliana* was recognized to be a versatile tool for genetic studies by Laibach [57] when he was a student with Strasburger in Bonn, Germany. More than 30 years later in 1943 – then Professor of Botany in Frankfurt – he published an influential paper [58] clearly describing the favorable features making this plant a true model organism: (1) short generation time of only two months, (2) high seed yield, (3) small size, (4) simple cultivation, (5) self fertilization, but (6) easy crossing yielding fully fertile hybrids, (7) only five chromosomes in the haploid genome, and (8) the possibility of isolating spontaneous and induced mutants. An attempt by Rédei in the 1960s to convince funding agencies to develop *Arabidosis* as a plant model system was unsuccessful, mainly because geneticists at that time had no access to genes at the molecular level and therefore no reason to work with a plant irrelevant to agriculture.

With the development of molecular biology two further properties of the *A. thaliana* genome made this little weed the superior choice as an experimental system. Laibach had already noted in 1907 that *A. thaliana* contained only one third of the chromatin of related *Brassica* species [57]. Much later it became clear that this weed has: (1) the smallest genome of any higher plant [61] and (2) a small amount of repetitive DNA. Within the plant kingdom characterized by its large variation of genome sizes (see, for example, http://www.rbgkew.org.uk/cval/database1.html), mainly because of different amounts of repetitive DNA, these features support efficient map-based cloning of genes for a detailed elucidation of their function at the molecular level. A first set of tools for that purpose became available during the 1980s with a comprehensive genetic map containing 76 phenotypic markers obtained by mutagenesis, RFLP maps, *Agro-*

bacterium-mediated transformation, and cosmid and yeast artificial chromosome (YAC) libraries covering the genome severalfold with only a few thousand clones.

It was soon realized that projects involving resources shared by many laboratories would profit from centralized collection and distribution of stocks and related information. "The Multinational Coordinated *Arabidopsis thaliana* Genome Research Project" was launched in 1990 and, as a result, two stock centers in Nottingham, UK, and Ohio, USA, and the *Arabidopsis* database at Boston, USA [62] were created in 1991. With regard to seed stocks these centers succeeded an effort already started by Laibach in 1951 [59] and continued by Röbbelen and Kranz. The new, additional collections of clones and clone libraries provided the basis for the genome sequencing project later on. With increased use of the internet at that time the database soon became a central tool for data storage and distribution and has ever since served the community as a central one-stop shopping point for information, despite its move to Stanford and its restructuring to become "The *Arabidopsis* Information Resource" (TAIR; http://www.arabidopsis.org).

In subsequent years many research tools were improved and new ones were added – mutant lines based on insertions of T-DNA and transposable elements were created in large numbers, random cDNA clones were sequenced partially, maps became available for different types of molecular marker such as RAPD, CAPS, microsatellites, AFLP, and SNP which were integrated with each other, recombinant inbred lines were established to facilitate the mapping process, a YAC library with large inserts and BAC and P1 libraries were constructed, physical maps based on cosmids, YAC, and BAC were built, and tools for their display were developed.

1.3.2.2

Structure of the Genome Project

In August 1996 the stage was prepared for large-scale genome sequencing. At a meeting in Washington DC representatives of six research consortia from North America, Europe, and Japan launched the "*Arabidopsis* Genome Initiative". They set the goal of sequencing the complete genome by the year 2004 and agreed on the strategy, the distribution of tasks, and guidelines for the creation and publication of sequence data. The genome of the ecotype Columbia was chosen for sequencing, because all large insert libraries had been prepared from this line and it was one of the most prominent ecotypes for all kinds of experiment worldwide besides Landsberg erecta (Ler). Because Ler is actually a mutant isolated from an inhomogeneous sample of the Landsberg ecotype after X-ray irradiation [63], it was not suitable to serve as a model genome. The sequencing strategy rested on BAC and P1 clones from which DNA can be isolated more efficiently than from YAC and which, on average, contain larger inserts than cosmids. This strategy was chosen even though most attempts to create physical maps had been based on cosmid and YAC clones at that time and that initial sequencing efforts had employed mostly cosmids. BAC and P1 clones for sequencing were chosen *via* hybridization to YAC and molecular markers of known and well separated map positions. Later, information from BAC end sequences and fingerprint and hybridization data, created while genome sequencing was already in progress, were used to minimize redundant sequencing caused by clone overlap. This multinational effort has been very fruitful and led to complete sequences for two of the five *Arabidopsis* chromosomes, chromosomes 2 [64] and 4 [65], except for their rDNA repeats and the heterochromatic regions

around their centromers. The genomic sequence was completed by the end of the year 2000, more than three years ahead of the original timetable. Sequences of the mitochondrial [66] and the plastid genome [67] have also been determined, so complete genetic information was available for *Arabidopsis*.

1.3.2.3

Results from the Genome Project

The sequenced chromosomes have yielded no surprises with regard to their structural organization. With the exception of one sequenced marker, more than one hundred have been observed in the expected order. Repetitive elements and repeats of transposable elements are concentrated in the heterochromatic regions around the centromers where gene density and recombination frequency are below the average (22 genes/100 kbp, 1 cM/50 - 250 kbp). With a few minor exceptions these average values do not vary substantially in other regions of the chromosomes, which is in sharp contrast with larger plant genomes [68–70]. In addition, genomes such as that of maize do contain repetitive elements and transposons interspersed with genes [71], so *Arabidopsis* is certainly not a model for the structure of large plant genomes.

From the sequences available it has been calculated that the 120 Mbp gene containing part of the nuclear genome of *Arabidopsis* contains approximately 25,000 genes (chr. 2: 4037, chr. 4: 3744 annotated genes; [72]), whereas the mitochondrial and plastid genomes carry only 57 and 78 genes on 366,924 and 154,478 bp of DNA, respectively. Most of the organellar proteins must therefore be encoded in the nucleus and are targeted to their final destinations *via* N-terminal transit peptides. It has recently been estimated that 10 or 14 % of the nuclear genes encode proteins located in mitochondria or plastids, respectively [73]. Only for a fraction of the predicted plastid proteins could homologous cyanobacterial proteins be identified [64, 65]; even so, lateral gene transfer from the endosymbiotic organelle to the nucleus has been assumed to be the main source of organellar proteins. These data indicate that either the large evolutionary distance between plants and cyanobacteria prevents the recognition of orthologs and/or many proteins from other sources in the eukaryotic cell have acquired plastid transit peptides, as suggested earlier on the basis of Calvin cycle enzymes [74]. A substantial number of proteins without predicted transit peptides but with higher homology to proteins from cyanobacteria than to any other organism have also been recognized [64, 65], indicating that plastids, at least, have contributed a significant part of the protein complement of other compartments. These data clearly show that plant cells have become highly integrated genetic systems during evolution, assembled from the genetic material of the eukaryotic host and two endosymbionts [75]. That this system integration is an ongoing process is revealed by the many small fragments of plastid origin in the nuclear genome [76] and the unexpected discovery of a recent gene-transfer event from the mitochondrial to the nuclear genome. Within the genetically defined centromer of chromosome 2, a 270-kbp fragment 99 % identical with the mitochondrial genome has been identified and its location confirmed *via* PCR across the junctions with unique nuclear DNA [64]. For a comprehensive list of references on *Arabidopsis thaliana* readers are referred to Schmidt [77]. This contribution cites only articles not listed by Schmidt or which are of utmost importance to the matters discussed here.

1.3.2.4

Follow-up Research in the Postgenome Era

Potential functional assignments can be made for up to 60 % of the predicted proteins on the basis of sequence comparisons, identification of functional domains and motifs, and structural predictions. Interestingly, 65 % of the proteins have no significant homology with proteins of the completely sequenced genomes of bacteria, yeast, *C. elegans*, and *D. melanogaster* [64], clearly reflecting the large evolutionary distance of the plant from other kingdoms and the independent development of multicellularity accompanied by a large increase in gene number. The discovery of protein classes and domains specific for plants, e.g. Ca^{2+}-dependent protein kinases containing four EF-hand domains [65] or the B3 domain of *ABI3*, *VP1*, and *FUS3* [78], and the significantly different abundance of several proteins or protein domains compared with *C. elegans* or *D. melanogaster*, for example myb-like transcription factors [79], C3HC4 ring finger domains, and P450 enzymes [65], further support this notion. Already the larger number of genes in the *Arabidopsis* genome (approx. 25,000) compared with the genomes of *C. elegans* (approx. 19,000) and *D. melanogaster* (approx. 14,000) seems indicative of different ways of evolving organisms of comparable complexity. Currently, the underlying reason for this large difference is unclear. It might simply reflect a larger proportion of duplicated genes, as it does for *C. elegans* compared with *D. melanogaster* [80]. The large number of observed tandem repeats [64, 65] and the large duplicated segments in chromosomes 2 and 4, 2.5 Mbp in total, and in chromosomes 4 and 5, a segment containing 37 genes [65], seem to favor this explanation. But other specific properties of plants, for example autotrophism, non-mobile life, rigid body structure, continuous organ development, successive gametophytic and sporophytic generations, and, compared with animals, different ways of processing information and responding to environmental stimuli and a smaller extent of combinatorial use of proteins, or any combination of these factors, might also contribute to their large number of genes. Functional assignment of proteins is not currently sufficiently advanced to enable us to distinguish between all these possibilities.

The data currently available indicate that the function of many plant genes is different from those of animals and fungi. Besides the basic eukaryotic machinery which was present in the last common ancestor before endosymbiosis with cyanobacteria created the plant kingdom, and which can be delineated by identification of orthologs in eukaryotic genomes [65], all the plant-specific functions must be elucidated. Because more than 40 % of all predicted proteins from the genome of *Arabidopsis* have no assigned function and many others have not been investigated thoroughly, this will require an enormous effort using different approaches. New techniques from chip-based expression analysis and genotyping to high-throughput proteomics and protein–ligand interaction studies and metabolite profiling have to be applied in conjunction with identification of the diversity present in nature or created *via* mutagenesis, transformation, gene knock-out, etc. Because multicellularity was established independently in all kingdoms, it might be wise to sequence the genome of a unicellular plant. Its gene content would help us identify all the proteins required for cell-to-cell communication and transport of signals and metabolites. To collect, store, evaluate, and access all the relevant data from these various approaches, many of which have been performed in parallel and de-

signed for high-throughput analyses, new bioinformatic tools must be created. To coordinate such efforts, a first meeting of "The Multinational Coordinated *Arabidopsis* 2010 Project" (http://www.arabidopsis.org/workshop1.html), with the objective of functional genomics and creation of a virtual plant within the next decade, was held in January 2000 at the Salk Institute in San Diego, USA. From just two examples it is evident that the rapid progress in *Arabidopsis* research will continue in the future. A centralized facility for chip-based expression analysis has been set-up already in Stanford, USA, and a company provides access to 39,000 potential single-nucleotide polymorphisms, which will speed up mapping and genotyping substantially. At the end the question remains the same as in the beginning – why should all these efforts be devoted to a model plant irrelevant to agriculture? The answer can be given again as a question – which other plant system would provide better tools for tackling all the basic questions of plant development and adaptation than *Arabidopsis thaliana*?

1.3.3 The Roundworm *Caenorhabditis elegans*

1.3.3.1 The Organism

The free-living nematode *Caenorhabditis elegans* (Fig. 1.6) is the first multicellular animal whose genome has been completely sequenced. This worm, although often viewed as a featureless tube of cells, has been studied as a model organism for more than 20 years. In the 1970s and 1980s, the complete cell lineage of the worm from fertilized egg to adult was determined by microscopy [81], and later the entire nervous system was reconstructed [82]. It has proved to have several advantages as an object of biological study, for example simple growth conditions, rapid generation time with an invariant cell lineage, and well developed genetic and molecular tools for its manipulation. Many of the discoveries made with *C. elegans* are particularly relevant to the study of higher organisms, because it shares many of the essential biological features, for example neuronal growth, apoptosis, intra- and intercellular signaling pathways, food digestion, etc. that are in the focus of interest of, for example, human biology.

Fig. 1.6 The freeliving bacteriovorous soil nematode *Caenorhabditis elegans* which is a member of the Rhabditidae (found at http://www.nematodes.org).

A special review by the *C. elegans* Genome Consortium [17] gives an interesting summary of how "the worm was won" and examines some of the findings from the near-complete sequence data.

The *C. elegans* genome was deduced from a clone-based physical map to be 100 Mb. This map was initially based on cosmid clones using a fingerprinting approach. Later yeast artificial chromosome (YAC) clones were incorporated to bridge the gaps between cosmid contigs, and provided coverage of regions that were not represented in the cosmid libraries. By 1990, the physical map consisted of fewer than 20 contigs and was useful for rescue experiments that were able to locate a phenotype of interest to a few kilobases of DNA [83]. Alignment of the existing genetic and physical maps into the *C. elegans* genome map

was greatly facilitated by cooperation of the entire "worm community". When the physical map had been nearly completed the effort to sequence the entire genome became both feasible and desirable. By that time this attempt was significantly larger than any sequencing project before and was nearly two orders of magnitude more expensive than the mapping effort.

1.3.3.2

The Structure of the Genome Project

In 1990 a three-year pilot project for sequencing 3 Mb of the genome was initiated as a collaboration between the Genome-Sequencing Center in St Louis, USA and the Sanger Center in Hinxton, UK. Funding was obtained from the NIH and the UK MRC. The genome sequencing effort initially focused on the central regions of the five autosomes which were well represented in cosmid clones, because most genes known at that time were contained in these regions. At the beginning of the project, sequencing was still based on standard radioisotopic methods, with "walking" primers and cosmid clones as templates for the sequencing reactions. A severe problem of this primer-directed sequencing approach on cosmids were, however, multiple priming events because of repetitive sequences and efficient preparation of sufficient template DNA. To address these problems the strategy was changed to a more classic shotgun sequencing strategy based on cosmid subclones generally sequenced from universal priming sites in the subcloning vectors. Further developments in automation of the sequencing reactions, fluorescent sequencing methods, improvements in dye-terminator chemistry [84], and the generation of assembly and contig editing programs, led to the phasing out of the instrument in favor of four-color, single-lane sequencing. The finishing phase then used a more ordered, directed sequencing strategy as well as the walking approach, to close specific remaining gaps and resolve ambiguities. Hence, the worm project grew into a collaboration among *C. elegans* Sequencing Consortium members and the entire international community of *C. elegans* researchers. In addition to the nuclear genome-sequencing effort other researchers sequenced its 15-kb mitochondrial genome and performed extensive cDNA analyses that facilitated gene identification.

The implementation of high-throughput devices and semi-automated methods for DNA purification and sequencing reactions led to overwhelming success in scaling up of the sequencing. The first 1 Mb threshold of *C. elegans* finished genome sequences was reached in May 1993. In August 1993, the total had already increased to over 2 Mb [85], and by December 1994 over 10 Mb of the *C. elegans* genome had been completed. Bioinformatics played an increasing role in the genome project. Software developments made the processing, analysis, and editing of thousands of data files per day a manageable task. Indeed, many of the software tools developed in the *C. elegans* project, for example ACeDB, PHRED, and PHRAP, have become key components in the current approach to sequencing the human genome.

The 50 Mb mark was passed in August 1996. At this point, it became obvious that 20 % of the genome was not covered by cosmid clones, and a closure strategy was implemented. For gaps in the central regions, either long-range PCR was used, or a fosmid library was probed in search of a bridging clone. For the remaining gaps in the central regions, and for regions of chromosomes contained only in YAC, purified YAC DNA was used as the starting material for shotgun sequencing. All of these final regions have been essentially completed with

the exception of several repetitive elements. The final genome sequence of the worm is, therefore, a composite from cosmids, fosmids, YAC, and PCR products. The exact genome size was approximate for a long time, mainly because of repetitive sequences that could not be sequenced in their entirety. Telomeres were sequenced from plasmid clones provided by Wicky et al. [86]. Of twelve chromosome ends, nine have been linked to the outermost YAC on the physical map.

1.3.3.3

Results from the Genome Project

Analysis of the approximately 100 Mb of the total *C. elegans* genome revealed nearly 20,000 predicted genes, considerably more than expected before sequencing began [87, 88], with an average density of one predicted gene per 5 kb. Each gene was estimated to have an average of five introns and 27 % of the genome residing in exons. The number of genes is approximately three times the number found in yeast [89] and approximately one-fifth to one-third the number predicted for the human genome [17].

Interruption of the coding sequence by introns and the relatively low gene density make accurate gene prediction more challenging than in microbial genomes. Valuable bioinformatics tools have been developed and used to identify putative coding regions and to provide an initial overview of gene structure. To refine computer-generated gene structure predictions, the available EST and protein similarities and the genomic sequence data from the related nematode *Caenorhabditis briggsae* were used for verification. Although approximately 60 % of the predicted genes have been confirmed by EST matches [17, 90], recent analyses have revealed that many predicted genes needed corrections in their intron-exon structures [91].

Similarities to known proteins provide an initial glimpse into the possible function of many of the predicted genes. Wilson [17] reported that approximately 42 % of predicted protein products have cross-phylum matches, most of which provide putative functional information [92]. Another 34 % of predicted proteins match other nematode proteins only, a few of which have been functionally characterized. The fraction of genes with informative similarities is far less than the 70 % observed for microbial genomes. This might reflect the smaller proportion of nematode genes devoted to core cellular functions [89], the comparative lack of knowledge of functions involved in building an animal, and the evolutionary divergence of nematodes from other animals extensively studied so far at the molecular level. Interestingly, genes encoding proteins with cross-phylum matches were more likely to have a matching EST (60 %) than those without cross-phylum matches (20 %). This observation suggested that conserved genes are more likely to be highly expressed, perhaps reflecting a bias for "house-keeping" genes among the conserved set. Alternatively, genes lacking confirmatory matches might be more likely to be false predictions, although the analyses did not suggest this.

In addition to the protein-coding genes, the genome contains several hundred genes for noncoding RNA. 659 widely dispersed transfer RNA genes have been identified, and at least 29 tRNA-derived pseudogenes are present. Forty-four percent of the tRNA genes were found on the X chromosome, which represents only 20 % of the total genome. Several other noncoding RNA genes, for example those for splicosomal RNA, occur in dispersed multigene families. Several RNA genes are located in the introns of protein coding genes, which might indicate RNA gene transposition [17]. Other noncoding RNA genes are located in long tandem

repeat regions; the ribosomal RNA genes were found solely in such a region at the end of chromosome I, and the 5S RNA genes occur in a tandem array on chemomosome V.

Extended regions of the genome code for neither proteins nor RNA. Among these are regions that are involved in gene regulation, replication, maintenance, and movement of chromosomes. Furthermore, a significant fraction of the *C. elegans* genome is repetitive and can be classified as either local repeats (e.g. tandem, inverted, and simple sequence repeats) or dispersed repeats. Tandem and inverted repeats amount to 2.7 and 3.6 % of the genome, respectively. These local repeats are distributed non-uniformly throughout the genome relative to genes. Not surprisingly, only a small percentage of tandem repeats are found within the 27 % of the protein coding genes. Conversely, the density of inverted repeats is higher in regions predicted as intergenic [17]. Although local repeat structures are often unique in the genome, other repeats are members of families. Some repeat families have a chromosome-specific bias in representation. Altogether 38 dispersed repeat families have been recognized. Most of these dispersed repeats are associated with transposition in some form [93], and include the previously described transposons of *C. elegans*. In addition to multiple-copy repeat families, a significant number of simple duplications have been observed to involve segments that range from hundreds of bases to tens of kilobases. In one instance a segment of 108 kb containing six genes was duplicated tandemly with only ten nucleotide differences observed between the two copies. In another example, immediately adjacent to the telomere at the left end of chromosome IV, an inverted repeat of 23.5 kb was present, with only eight differences found between the two copies. There are many instances of smaller duplications, often separated by tens of kilobases or more that might contain a coding sequence. This could provide a mechanism for copy divergence and the subsequent formation of new genes [17].

The GC content is more or less uniform at 36 % throughout all chromosomes, unlike human chromosomes that have different isochores [94]. There are no localized centromers, as are found in most other metazoa. Instead, the extensive highly repetitive sequences that are involved in spindle attachment in other organisms might be represented by some of the many tandem repeats found scattered among the genes, particularly on the chromosome arms. Gene density is also uniform across the chromosomes, although some differences are apparent, particularly between the centers of the autosomes, the autosome arms, and the X chromosome.

More striking differences become evident on examination of other features. Both inverted and tandem repeat sequences are more frequent on the autosome arms than in the central regions or on the X chromosome. This abundance of repeats on the arms is probably the reason for difficulties in cosmid cloning and sequence completion in these regions. The fraction of genes with cross-phylum similarities tends to be smaller on the arms as does the fraction of genes with EST matches. Local clusters of genes also seem to be more abundant on the arms.

1.3.3.4 Follow-up Research in the Postgenome Era

Although sequencing of the *C. elegans* genome has essentially been completed, analysis and annotation will continue and – hopefully – be facilitated by further information and better technologies as they become available. The recently completed genome

sequencing of *C. briggsae* enabled comparative analysis of functional sequences of both genomes. The genomes are characterized by striking conservation of structure and function. Thus, many features could be verified in general. Although several specific characteristics had to be corrected for *C. elegans*, however, for example the number of predicted genes, the noncoding RNA, and repetitive sequences [95], it is now possible to describe many interesting features of the *C. elegans* genome, on the basis of analysis of the completed genome sequence.

The observations and findings of the *C. elegans* genome project provide a preliminary glimpse of the biology of metazoan development. There is much left to be uncovered and understood in the sequence. Of primary interest, all of the genes necessary to build a multicellular organism are now essentially available, although their exact boundaries, relationships, and functional roles have to be elucidated more precisely. The basis for a better understanding of the regulation of these genes is also now within our grasp. Furthermore, many of the discoveries made with *C. elegans* are relevant to the study of higher organisms. This extends beyond fundamental cellular processes such as transcription, translation, DNA replication, and cellular metabolism. For these reasons, and because of its intrinsic practical advantages, *C. elegans* has proved to be an invaluable tool for understanding features such as vertebrate neuronal growth and pathfinding, apoptosis and intra- and intercellular signaling pathways.

1.3.4 The Fruitfly *Drosophila melanogaster*

1.3.4.1 The Organism

In 1910, T.H. Morgan started his analysis of the fruit fly *Drosophila melanogaster* and identified the first white-eyed mutant fly strain [96]. Now, less than 100 years later, almost the complete genome of the insect has been sequenced, offering the ultimate opportunity to elucidate processes ranging from the development of an organism to its daily behavior.

Drosophila (Fig. 1.7), a small insect with a short lifespan and very rapid, well-characterized development is one of the best analyzed multicellular organisms. During the last century, more than 1300 genes – mostly based on mutant phenotypes – were genetically identified, cloned, and sequenced. Surprisingly, most were found to have counterparts in other metazoa including even humans. Soon it became evident that not only are the genes conserved among species, but also the functions of the encoded proteins. Processes like the development of limbs, the nervous system, the eyes and the heart, or the presence of circadian rhythms and innate immunity are highly

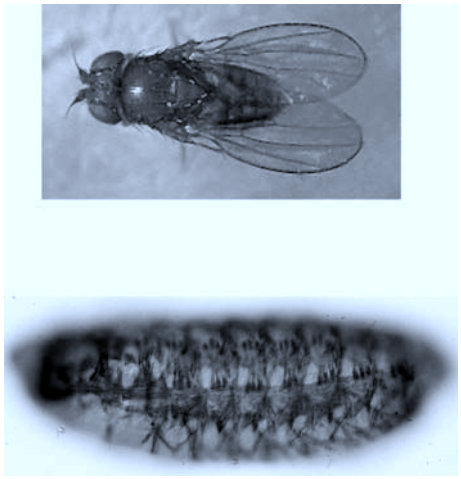


Fig. 1.7 The fruit fly *Drosophila melanogaster*. (a) Adult fly. (b) Stage 16 embryo. In dark-brown, the somatic muscles are visualized by using an β3-tubulin-specific antibody; in red-brown, expression of β1-tubulin in the attachment sites of the somatic muscles is shown.

conserved, even to the extent that genes taken from human sources can supplement the function of genes in the fly [80]. Early in the 1980s identification of the HOX genes led to the discovery that the anterior–posterior body patterning genes are conserved from fly to man [97]. Furthermore, genes essential for the development and differentiation of muscles like twist, MyoD, and MEF2 act in most of the organisms analyzed so far [98]. The most striking example is the capability of a certain class of genes, the PAX-6 family, to induce ectopic eye development in *Drosophila*, irrespective of the animal source from which it was taken [99]. Besides these developmental processes, human disease networks involved in replication, repair, translation, metabolism of drugs and toxins, neural disorders like Alzheimer's, and also higher order functions like memory and signal transduction cascades have also been shown to be highly conserved [97, 100, 101]. It is quite a surprising conclusion that although simply an insect, Drosophila is capable of serving as a model system even for complex human diseases.

The genomic organization of *Drosophila melanogaster* has been known for many years. The fly has four chromosomes, three large and one small, with an early estimate of 1.1×10^8 bp. Using polytene chromosomes as tools, in 1935 and 1938 Bridges published maps of such accuracy they are still used today [96]. Making extensive use of chromosomal rearrangements he constructed cytogenetic maps that assigned genes to specific sections and even specific bands. With the development of techniques such as *in-situ* hybridization to polytene chromosomes, genes could be mapped with a resolution of less than 10 kb. Another major advantage of *Drosophila* is the presence of randomly shuffled chromosomes, the so-called balancers, which enable easy monitoring and pursuit of given mutations on the homologous chromosome and guarantee the persistence of such a mutation at the chromosomal place where it has initially occurred, by suppression of meiotic recombination.

1.3.4.2

Structure of the Genome Project

The strategy chosen for sequencing was the whole-genome-shotgun sequencing (WGS) technique. For this technique, the whole genome is broken into small pieces, subcloned into suitable vectors and sequenced. For the *Drosophila* genome project, libraries containing 2 kb, 10 kb, and 130 kb inserts were chosen as templates. They were sequenced from the ends, and assembled by pairs of reads, called mates, from the ends of the 2 kb and 16 kb inserts [102]. Assembly of the sequences was facilitated by the absence of interspersed repetitive elements like the human ALU-repeat family. The ends of the large BAC clones were taken to unequivocally localize the sequences into large contigs and scaffolds. Three million reads of ~500 bp were used to assemble the *Drosophila* genome. The detailed strategies, techniques, and algorithms used are described in an issue of Science [103] published in March 2000. The work was organized by the Berkeley *Drosophila* Genome Project (BDGP) and performed at the BDGP, the European DGP, the Canadian DGP and at Celera Genomics under the auspices of the federally funded Human Genome Project. The combined DGP produced all the genomic resources and finished 29 Mbp of sequences. In total, the *Drosophila* genome is ~180 Mbp in size, a third of which is heterochromatin. From the ~120 Mbp of euchromatin, 98 % are sequenced with an accuracy of at least 99.5 %. Because of the structure of the heterochromatin, which is mainly built up of re-

petitive elements, retrotransposons, rRNA clusters, and some unique sequences, most of it could not be cloned or propagated in YAC, in contrast with the *C. elegans* genome project [102].

1.3.4.3

Results of the Genome Project

As a major result, the number of genes was calculated to be ~13,600 using a *Drosophila*-optimized version of the program GENIE [100]. Over the last 20 years, 2500 of these have already been characterized by the fly community, and their sequences have been made available continuously during the ongoing of the fly project. The average density is one gene in 9 kb, but ranges from none to nearly 30 genes per 50 kbp, and, in contrast with *C. elegans*, gene-rich regions are not clustered. Regions of high gene density correlate with G + C-rich sequences. Computational comparisons with known vertebrate genes have led to both expected findings and surprises. So genes encoding the basic DNA replication machinery are conserved among eukaryotes; especially, all the proteins known to be involved in recognition of the replication start point are present as single-copy genes, and the ORC3 and ORC6 proteins are closely similar to vertebrate proteins but much different from those in yeast and, apparently, even more different from those in the worm. Focusing on chromosomal proteins, the fly seems to lack orthologs to most of the mammalian proteins associated with centromeric DNA, for example the CENP-C/MIF-2 family. Furthermore, because *Drosophila* telomeres lack the simple repeats characteristic of most eukaryotic telomeres, the known telomerase components are absent. Concerning gene regulation, RNA polymerase subunits and co-factors are more closely related to their mammalian counterparts than to yeast; for example, the promoter interacting factors UBF and TIF-IA are present in *Drosophila* but not in yeast. The overall set of transcription factors in the fly seems to comprise approximately 700 members, about half of which are zinc-finger proteins, whereas in the worm only one-third of 500 factors belong to this family. Nuclear hormone receptors seem to be more rare, because only four new members were detected, bringing the total to 20 compared with more than 200 in *C. elegans*. As an example of metabolic processes, iron pathway components were analyzed. A third ferritin gene has been found that probably encodes a subunit belonging to cytosolic ferritin, the predominant type in vertebrates. Two newly discovered transferrins are homologs of the human melanotransferrin p97, which is of especial interest, because the iron transporters analyzed so far in the fly are involved in antibiotic response rather than in iron transport. Otherwise, proteins homologous to transferrin receptors seem to be absent from the fly, so the melanotransferrin homologs might mediate the main insect pathway for iron uptake (all data taken from Adams et al. [100]). The sequences and the data compiled are freely available to the scientific community on servers in several countries (see: http://www.fruitfly.org). In addition, large collections of EST (expressed sequence tags), cDNA libraries, and genomic resources are available, as are data bases presenting expression patterns of identified genes or enhancer trap lines, for example flyview at the University of Münster (http://pbio07.uni-muenster.de), started by the group of W. Janning. Furthermore, by using more advanced transcription profiling approaches in combination with lower stringency gene prediction programs, approximately 2000 new genes could be detected, according to a recent report [104].

1.3.4.4

Follow-up Research in the Postgenome Era

Comparative analysis of the genomes of *Drosophila melanogaster, Caenorhabditis elegans*, and *Saccharomyces cerevisiae* has been performed [80]. It showed that the core proteome of *Drosophila* consists of 9453 proteins, which is only twice the number for *S. cerevisiae* (4383 proteins). Using stringent criteria, the fly genome is much closer related to the human genome than to that of *C. elegans* one. Interestingly, the fly has orthologs to 177 of 289 human disease genes examined so far and provides an excellent basis for rapid analysis of some basic processes involved in human disease. Furthermore, hitherto unknown counterparts were found for human genes involved in other disorders, for example, *menin*, *tau*, *limb girdle muscular dystrophy type 2B*, *Friedrich ataxia*, and *parkin*. Of the cancer genes surveyed, at least 68 % seem to have *Drosophila* orthologs, and even a p53-like tumor-suppressor protein was detected. All of these fly genes are present as single copies and can be genetically analyzed without uncertainty about additional functionally redundant copies. Hence, all the powerful *Drosophila* tools such as genetic analysis, exploitation of developmental expression patterns, loss-of-function, and gain-of-function phenotypes can be used to elucidate the function of human genes that are not yet understood. A very recent discovery, the function of a new class of regulatory RNA termed microRNA, has rapidly attracted attention to the non-coding parts of the genome [105]. Screens for miRNA targets in *Drosophila* [106] have already led to identification of a large collection of genes and will certainly be very helpful in elucidating the role of the miRNA during development. Furthermore, genome-wide protein–protein-interaction studies using the yeast two-hybrid system have been performed [107] and led to a draft map of about 20,000 interactions between more than 7,000 proteins. These results must certainly be refined but clearly emphasize the importance of genome-wide screening of both proteins and RNA in combination with elaborated computational methods.

1.4 Conclusions

This book chapter summarizes the genome projects of selected model organisms with completed, or almost completed, genomic sequences. The organisms represent the major phylogenetic lineages, the eubacteria, archaea, fungi, plants, and animals, thus covering unicellular prokaryotes with singular, circular chromosomes, and uni- and multicellular eukaryotes with multiple linear chromosomes. Their genome sizes range from ~2 to ~180 Mbp with estimated gene numbers ranging from 2000 to 25,000 (Tab. 1.3).

The organization of the genome of each organism has its own specific and often unexpected characteristics. In *B. subtilis*, for example, a strong bias in the polarity of transcription of the genes with regard to the replication fork was observed, whereas in *E. coli* and *S. cerevisiae* the genes are more or less equally distributed on both strands. Furthermore, although insertion sequences are widely distributed in bacteria, none was found in *B. subtilis*. In *A. thaliana*, an unexpectedly high percentage of proteins showed no significant homology with proteins of organisms outside the plant kingdom, and thus are obviously specific to plants. Another interesting observation resulting from the genome projects concerns gene density. In prokaryotic organisms, i.e.

Table 1.3 Summary of information on the genomes of model organisms described in this chapter (status April 2004).

Organism	*Genome structure*	*Genome size (kb)*	*Estimated no. of genes/ORF*
Escherichia coli	1 chromosome, circular	4600	4400
Bacillus subtilis	1 chromosome, circular	4200	4100
Archaeoglobus fulgidus	1 chromosome, circular	2200	2400
Saccharomyces cerevisiae	16 chromosomes, linear	12,800	6200
Arabidopsis thaliana	5 chromosomes, linear	130,000	25,000
Caenorhabditis elegans	6 chromosomes, linear	97,000	19,000
Drosophila melanogaster	4 chromosomes, linear	180,000	16,000

eubacteria and archaea, genome sizes vary substantially. Their gene density is, however, relatively constant at approximately one gene per kbp. During evolution of eukaryotic organisms, genome sizes grew, but gene density decreased from one gene in two kbp in *Saccharomyces* to one gene in 10 kbp in *Drosophila*. This led to the surprising observation that some bacterial species can have more genes than lower eukaryotes and that the number of genes in *Drosophila* is only approximately three times higher than the number in *E. coli*.

Another interesting question concerns the general homology of genes or gene products between model organisms. Comparison of protein sequences has shown that some gene products can be found in a wide variety of organisms.

Thus, comparative analysis of predicted protein sequences encoded by the genomes of *C. elegans* and *S. cerevisiae* suggested that most of the core biological functions are conducted by orthologous proteins that occur in comparable numbers in these organisms. Furthermore, comparing the yeast genome with the catalog of human sequences available in the databases revealed that a significant number of yeast genes have homologs among human genes of unknown function. *Drosophila* is of special importance in this respect, because many human disease networks have been shown to be highly conserved in the fruitfly. Hence, the insect is, in an outstanding manner, capable of serving as a model system even for complex human diseases. Because the respective genes in the fly are present as single copies, they can be genetically analyzed much more easily.

Genome research of model organisms has just begun. At the time when this article was written, more than 100 genome sequences, mainly from prokaryotes, have been published and more than 200 genomes are currently being sequenced worldwide. The full extent of this broad and rapidly expanding field of genome research on model organisms cannot be covered by a single book chapter. The World-Wide Web, however, is an ideal platform for making available the outcome of large genome projects in an appropriate and timely manner. Beside several specialized entry points, two web pages are recommended as an introduction with a broad overview of model-organism research – the WWW Virtual Library: Model Organisms (http://ceolas.org/VL/mo) and the WWW Resources for Model Organisms (http://genome.cbs.dtu.dk/gorm/modelorganism.html). Both links are excellent starting sites for detailed information on model-organism research.

References

1 Berg, P., Baltimore, D., Brenner, S., Roblin, R.O. 3rd, Singer, M.F. (1975), Asilomar conference on recombinant DNA molecules. *Science* 188, 991–994.

2 Blattner, F.R., Plunkett, G. III, Bloch, C.A., Perna, N.T., Burland, V. et al. (1997), The complete genome sequence of *Escherichia coli* K-12, *Science* 277, 1453–1474.

3 Yamamoto, Y., Aiba, H., Baba, T., Hayashi, K., Inada, T. et al. (1997), Construction of a contiguous 874-kb sequence of the *Escherichia coli* -K12 genome corresponding to 50.0–68.8 min on the linkage map and analysis of its sequence features. *DNA Res.* 4, 91–113.

4 Kröger, M., Wahl, R. (1998) Compilation of DNA sequences of *Escherichia coli* K12: description of the interactive databases ECD and ECDC, *Nucl. Acids Res.* 26, 46–49.

5 Thomas, G.H. (1999), Completing the *E. coli* proteome: a database of gene products characterised since completion of the genome sequence, *Bioinformatics* 15, 860–861.

6 Taylor, A.L., Thoman, M.S. (1964) The genetic map of Escherichia coli K-12. *Genetics* 50, 659–677.

7 Bachmann, B.J. (1983) Linkage map of *Escherichia coli* K-12, edition 7. *Microbiol. Rev.* 47, 180 –230.

8 Neidhard, F.C. (1996) *Escherichia coli and Salmonella typhimurium – Cellular and Molecular Biology*, 2nd edn. American Society for Microbiology, Washington DC.

9 Kohara, Y., Akiyama, K., Isono, K. (1987), The physical map of the whole *E. coli* chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* 50, 495–508.

10 Roberts, R. (2000) Database issue of Nucleic Acids Research, *Nucl. Acids Res.* 28, 1–382.

11 Datsenko, K.A., Wanner, B.L. (2000), One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *PNAS* 97, 6640–6645.

12 Karp, P.D., Riley, M.R., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., Gama-Castro, S. (2002), The EcoCyc Database. *Nucl. Acids Res.* 30, 56–58.

13 Sekowska, A. (1999), Une rencontre du métabolisme du soufre et de l'azote: le métabolisme des polyamines chez *Bacillus subtilis*, thesis in *Génétique cellulaire et moléculaire*. Versailles-Saint-Quentin: Versailles, France. p. 202.

14 Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G. et al. (1997), The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249–256.

15 Moreno, M.S., Schneider, B.L., Maile, R.R., Weyler, W., Saier, M.H., Jr. (2001), Catabolite repression mediated by the CcpA protein in *Bacillus subtilis*: novel modes of regulation revealed by whole-genome analyses. *Mol. Microbiol.* 39, 1366–1381.

16 Fabret, C., Quentin, Y., Guiseppi, A., Busuttil, J., Haiech, J., Denizot, F. (1995), Analysis of errors in finished DNA sequences: the surfactin operon of *Bacillus subtilis* as an example. *Microbiology* 141, 345–350.

17 Wilson, R.K. (1999), How the worm was won. The *C. elegans* genome sequencing project. *Trends Genet.* 15, 51–58.

18 Arias, R.S., Sagardoy, M.A., van Vuurde, J.W. (1999), Spatio-temporal distribution of naturally occurring *Bacillus* spp. and other bacteria on the phylloplane of soybean under field conditions. *J. Basic Microbiol.* 39, 283–292.

19 Rocha, E.P. (2003), DNA repeats lead to the accelerated loss of gene order in bacteria. *Trends Genet*. 19, 600–603.

20 Rocha, E.P., Danchin, A., Viari, A. (1999), Universal replication biases in bacteria. *Mol. Microbiol*. 32, 11–16.

21 Schofield, M.J. and Hsieh, P. (2003), DNA mismatch repair: molecular mechanisms and biological function. *Annu. Rev. Microbiol*. 57, 579–608.

22 Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A. et al. (2001), Comparative genomics of Listeria species. *Science* 294, 849–852.

23 Rocha, E.P., Sekowska, A., Danchin, A. (2000), Sulphur islands in the *Escherichia coli* genome: markers of the cell's architecture? FEBS Lett. 476, 8–11.

24 Danchin, A., Guerdoux-Jamet, P., Moszer, I., Nitschke, P. (2000), Mapping the bacterial cell architecture into the chromosome. Philos. *Trans. R. Soc. Lond. B Biol. Sci*. 355, 179–190.

25 Gottesman, S., Roche, E., Zhou, Y., Sauer, R.T. (1998), The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. *Genes Dev*. 12, 1338–1347.

26 Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K. et al. (2003), Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* 100, 4678–4683.

27 Soma, A., Ikeuchi, Y., Kanemasa, S., Kobayashi, K., Ogasawara, N. et al. (2003), An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. *Mol. Cell*. 12, 689–698.

28 Rocha, E.P. and Danchin, A. (2003), Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res*. 31, 6570–6577.

29 Beckering, C.L., Steil, L., Weber, M.H., Volker, U., Marahiel, M.A. (2002), Genomewide transcriptional analysis of the cold shock response in *Bacillus subtilis*. *J. Bacteriol*. 184, 6395–6402.

30 Hoffmann, T., Schutz, A., Brosius, M., Volker, A., Volker, U., Bremer, E. (2002), High-salinity-induced iron limitation in *Bacillus subtilis*. *J. Bacteriol*. 184, 718–727.

31 Danchin, A. (1997), Comparison between the *Escherichia coli* and *Bacillus subtilis* genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP. *DNA Res*. 4, 9–18.

32 Klenk, H.-P., Clayton, R.A., Tomb, J.-F., White, O., Nelson, K.E. et al. (1997), The complete genome of the hyperthermophilic, sulfate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390, 364–370.

33 Stetter, K.O. (1988), *Archaeoglobus fulgidus* gen. nov., sp. nov.: a new taxon of extremely thermophilic archaebacteria. *System. Appl. Microbiol*. 10, 172–173.

34 Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G. et al. (1997), The complete genome sequence of the gastric pathogen *Heliobacter pylori*. *Nature* 388, 539–547.

35 Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D. et al. (1996), Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*, *Science* 273, 1058–1073.

36 Rabus, R., Ruepp, A., Frickey, T., Rattei, T., Fartmann, B. et al. (2004), The genome of Desulfotalea psychrophila, a sulphate-reducing bacterium from permanently cold Arctic sediments. *Environ. Microbiol*., in press.

37 Vorholt, J., Kunow, J., Stetter, K.O., Thauer, R.K. (1995), Enzymes and coenzymes of the carbon monoxide dehydrogenase pathway for autotrophic CO_2 fixation in *Archaeoglobus lithotrophicus* and the lack of carbon monoxide dehydrogenase in the heterotrophic *A. profundus*. *Arch. Microbiol*. 163, 112–118.

38 Stetter KO, Lauerer G, Thomm M, Neuner A (1987) Isolation of extremely thermophilic sulfate reducers: evidence for a novel branch of archaebacteria. *Science* 236, 822–824.

39 Hosfield, D.J., Frank, G., Weng, Y., Tainer, J.A., Shen, B. (1998), Newly discovered archaebacterial flap endonucleases show a structure-specific mechanism for DNA substrate binding and catalysis resembling human flap endonuclease-1. *J. Biol. Chem*. 273, 27154–27161.

40 Lyamichev, V., Mast, A.L., Hall, J.G., Prudent, J.R., Kaiser, M.W. et al. (1999), Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat. Biotechnol*. 17, 292–296.

41 Cooksey, R.C., Holloway, B.P., Oldenburg, M.C., Listenbee, S., Miller, C.W. (2000), Evaluation of the invader assay, a linear signal amplification method, fro identification of mutations associated with resistance of rifampin and isoniazid in *Mycobacterium*

tuberculosis, *Antimicrob. Agents Chemother.* 44, 1296–1301.

42 Tang, T.H., Bachellerie, J.P. Rozhdestvensky, T., Bortolin, M.L., Huber, H. et al., (2003), Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus, *Proc. Natl. Acad. Sci. USA* 99, 7536–7541.

43 Maisnier-Patin, S., Malandrin, L., Birkeland, N.K., Bernander, R. (2002), Chromosome replication patterns in the hyperthermophilic euryarchaea *Archaeoglobus fulgidus* and *Methanocaldococcus (Methanococcus) janashii*, *Mol. Microbiol.* 45, 1443–1450.

44 Goffeau, A. Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B. et al. (1996), Life with 6000 genes, *Science* 274, 546–567.

45 Anonymous (1997), Dictionary of the yeast genome, *Nature* 387 (suppl.).

46 Lindegren, C.C. (1949) The Yeast Cell, its Genetics and Cytology. Educational Publishers, St Louis, MI.

47 Mortimer, R.K., Contopoulou, R., King, J.S. (1992), Genetic and physical maps of *Saccharomyces cerevisiæ*, Edition 11. *Yeast* 8, 817–902.

48 Carle, G.F., Olson, M.V. (1985), An electrophoretic karyotype for yeast. *Proc. Natl. Acad. Sci. USA* 82, 3756.

49 Stucka, R., Feldmann, H. (1994), Cosmid cloning of Yeast DNA, in: *Molecular Genetics of Yeast – A Practical Approach* (Johnston, J. Ed.) Oxford Univ. Press, pp. 49–64.

50 Thierry, A., Dujon, B. (1992), Nested chromosomal fragmentation in yeast using the meganuclease *I-Sce* I: a new method for physical mapping of eukaryotic genomes. *Nucleic Acids Res.* 20, 5625–5631.

51 Louis, E.J., Borts, R.H. (1995), A complete set of marked telomeres in *Saccharomyces cerevisiae* for physical mapping and cloning. *Genetics* 139, 125–136.

52 Wolfe, K.H., Shields, D.C. (1997), Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713.

53 Foury, F., Roganti, T., Lecrenier, N., Pumelle, B. (1998), Yeast genes and human disease. *FEBS Lett.* 440, 325–331.

54 DeRisi, J.L., Iyer, V.R., Brown, P.O. (1997), Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686.

55 Goffeau, A. (2000), Four years of postgenomic life with 6,000 yeast genes. *FEBS Lett.* 480, 37–41.

56 Feldmann, H. (ed.) (2000), Genolevures: Genomic exploration of the hemiascomycetous yeasts. *FEBS Lett.* 487, no. 1 (http://cbi.labri.fr/genolevures).

57 Laibach, F. (1907). Zur Frage nach der Individualität der Chromosomen im Pflanzenreich. *Beih. Bot. Centralblatt, Abt. I*, 22, 191–210.

58 Laibach, F. (1943) *Arabidopsis thaliana* (L.) Heynh. als Objekt für genetische und entwicklungsphysiologische Untersuchungen. *Bot. Arch.* 44, 439–455.

59 Laibach, F. (1951) Über sommer- und winterannuelle Rassen von *Arabidopsis thaliana* (L.) Heynh. Ein Beitrag zur Atiologie der Blütenbildung, *Beitr. Biol. Pflanzen* 28, 173–210.

60 Müller, A. (1961) Zur Charakterisierung der Blüten und Infloreszenzen von *Arabidopsis thaliana* (L.) Heynh. *Die Kulturpflanze* 9, 364–393.

61 Arumuganathan, K., Earle E.D. (1991), Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* 9, 208–218.

62 Cherry, J.M., Cartinhour, S.W., Goodmann, H.M. (1992), AAtDB, an *Arabidopsis thaliana* database. *Plant Mol. Biol. Rep.* 10, 308–309, 409–410.

63 Rédei, G.P. (1992), A heuristic glance at the past of *Arabidopsis* genetics, in: *Methods in Arabidopsis Research* (Koncz, C., Chua, N.-H., Schell, J. Eds.), World Scientific Publishing, Singapore, pp. 1–15.

64 Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.I. et al. (1999), Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 402, 761–768.

65 Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G. et al. (1999), Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* 402, 769–777.

66 Unseld, M., Marienfeld, J.R., Brandt, P., Brennicke, A. (1997), The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature Genet.* 15, 57–61.

67 Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Tabata, S. (1999), Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* 6, 283–290.

68 Gill, K.S., Gill, B.S., Endo, T.R., Taylor, T. (1996a), Identification and high density mapping of gene-rich regions in the chromosome group *5* of wheat. *Genetics* 143, 1001–1012.

69 Gill, K.S., Gill, B.S., Endo, T.R., Taylor, T. (1996b), Identification and high density mapping of gene–rich regions in the chromosome group *1* of wheat. *Genetics* 144, 1883–1891.

70 Künzel, G., Korzun, L., Meister, A. (2000), Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* 154, 397–412.

71 SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., Bennetzen, J.L. (1996), Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274 765–768.

72 Meyerowitz, E.M. (2000), Today we have the naming of the parts. *Nature* 402, 731–732.

73 Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G. (2000), Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016.

74 Martin, W., Schnarrenberger, C. (1997), The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Curr. Genet.* 32, 1–8.

75 Herrmann, R.G. (2000), Organelle genetics – part of the integrated plant genome. *Vortr. Pflanzenzüchtg.* 48, 279–296.

76 Bevan, M., Bennetzen, J.L., Martienssen, R. (1998), Genome studies and molecular evolution. Commonalities, contrasts, continuity and change in plant genomes. *Curr Opin Plant Biol.* 101–102.

77 Schmidt, R. (2000), The *Arabidopsis* genome. *Vortr. Pflanzenzüchtg.* 48, 228–237.

78 Suzuki, M., Kao, C.Y., McCarty, D.R. (1997), The conserved B3 domain of VIVIPAROUS1 has a cooperative DNA binding activity. *Plant Cell* 9, 799–807.

79 Jin, H., Martin, C. (1999), Multifunctionality and diversity within the plant MYB-gene family. *Plant Mol. Biol.* 41, 577–585.

80 Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R. et al. (2000), Comparative genomics of eukaryotes. *Science* 287, 2204–2215.

81 Sulston, J. (1988), Cell Lineage; In: *The Nematode Caenorhabditis elegans* (Wood, W.B. Ed.) pp. 123–155, CSHL Press

82 Chalfie, M., White, J. (1988), The Nervous system, In: *The Nematode Caenorhabditis elegans* (Wood, W.B. Ed.) pp. 337–391, CSHL Press.

83 Coulson, A., Waterston, R., Kiff, J., Sulston, J., Kohara, Y. (1988), Genome linking with yeast artificial chromosomes. *Nature* 335, 184–186.

84 Lee, L.G., Connell, C.R., Woo, S.L., Cheng, R.D., McArdle, B.F., Fuller, C.W., Halloran, N.D., Wilson, R.K. (1992), DNA sequencing with dye-labeled terminators and T7 DNA polymerase: Effect of dyes and dNTPs on incorporation of dye-terminators, and probability analysis of termination fragments. *Nucleic Acids Res.* 20, 2471–2483.

85 Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M. et al. (1994), 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans. Nature* 368, 32–38.

86 Wicky, C., Villeneuve, A.M., Lauper, N., Codourey, L., Tobler, H., Muller, F. (1996), Telomeric repeats (TTAGGC) are sufficient for chromosome capping in *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA* 93, 8983–8988.

87 Herman, R.K. (1988), Genetics, in: *The Nematode Caenorhabditis elegans* (Wood, W.B. Ed.) pp. 17–45, CSHL Press.

88 Waterston, R., Sulston, J. (1995), The genome of *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA* 92, 10836–10840.

89 Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V. et al. (1998), Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282, 2022–2027.

90 Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., Spieth, J. (2001), WormBase: network access to the genome and biology of *Caenorhabditis elegans. Nucleic Acids Res.* 29, 82–86.

91 Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M. et al. (2003), C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genet*. 34, 35–41.

92 Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., Claverie, J.M. (1993), Ancient conserved regions in new gene sequences and the protein database. *Science* 259, 1711–1716.

93 Smit, A.F. (1996), The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev*. 6, 743–748.

94 Bernardi, G. (1995), The human genome: organization and evolutionary history. *Annu. Rev. Genet*. 29, 445–476.

95 Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R. et al. (2003), The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLOS Biology*, 1, 166–192.

96 Rubin, G.M., Lewis, E.B. (2000), A brief history of *Drosophila´s* contributions to genome research. *Science* 287, 2216–2218.

97 Veraksa, A., Del Campo, M., McGinnis, W. (2000), Developmental patterning genes and their conserved functions: From model organisms to humans. *Mol. Genet. Metab*. 69, 85–100.

98 Zhang, J.M., Chen, L., Krause, M., Fire, A., Paterson, B.M. (1999), Evolutionary conservation of MyoD function and differential utilization of E proteins. *Dev. Biol*. 208, 465–472.

99 Halder, G., Callaerts, P., Gehring, W. J. (1995), Induction of ectopic eyes by targeted expression of the *eyeless* gene in *Drosophila*. *Science* 267, 1788–1792.

100 Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D. et al., (2000), The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195.

101 Mayford, M., Kandel, E.R. (1999), Genetic approaches to memory storage. *Trends Genet*. 15, 463–470.

102 Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P. et al. (2000), A whole-genome assembly of *Drosophila*. *Science* 287, 2196–2204.

103 Anonymous (2000) The Drosophila Genome, *Science* 287, issue 5461.

104 Hild, M., Beckmann, B., Haas, S.A., Koch, B., Solovyev, V. et al. (2003), An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol*. 5, R3.

105 Lai Eric C. (2003), microRNAs: Runts of the genome assert themselves. *Curr. Biol*. 13, R925–R936.

106 Stark, A., Brennecke, J., Russell, R.B., Cohen, S.M. (2003), Identification of *Drosophila* microRNA targets. *PLOS Biology*, 1, 397–409.

107 Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B. et al. (2003), A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736.

2
Environmental Genomics: A Novel Tool for Study of Uncultivated Microorganisms

Alexander H. Treusch and Christa Schleper

Abstract

The vast majority of microbial species has not been cultivated in the laboratory to date and is, therefore, not easily amenable to physiological and genomic characterization. A novel means of addressing this problem is the environmental genomic or metagenomic approach in which whole-community DNA is extracted from environmental samples, purified from contaminating substances, and cloned into large DNA-libraries. These libraries can be screened for genome fragments of specific lineages or can be analyzed in large-scale sequencing approaches to gain insights into genome structure and gene content of specific lineages or even of whole microbial communities. Here we summarize the methodology and the information gained so far from analyses of genome fragments of uncultivated microorganisms, alongside biotechnological aspects of environmental genomics.

2.1
Introduction: Why Novel Approaches to Study Microbial Genomes?

The analysis of prokaryotic genomes has proven to be a powerful tool for characterizing microorganisms. It paves the way for prediction of physiological pathways and fuels studies on genome structure, horizontal gene transfer, and gene regulation. Currently there are more than 160 fully sequenced prokaryotic genomes with more than 180 in progress [1]. All the organisms targeted with these genome projects can be grown in pure cultures or, for intracellular symbionts, can be physically separated from their environment. The approximately 5000 described species [2], however, represent only a minor fraction of the prokaryotic diversity found in natural environments [3]. On the basis of amplification of 16S rRNA genes directly from environmental samples, it has be shown that the actual microbial diversity is much larger. Sequence determination of the cloned 16S rDNA products and reconstruction of phylogenetic trees resulted in a large and comprehensive dataset [4]. Only half of today's defined 52 bacterial phyla have cultivated representatives, the others are represented

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

solely by 16S rDNA sequences from environmental surveys [5]. Although, an impressive 112,000 different bacterial rDNA genes are currently found in the public databases [6], these still reflect only a minor fraction of the existing diversity. Estimates of the number of prokaryotic species living on the planet range between 1.5 and 14 million [7], with as many as 500,000 living in a single soil sample [8].

Cultivation techniques traditionally used in microbiology for isolation of novel species have been selective and mostly favored fast-growing organisms [9, 10]. To overcome such biases, novel approaches have recently been developed that are based on the use of low-nutrition media or dilute media [11–13] or on the simulation of the natural environment, abolishing classical Petri dishes or liquid cultures [14, 15]. Although these novel cultivation approaches have proven successful and have led to the isolation of several novel species even from phyla that have solely been predicted in 16S rRNA surveys, they will not lead to a sufficient analysis of naturally occurring microorganisms. How can we access the physiology of microorganisms that cannot be cultured in the laboratory but that play potentially important roles in the environment? Is it possible to study complete naturally occurring microbial communities and the interaction between different species in specific environments? How can we comprehensively monitor shifts in microbial communities that occur when environmental factors are changed? What is the extent of genomic variation and microheterogeneity in one microbial species in its environment? Can we exploit the huge number of uncultivated microorganisms for biotechnological applications?

Recent advances in environmental genomic studies demonstrate that this novel approach has an enormous potential to address most of these questions, circumventing the need for cultivation of the organisms under investigation. This approach can, furthermore, serve as a basis for functional genomic studies of naturally occurring microbial communities.

2.2 Environmental Genomics: The Methodology

Inspired by rapid advances in genomic analysis of cultivated microorganisms, DeLong and collaborators were the first to apply genomic approaches to uncultivated microorganisms by studying marine planktonic and symbiotic crenarchaeota [16–18]. The environmental genomic approach depends on isolation of DNA directly from the ecosystem. This DNA is cloned into *Escherichia coli* vectors, by procedures quite similar to those used in genome projects of cultivated organisms. The major challenge of the environmental genomic approach is the purification of the DNA from contaminating substances that prevent following molecular biological methods. This is particularly true for soils, which contain large and varying amounts of humic substances like humic and fulvic acids; these co-purify with DNA and severely inhibit procedures used in molecular biology [19]. Humic substances are difficult to separate from DNA, and special procedures had to be developed before DNA from soil samples could be used for environmental genomic studies, in particular when high-molecular-weight DNA was extracted to construct large-insert libraries [20].

The second challenge is the production of large DNA libraries, that must be much more complex than those of single organisms, and identification of genome fragments of distinct organisms therein. If large-insert libraries are used, this can be

done by searching for phylogenetic "marker" genes, like those encoding 16S rRNA, 23S rRNA, DNA-polymerase, RNA-polymerase, etc. [18, 21]. The identification of one of those genes on a large genomic fragment in the environmental library enables the association of this fragment and its encoded genes with one specific lineage based on the marker gene phylogeny. The growing number of sequenced genomes from cultivated species is of great help in the identification of suitable marker genes for this purpose. Another approach that has recently been initiated involves large-scale shotgun sequencing from environmental libraries and assembly of the random sequence reads into partial or even complete genomes.

The approaches used in environmental genomics are summarized in Fig. 2.1 – DNA from an environmental sample is isolated and purified from contaminating substances. Depending on the type of study, different sizes of DNA-fragments are cloned. For large fragments, cosmid, fos-

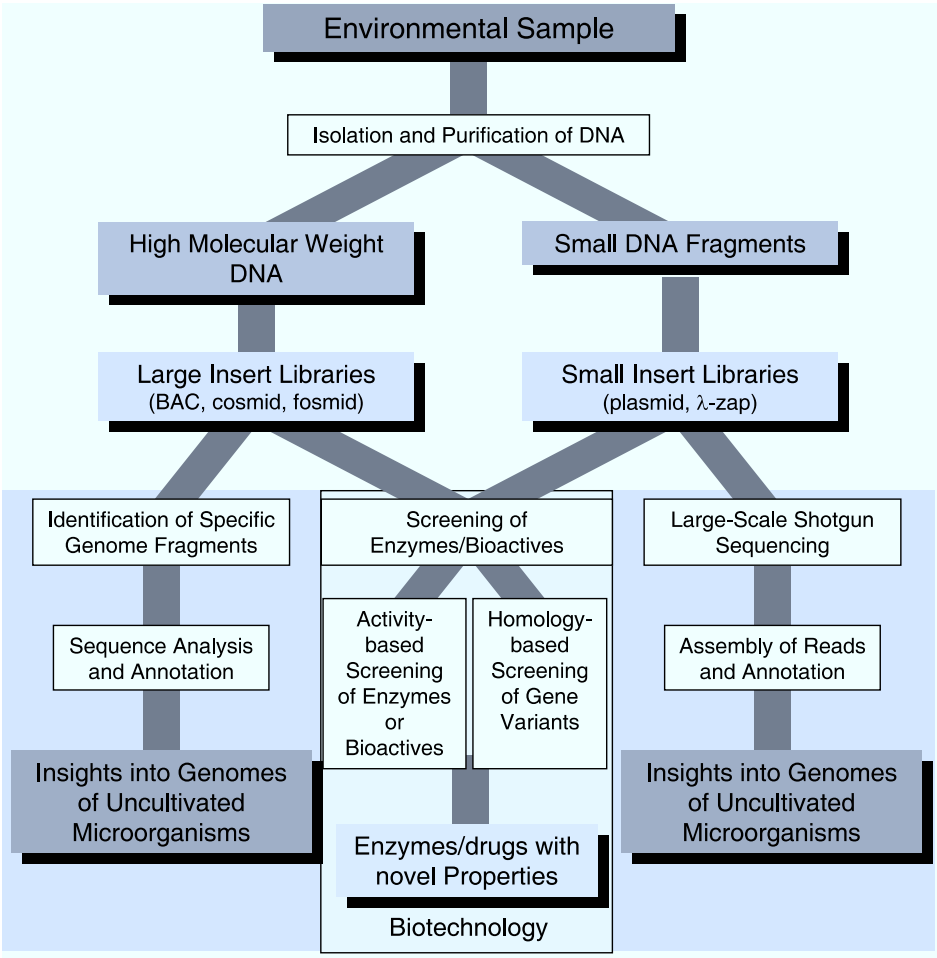


Fig. 2.1 Flow-chart of different techniques used in environmental genomic studies.

mid, and BAC vectors are used to generate large insert libraries (>30 kb insert size). The libraries are screened for genome fragments of uncultivated microorganisms by PCR or hybridization using primers and probes specific for phylogenetic marker genes. Alternatively, random sequences can be created from the inserts for the identification of phylogenetic marker genes in the large-insert libraries [22]. Either way, the sequences of the identified genome fragments of specifically targeted organisms can subsequently be analyzed and annotated.

Alternatively, small fragments (<10 kb) are cloned into plasmids and are used for large-scale shotgun-sequencing approaches [23, 24]. This novel approach recently proved to be a powerful tool, but the major drawback is the need to assemble the sequences in contigs and scaffolds. In contrast with the analysis of genomes from cultivated species this is a difficult task, not only because of the enormous number of different genomes present in the libraries, but also because of the genomic microheterogeneity that occurs in populations of single species. Confidence in these sequences will always be less than in those generated from larger genomic fragments.

Large and small insert libraries (often also referred to as "metagenomic" libraries [25]) are also very useful for biotechnological applications. They can be screened for novel enzymes or drugs in two different ways. First, activity-based screening in expression libraries can lead to identification of novel gene products. This technique relies on the expression of the environmental genes via strong promoters that are located in the vector of the surrogate host organism or via endogenous promoters encoded on the environmental sequences. Second, homology-based screening is used; in this degenerate primers target conserved regions of the desired enzyme gene for screening by PCR. After identification, those genes have to be expressed to prove their activity. The major drawback of this method, in contrast with activity-based screening, is that "only" variants of known enzymes and not completely novel ones can be found. Although for many biotechnological applications small insert libraries are often sufficient to search biosynthetic clusters of secondary metabolites, for example, large insert libraries are needed.

There are, of course, many different possible applications for the environmental genomic approach. Some examples are summarized in the following sections.

2.3 Where it First Started: Marine Environmental Genomics

The environmental genomic approach was first applied to marine ecosystems. Stein et al. [16] identified a 38.5 kb genomic fragment of an uncultivated mesophilic crenarchaeote within a 3552 clones containing fosmid library (average insert size 40 kb) from marine plankton. It was the first genomic information obtained from this group of Archaea, which had been predicted only in 16S rRNA surveys. The same technology was used to obtain information about the uncultivated crenarchaeal symbiont *Cenarchaeum symbiosum*, which was discovered in the marine sponge *Axinella mexicana* [26]. *C. symbiosum* DNA-containing clones were identified in a mixed metagenomic library using the 16S rDNA as marker. The analyzed sequences gave first insights into the genomic structure and protein-encoding genes of this organism and into the genomic microheterogeneity of this species [18]. A gene encoding a B-family DNA polymerase was expressed in *E. coli*

and its activity at moderate temperatures proved adaptation of *C. symbiosum* to the mesophilic growth temperatures of its host. This gave an important hint of the existence of mesophilic crenarchaeota whose cultivated relatives are exclusively extreme thermophiles [17]. The whole genome of *C. symbiosum* is currently being determined [27].

Environmental genomic studies showed their full potential for answering ecological questions when Beja et al. [28] discovered a new type of rhodopsin, named proteorhodopsin, encoded on a 130 kb genomic fragment from an uncultivated marine γ-proteobacterium of the SAR86 group. In biochemical characterizations this proteorhodopsin exhibited the same light-driven electron pumping as the bacteriorhodopsin known from halophilic archaea. These data and subsequent studies showed that there is high diversity within the proteorhodopsins, including different spectral tunings which reflect adaptation to different habitats [29]. A local and taxonomic wide distribution of proteorhodopsins further implied that this novel and so far completely overlooked type of phototrophy plays an important role in the ocean [30].

Beja et al. also constructed two BAC libraries, comprising 1632 and 4608 clones, with average insert sizes of 60 kb and 80 kb, respectively [21]. In these libraries they identified one genomic fragment of a planktonic euryarchaeote beside several bacterial genomic fragments (*Roseobacter* group, *Bacteroidetes* group, SAR11, SAR86, SAR116) using group-specific 16S rDNA and 23S rDNA and 16S-ITS-23S operon-specific primers. The 60 kb sequence of the marine euryarchaeotic genomic fragment was determined and contained 5S and 23S rRNA genes not linked to those of the 16S rRNA, alongside 38 protein encoding genes. Identification of crenarchaeal 16S rRNA gene-containing fosmids generated from microbial biomass of Antarctic surface water picoplankton opened the possibility for comparative genomic analyses of marine group I crenarchaeal genomic fragments [31]. The authors found similarities but also many differences between genome structure and gene content in their 16S rDNA-containing fragments and those identified in the study of Stein et al. [16] and from *Cenarchaeum symbiosum* [18]. Furthermore, the differences between protein-encoding genes and non-coding regions on fosmids with identical 16S rRNA genes indicated the genomic microheterogeneity found in this population. Within a cosmid library with 200–250 Mbp of stored DNA from 500-m-deep marine plankton, Lopez-Garcia et al. identified six genomic fragments that contained 16S rRNA genes of group II marine euryarchaeota and one fragment with the 16S rDNA of a marine group I crenarchaeote [32]. The crenarchaeal fragment was completely sequenced and analyzed, revealing potential horizontal gene transfers between eury- and crenarchaeota and between Archaea and Bacteria.

Recently a large whole-genome shotgun-sequencing project of microbial plankton from the Sargasso Sea has delivered more than 1.045 Gbp of non-redundant sequence from this ecosystem [24]. On the basis of sequence relatedness the authors estimated that their data were derived from approximately 1800 different species. Around 25% of the sequence reads could be assembled into 333 scaffolds spanning 30.9 Mbp with at least 3× coverage, of which 21 scaffolds had over 14× coverage and 9.35 Mbp of sequence. Most abundant in their data set, as determined by phylogenetic reconstructions with multiple phylogenetic markers, were sequences from species belonging to the phylum Proteobacteria, followed by those belonging to the Cyanobacteria and

other phyla. Altogether they identified 1.2 million genes, including 1164 small subunit rRNA genes and 782 genes for new photorhodopsin-like photoreceptors, showing the potential of the shotgun sequencing approach.

2.4 Environmental Genomics of Defined Communities: Biofilms and Microbial Mats

Biofilms and microbial mats are communities whose diversity can reach from a single species to a multitude of different organisms living together in a single habitat attached to a surface [33]. They are found in many habitats, ranging from human skin, intestinal stems, marine and freshwater sediments to acidic and thermophilic surfaces. Often the microbial species building the biofilm or mat cannot be cultured separately. In environmental genomic approaches these communities can be studied without laboratory cultivation.

For example, Schmeisser et al. [34] used different strategies to characterize a drinking water biofilm growing on rubber-coated valves. Beside an inventory of the 16S rDNA present in the biofilm they did a snapshot-sequencing approach of a small insert metagenomic library created from mixed DNA of the species occurring in the biofilm. From a cosmid library prepared from the same DNA the complete sequences of four clones were determined. This study gave initial insights into species composition and some genome fragments of this biofilm. No evidence of a potential health risk of the biofilm were found at the DNA- and protein sequence level.

A large-scale shotgun-sequencing approach on the metagenome of an acid mine drainage biofilm gave deep insight into the community structure and metabolisms of the different species [23]. Producing over 100,000 high-quality sequence reads from a small insert plasmid library (average insert size 3.2 kb) enabled the authors to reconstruct near-complete genomes of *Leptospirillum* (group II) and *Ferroplasma* (type II). In addition, three other genomes were partially recovered. On the basis of this comprehensive data set, potential metabolic pathways of the different species were predicted and information about their carbon, nitrogen, and energy metabolism was obtained. The population structure of the different species could also be analyzed, because every sequence read stemmed from a different individual of the population. Analysis of single-nucleotide polymorphisms revealed a mosaic genome structure of the *Ferroplasma* type II population that seems to have evolved from three ancestral strains by homologous recombination. Altogether, the data gave deep insight into the genomic potential of the biofilm. The study also showed that it is possible to assemble genomes in environmental genomic approaches, at least if there are only few species with low heterogeneity on the population level and if there are relatively similar abundances of the different species in the environment.

2.5 Environmental Genomics for Studies of Soil Microorganisms

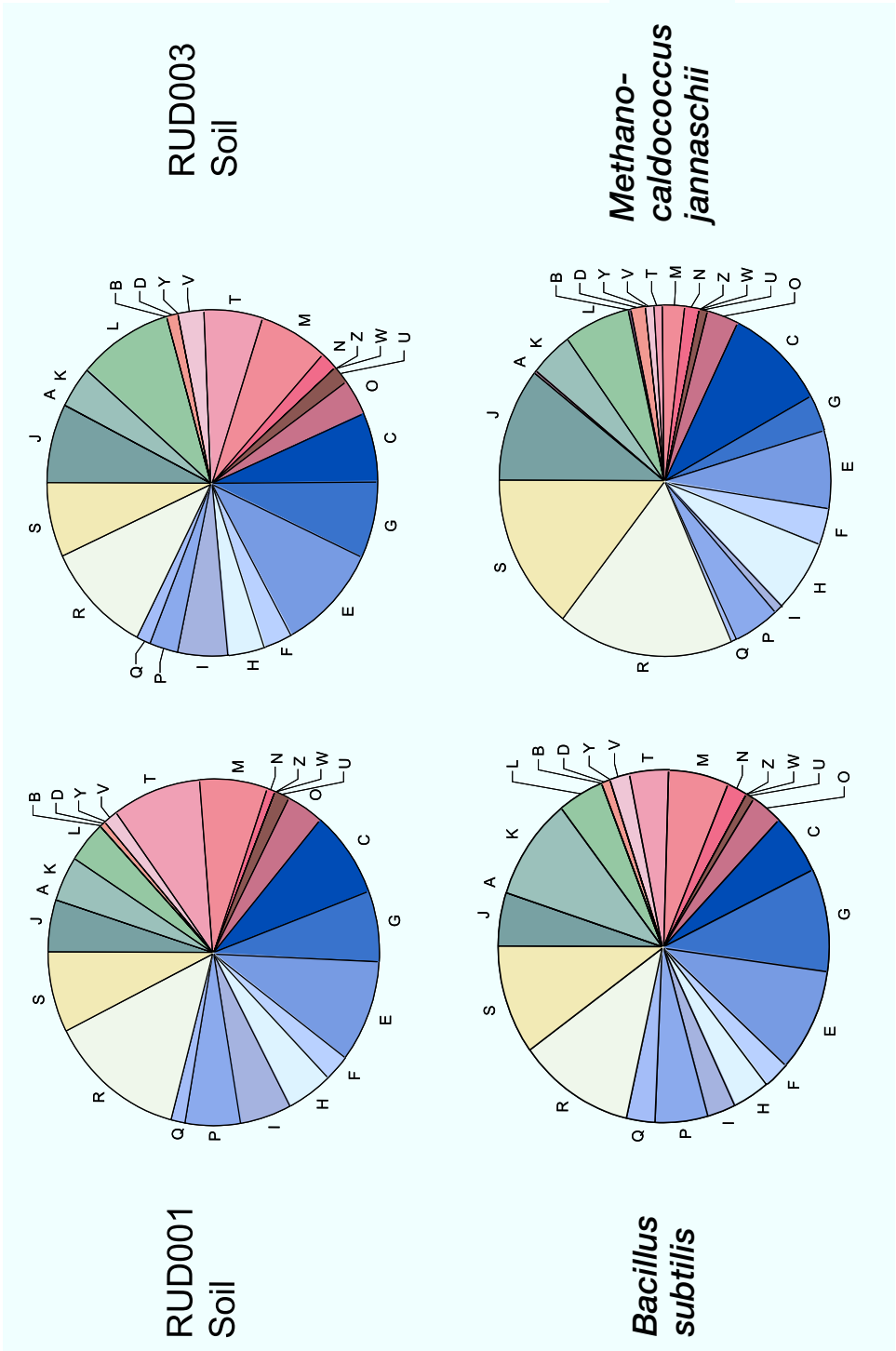
Soils are probably the most diverse ecosystems on the planet, with 12,000–18,000 different dominant species in one sample, as measured by means of DNA–DNA re-association kinetics [35]. Taking into account the different species frequencies by using estimates from animal and plant ecologists, one soil sample might even harbor up to 500,000 different species if 20,000 common

ones are estimated [8]. With this large diversity, environmental genomic approaches in soil might not be suitable for describing the complete set of genomes occurring in those habitats, but genome fragments of specific groups can be targeted. The use of soil samples is particularly challenging, because DNA preparations from soils are heavily contaminated with humic substances that tend to bind to nucleic acids and need to be specifically purified before cloning. Nevertheless, after development of purification methods, even for high molecular weight DNA, environmental genomic approaches have proved to be useful for the characterization of soil microorganisms.

Although many procedures have been established for purification of DNA from humic substances [36–38], most are only suitable for the isolation of small DNA fragments. On the basis of the procedure of Zhou et al. [36], Rondon et al. [39] developed a method for purifying high-molecular-weight DNA. They constructed two environmental BAC libraries with 27 kb (3600 clones) and 44.5 kb (25,600 clones) average insert sizes from soil. Screening resulted in identification of 16S rRNA gene-containing clones of *Proteobacteria*, *Cytophagales*, low G + C gram positives and *Acidobacteria*.

Another procedure for removal of humic acids and recovery of high-molecular-weight DNA was developed in our laboratory [20]. The key procedure is a two-phase pulse-field electrophoresis, with the first phase containing polyvinylpyrrolidone (PVP) and the second without PVP. The DNA is purified from humic substances, that are retarded by the PVP in the first phase and cleaned from PVP, that itself inhibits enzymatic reaction, in the second phase. The procedure enables isolation of highly pure, high-molecular-weight DNA in large amounts, by minimizing shearing effects. The resulting environmental fosmid libraries were used to characterize uncultivated Acidobacteria and Archaea [20, 22, 40]. Within three large insert environmental genomic libraries from two different soils, comprising altogether 75,678 clones with over 3 Gbp of stored DNA, 22 genomic fragments belonging to species of the phylum *Acidobacteria* were identified on the basis of 16S rRNA genes [40]. Four out of six sequenced fragments were placed within group V, and two within group III of the Acidobacteria, as determined by 16S rDNA comparisons. Sequences of fragments of the different groups varied in the G + C content of their protein-encoding regions by more than 10%, which was in good correlation with their phylogenetic placement. Although most encoded proteins could be assigned to "housekeeping" functions, some indicated special traits, e.g. a putative β-1′,2′-glucan synthetase, an intracellular polyhydroxybutyrate depolymerase, and an operon with low but significant similarity to a biosynthetic cluster of *Streptomyces lincolnensis*. Furthermore, a cluster of genes was highly homologous and syntenic to a gene cluster found in species of the *Rhizobiales*, implying a recent horizontal gene transfer between these bacterial lineages. Together with a study by Liles et al. [41], who identified and analyzed a 25-kb acidobacterial genomic fragment within a metagenomic soil library, these are the first reports of genomic information from species of the novel bacterial phylum *Acidobacteria*.

The same soil libraries were also characterized in a random end-sequencing approach to compare their content and to identify biases introduced by the cell lysis and cloning procedures [22]. On the basis of 5376 sequence tags of approximately 700 bp length that cover a total of ca. 4 Mbp, we showed that mostly bacterial and to a much lesser extent archaeal and eukaryotic genome fragments (ca. 1% each) had been


RUD003
Soil
Methano-
caldococcus
jannaschii
RUD001
Soil
Bacillus
subtilis

captured in the libraries. The diversity of putative protein-encoding genes, as reflected by their distribution into different COG clusters, was comparable with that encoded in complete genomes of cultivated microorganisms (Fig. 2.2). A huge variety of genomic fragments had been captured in the libraries, as seen by comparison with sequences in the public databases and by the large variation in G + C content. The study dissected differences in the libraries which relate to the different ecosystems analyzed and to biases introduced by different DNA preparations. Furthermore, a range of taxonomic marker genes (other than 16S rRNA) were identified that enabled assignment of genome fragments to specific lineages. The complete sequences of two genome fragments identified as being affiliated with archaea based on a gene encoding a CDC48 homolog and a thermosome subunit, respectively, together with a 16S rRNA containing fosmid gave first insights into the genomes of uncultivated crenarchaeota from soil [20, 22].

Together these studies demonstrate that environmental genomic approaches can also be applied to very diverse microbial populations such as those found in soil. Further, not only the most abundant microbial groups, e.g. acidobacteria [40], can be targeted with this approach, but also groups of organisms that occur in relatively low abundance (<5 %), e.g. mesophilic crenarchaeota [20, 22].

2.6 Biotechnological Aspects

Environmental genomic libraries can also serve as a fruitful resource for biotechnology. They enable researchers to identify genes for enzymes with novel properties without the need to cultivate the organisms. Soil habitats are of special interest, because many organisms in soils are capable of natural-product biosynthesis. These compounds often can be used as antimicrobials, cancer chemotherapeutics, and for immunomodulating and other pharmaceutical applications [42]. Wang et al. were the first to show that it is possible to isolate novel metabolites from soil DNA recombinants [43]. Using DNA isolated directly from soil samples by the method of Yap [44], they constructed and screened 1000 *Streptomyces lividans* recombinants to identify 18 clones that produced small molecules of the terragine/norcardamine family. Rondon et al. identified clones expressing DNAse, antibacterial, lipase and amylase activity [39]. In a cosmid library constructed from soil DNA, Brady et al. [45] identified a blue-colored clone that produced deoxyviolacein and the broad spectrum antibiotic violacein. The recombinant cosmid harbored the whole gene clusters responsible for the synthesis of these small molecules. The search for antimicrobial small molecules was the motivation for MacNeil et al. [42] to construct a BAC

Fig. 2.2 Distribution of BLASTX similarities among different COG categories (clusters of orthogologous genes, A: RNA processing and modification, B: chromatin structure and dynamics, etc., see www.NCBI.nlm.nih.gov). Datasets are from random sequence tags of two environmental fosmid libraries (RUD001 and RUD003) made from the same ecosystem but with different G + C content, compared with the distribution of predicted ORF from two complete microbial genomes, *B. subtilis* (heterotrophic soil bacterium) and *M. jannaschii* (autotrophic, hyperthermophilic archaeon). Note that although the soil libraries consist of a mixture of genomes from many different microorganisms, the distribution into COG is similar to that of genomes from cultivated microorganisms, demonstrating that some sort of "average" genome content is represented in the environmental libraries (details are given elsewhere [22].

library from soil DNA. Of the resulting clones with an average insert size of 37 kb ranging from 5 to 120 kb, 12,000 were picked and screened for antimicrobial activity. Altogether four clones could be identified, resulting in a hit rate of one antibacterial clone per 60 Mbp of soil DNA. One of the clones, producing a purple pigment, was further characterized and shown to produce indirubine, an antileukemic drug. Especially for identification of novel bioactive compounds Courtois et al. [46] constructed a 5000-clone library in an *Escherichia coli–Streptomyces lividans* shuttle cosmid vector using DNA extracted from soil microorganisms by a Nycodenz density gradient. The resulting library was screened by PCR to identify potential polyketide synthase gene clusters. Together with activity-based screening at least 13 bioactive compound-producing clones could be identified. Many other novel enzymes can be identified within environmental libraries, for example those needed for 4-hydroxybutyrate utilization [47], lipases and esterases [48].

2.7 Conclusions and Perspectives

Environmental genomics has proven to be a powerful tool for studying uncultivated microorganisms in marine and soil ecosystems and in biofilms. The development of novel technologies has helped overcome problems first encountered, mainly that of obtaining highly pure and sufficient DNA for constructing the complex libraries. Although the first studies in environmental genomics are very convincing, they also show the limitations of the approach in its current form – the more complex a microbial population, the less likely is the recovery of whole genomes. Partly-recovered genomes give interesting insight but do not tell the whole story. In future studies it might be interesting to construct biased or focused libraries in which genome fragments of the organisms of interest will be enriched before cloning. With falling costs for sequencing, shotgun approaches and large-scale sequencing efforts will also be more widely used. Environmental genomic studies will serve as a basis for studying activities of microorganisms in their environment by using functional genomic tools. High-throughput methods based on DNA arrays have already been applied to environmental questions [49]. Beside identification and genotyping of bacteria [50, 51], population genetics [52], and the detection and quantification of functional genes in the environment [53, 54] are the most promising applications. Together, these approaches will lay the ground for studying microbial physiology, species interactions, and networks of dependences between organisms in naturally occurring microbial populations.

References

1 www.ncbi.nlm.nih.gov
2 DSMZ (2004). Bacterial Nomenclature Up-to-date [http://www.dsmz.de/bactnom/bactname.htm]
3 Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* **276**:734–740.
4 Ludwig, W. (1999) http://www.mikrobiologie.tu-muenchen.de/pub/ARB, Department of microbiology, TU-München, Munich.
5 Rappe, M.S., and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu Rev Microbiol* **57**:369–394.
6 The Ribosomal Data Base Project II, June 2004, http://rdp.cme.msu.edu
7 Palleroni, N.J. (1997) Prokaryotic diversity and the importance of culturing. *Antonie Van Leeuwenhoek* **72**:3–19.
8 Dykhuizen, D.E. (1998) Santa Rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek* **73**:25–33.
9 Ferguson, R.L., Buckley, E.N., and Palumbo, A.V. (1984) Response of marine bacterioplankton to differential filtration and confinement. *Appl Environ Microbiol* **47**:49–55.
10 Eilers, H., Pernthaler, J., and Amann, R. (2000) Succession of pelagic marine bacteria during enrichment: a close look at cultivation-induced shifts. *Appl Environ Microbiol* **66**:4634–4640.
11 Janssen, P.H., Yates, P.S., Grinton, B.E., Taylor, P.M., and Sait, M. (2002) Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. *Appl Environ Microbiol* **68**:2391–2396.
12 Sait, M., Hugenholtz, P., and Janssen, P.H. (2002) Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. *Environ Microbiol* **4**:654–666.
13 Joseph, S.J., Hugenholtz, P., Sangwan, P., Osborne, C.A., and Janssen, P.H. (2003) Laboratory cultivation of widespread and previously uncultured soil bacteria. *Appl Environ Microbiol* **69**:7210–7215.
14 Kaeberlein, T., Lewis, K., and Epstein, S.S. (2002) Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment. *Science* **296**:1127–1129.
15 Zengler, K., Toledo, G., Rappe, M., Elkins, J., Mathur, E.J., Short, J.M., and Keller, M. (2002) Cultivating the uncultured. *Proc Natl Acad Sci U S A* **99**:15681–15686.
16 Stein, J.L., Marsh, T.L., Wu, K.Y., Shizuya, H., and DeLong, E.F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178**:591–599.
17 Schleper, C., Swanson, R.V., Mathur, E.J., and DeLong, E.F. (1997) Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon Cenarchaeum symbiosum. *J Bacteriol* **179**:7803–7811.
18 Schleper, C., DeLong, E.F., Preston, C.M., Feldman, R.A., Wu, K.Y., and Swanson, R.V. (1998) Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon Cenarchaeum symbiosum. *J Bacteriol* **180**:5003–5009.

19 Tebbe, C.C., and Vahjen, W. (1993) Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Appl Environ Microbiol* **59**:2657–2665.

20 Quaiser, A., Ochsenreiter, T., Klenk, H.P., Kletzin, A., Treusch, A.H., Meurer, G. et al. (2002) First insight into the genome of an uncultivated crenarchaeote from soil. *Environ Microbiol* **4**:603–611.

21 Beja, O., Suzuki, M.T., Koonin, E.V., Aravind, L., Hadd, A., Nguyen, L.P. et al. (2000b) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**:516–529.

22 Treusch, A.H., Kletzin, A., Raddatz, G., Ochsenreiter, T., Quaiser, A., Meurer, G. et al. (2004) Characterization of Large-Insert DNA Libraries from Soil for Environmental Genomic Studies of Archaea. *Environ Microbiol* **6**:970–980.

23 Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M. et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**:37–43.

24 Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**:66–74.

25 Riesenfeld, C.S., Schloss, P.D., Handelsman, J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* **38**:525–552.

26 Preston, C.M., Wu, K.Y., Molinski, T.F., and DeLong, E.F. (1996) A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc Natl Acad Sci U S A* **93**:6241–6246.

27 DeLong, E. (2003) Oceans of Archaea. *ASM News* **69**:503–511.

28 Beja, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P. et al. (2000a) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**:1902–1906.

29 Beja, O., Spudich, E.N., Spudich, J.L., Leclerc, M., and DeLong, E.F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411**: 786–789.

30 de la Torre, J.R., Christianson, L.M., Beja, O., Suzuki, M.T., Karl, D.M., Heidelberg, J., and DeLong, E.F. (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc Natl Acad Sci U S A* **100**:12830–12835.

31 Beja, O., Koonin, E.V., Aravind, L., Taylor, L.T., Seitz, H., Stein, J.L. et al. (2002) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl Environ Microbiol* **68**:335–345.

32 Lopez-Garcia, P., Brochier, C., Moreira, D., and Rodriguez-Valera, F. (2004) Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ Microbiol* **6**:19–34.

33 Guerrero, R., Piqueras, M., and Berlanga, M. (2002) Microbial mats and the search for minimal ecosystems. *Int Microbiol* **5**:177–188.

34 Schmeisser, C., Stockigt, C., Raasch, C., Wingender, J., Timmis, K.N., Wenderoth, D.F. et al. (2003) Metagenome survey of biofilms in drinking-water networks. *Appl Environ Microbiol* **69**:7298–7309.

35 Torsvik, V., Sorheim, R., and Goksoyr, J. (1996) Total bacterial diversity in soil and sediment communities – a review. *J Ind Microbiol* **17**:170–178.

36 Zhou, J., Bruns, M.A., and Tiedje, J.M. (1996) DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**:316–322.

37 Krsek, M., and Wellington, E.M. (1999) Comparison of different methods for the isolation and purification of total community DNA from soil. *J Microbiol Methods* **39**:1–16.

38 Miller, D.N., Bryant, J.E., Madsen, E.L., and Ghiorse, W.C. (1999) Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl Environ Microbiol* **65**:4715–4724.

39 Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R. et al. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66**:2541–2547.

40 Quaiser, A., Ochsenreiter, T., Lanz, C., Schuster, S.C., Treusch, A.H., Eck, J., and Schleper, C. (2003) Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Mol Microbiol* **50**:563–575.

41 Liles, M.R., Manske, B.F., Bintrim, S.B., Handelsman, J., and Goodman, R.M. (2003) A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol* **69**:2684–2691.

42 MacNeil, I.A., Tiong, C.L., Minor, C., August, P.R., Grossman, T.H., Loiacono, K.A. et al. (2001) Expression and isolation of antimicrobial small molecules from soil DNA libraries. *J Mol Microbiol Biotechnol* **3**:301–308.

43 Wang, G.Y., Graziani, E., Waters, B., Pan, W., Li, X., McDermott, J. et al. (2000) Novel natural products from soil DNA libraries in a streptomycete host. *Org Lett* **2**:2401–2404.

44 Yap, W.H., Li, X., Soong, T.W., and Davies, J. (1996) Genetic diversity of soil microorganisms assessed by analysis of *hsp70* (*dnaK*) sequences. *J Ind Microbiol* **17**:179–184.

45 Brady, S.F., Chao, C.J., Handelsman, J., and Clardy, J. (2001) Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. *Org Lett* **3**:1981–1984.

46 Courtois, S., Cappellano, C.M., Ball, M., Francou, F.X., Normand, P., Helynck, G. et al. (2003) Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol* **69**:49–55.

47 Henne, A., Daniel, R., Schmitz, R.A., and Gottschalk, G. (1999) Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl Environ Microbiol* **65**:3901–3907.

48 Henne, A., Schmitz, R.A., Bomeke, M., Gottschalk, G., and Daniel, R. (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on Escherichia coli. *Appl Environ Microbiol* **66**:3113–3116.

49 Guschin, D.Y., Mobarry, B.K., Proudnikov, D., Stahl, D.A., Rittmann, B.E., and Mirzabekov, A.D. (1997) Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. *Appl Environ Microbiol* **63**:2397–2402.

50 Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L., and Falkow, S. (2000) A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains. *Proc Natl Acad Sci U S A* **97**:14668–14673.

51 Cho, J.C., and Tiedje, J.M. (2001) Bacterial species determination from DNA–DNA hybridization by using genome fragments and DNA microarrays. *Appl Environ Microbiol* **67**:3677–3682.

52 Chakravarti, A. (1999) Population genetics – making sense out of sequence. *Nat Genet* **21**:56–60.

53 Wu, L., Thompson, D.K., Li, G., Hurt, R.A., Tiedje, J.M., and Zhou, J. (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl Environ Microbiol* **67**:5780–5790.

54 Taroncher-Oldenburg, G., Griner, E.M., Francis, C.A., and Ward, B.B. (2003) Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Appl Environ Microbiol* **69**:1159–1171.

3
Applications of Genomics in Plant Biology

Richard Bourgault, Katherine G. Zulak, and Peter J. Facchini

3.1
Introduction

It is often said that genomics has "revolutionized" plant biology [1]. Genomic technologies have dramatically enhanced our ability to characterize the genetic architecture of plant genomes and better understand the relationships between genotype and phenotype. Within every discipline of plant biology genomics has enabled parallel expression analysis of thousands of genes, characterization of the precise protein composition of specific tissues, and identification of the genetic basis for specific mutations. The genome of the model plant *Arabidopsis thaliana* has been almost completely sequenced providing plant biologists with a "blueprint" for the assembly and operation of a plant, and expediting application of novel technologies to functional analysis of plant genetic information. Several rapid and multiparallel approaches, broadly known as *functional genomics*, have become routine approaches for assigning functions to new genes. Such methods enable unprecedented and thorough analysis of the plant cellular components that determine gene function, specifically transcripts, proteins, and metabolites (Fig. 3.1). Similarly the phenotypic variations of entire mutant collections are being analyzed fast-

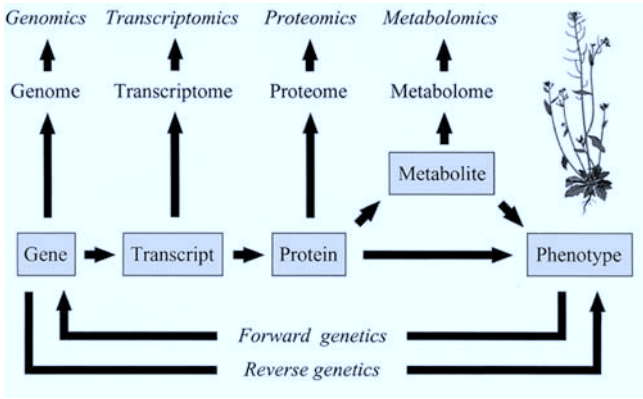


Fig. 3.1 Relationships among different levels of cellular function and corresponding genomics technologies.

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

er and more efficiently than ever before by use of *reverse genetic* approaches. Different methods, termed *transcriptomics*, *proteomics*, and *metabolomics*, have evolved into distinct strategies within the functional genomics platform and these approaches have been fully integrated into field of plant biology (Fig. 3.1). However, the role of genes with unknown functions cannot be fully determined and appreciated by use of a single genomic technique. Only consolidation of information collected using different genomic tools will facilitate the assignment of functions to the vast numbers of plant genes that remain uncharacterized. In this chapter the applications of genomic techniques to plant biology are discussed with a focus on the impact of genomics on strategies used to investigate development and function.

3.2 Plant Genomes

3.2.1 Structure, Size, and Diversity

All living organisms have a unique genome that can be defined as the full (haploid) complement of genetic information required to build and to maintain a functional life form. This genetic information, in the form of DNA, provides the code to produce the proteins that are involved in processes such as cellular metabolism, transcription, translation, defense, signaling, growth, protein fate, and transport. In plants, as in other eukaryotes, protein-coding genes have a coding region flanked by untranslated regions and can be split into exons and introns. In contrast with the function of prokaryotic operons, in which the transcription of multiple genes is often controlled by a single promoter, each eukaryotic gene has its own promoter.

Genome structure can be defined as the overall organization of genes in an organism. Although there are exceptions [**2**], in the vast majority of prokaryotes all genes are contained on a single circular chromosome, which is relatively small in size (4.6 megabasepairs [Mb] for *E. coli*; 4.2 Mb for *Bacillus subtillus*) [3]. In plants and other eukaryotes, nuclear genes are contained on much larger, linear chromosomes, which vary greatly in number and size from one species to another. Fig. 3.2 shows relative genome sizes in bacteria, yeast, animals and plants and also illustrates the broad range of genome size in the animal and plant kingdoms. In addition, many plants have elevated ploidy levels in which the entire diploid genome (originating from maternal and paternal sources) is present in multiple copies. It should be noted that a larger and more complex genome does not necessarily lead to a more complex organism. Much of the DNA present in plants and other organisms with large genomes is non-coding, repetitive DNA, and the total number of expressed genes is not as proportionally large as genome size might suggest [4]. For example, in rice 42.2 % of the 430 Mb genome was found to be in precise 20-nucleotide oligomer repeats situated in intergenic regions between genes [5].

In addition to chromosomes in the plant cell nucleus, the plastid (chloroplast) and mitochondrion are organelles that also carry their own genetic information. It is important to note that both organelles require many gene products encoded by nuclear DNA for them to take form and be functional. The plastid chromosome is in the form of a closed circle, typically about 120–160 kb in size, with genes densely arranged in an order that is highly con-

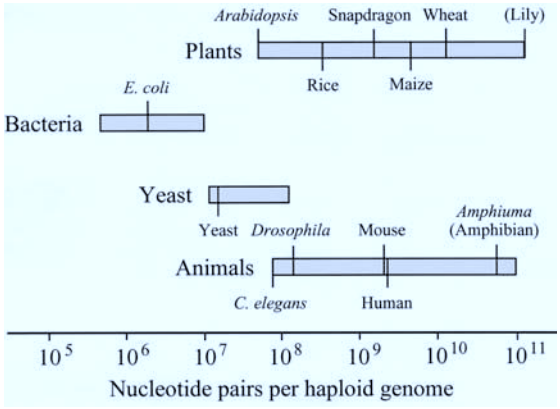


Fig. 3.2 Haploid genome size, in base pairs, of different organisms. Most eukaryotes have haploid genome sizes between 10^7 and 10^{11} base pairs of DNA. Plants have a full spectrum of haploid genome sizes compared with other eukaryotes.

served among plants. The mitochondrial chromosome can be circular or linear with genes being more widely dispersed than in the plastid. For example, in the *Arabidopsis* mitochondrial genome, coding regions make up less than 10 % of the total 367 kb of DNA and encode only 57 identified genes [6]. The size of the mitochondrial genome also varies greatly from one plant species to another. For example, the mitochondrial genome is 200 kb in *Oenothera spp.* and *Brassica spp.*, whereas it is up to 2600 kb in muskmelon [7]. This is in contrast to animal mitochondrial genomes, which are relatively compact (approx. 16 kb).

According to the endosymbiont hypothesis, the origin of the chloroplast and mitochondrion arose when an early protoeukaryotic organism engulfed an ancestral cyanobacterium and an ancestral α-proteobacterium. The photosynthetic cyanobacterium evolved into the plastid and the α-proteobacterium became the mitochondrion. Over the course of time and evolution, most of the DNA once present in the endosymbionts was transferred to the nuclear genome of the protoeukaryote [7]. This phenomenon has been confirmed by evidence from genome sequencing projects on rice and *Arabidopsis* [8].

3.2.2 Chromosome Mapping: Genetic and Physical

The location, or locus, of a specific gene within a genome can be described by using a genetic and/or physical map. A genetic linkage map locates a gene in relation to another gene or, using more current techniques, to a molecular marker on the same chromosome. Genetic mapping is based on the concept that greater distances between two linked genes create a higher probability that homologous chromatids will cross over in the region between the genes during meiosis. This produces a larger proportion of recombinant progeny compared with limited segregation of linked genes that are situated closer together. Recombination frequency can be measured by observing the co-inheritance of two particular phenotypes, or a phenotype and a molecular marker. A molecular marker is a site of heterozygosity for some types of neutral DNA variation which is not associated with any phenotypic variation. These variations are caused by differences in the location of restriction enzyme cleavage sites caused by single nucleotide changes, or by the presence of a variable number of short, tandem repeats. One genetic map unit (m.u.), also

called a centimorgan (cM), is defined as that distance between genes for which one product of meiosis out of 100 is recombinant [9]. In early plant genetic research, genetic mapping was conducted by crossing, and/or self-pollinating, heterozygous plants with a type of observable phenotype or mutation that could be followed in the progeny. This type of classical work provided detailed genetic maps for species such as *Arabidopsis* [10], tomato [11], and corn [12], but there were large gaps in the genomes to which no measurable phenotypes could be assigned. High-resolution maps can be made using molecular markers such as RFLP (restriction fragment length polymorphisms), SSLP (simple-sequence length polymorphisms), and RAPD (randomly amplified polymorphic DNA). Dense linkage maps developed using this approach continue to facilitate the positional cloning of important genes, enable the genetic dissection of quantitative trait loci, assist in phylogenetic comparisons between related species, and provide a critical guide for assembly of accurate physical maps.

A physical map is generated by directly examining the physical material of the genome, that is, its DNA composition. Genomic DNA is isolated, cleaved, and assembled into clones that are manageable in size for examination and manipulation. Very large fragments of DNA (~1 Mb) can be cloned into yeast vectors termed YAC (yeast artificial chomosomes) whereas smaller fragments can be placed into BAC (bacterial artificial chromosomes), PAC (phage P1-based artificial chromosomes), and TAC (transformation-competent artificial chromosomes). Still smaller fragments can be cloned into conventional plasmid DNA vectors. Although different strategies can be used to actually create the physical map, the result is the assembly of cloned fragments into overlapping groups called contigs. Ultimately, the aim is to represent a complete chromosomal segment with an overlapping set of contigs. Distances between genes or markers can then be defined in terms of "base pairs" as opposed to the more ambiguous "map unit", which is used to describe genetic, rather than physical distances. A physical map can then be integrated with molecular markers from a genetic map [9].

The highest resolution physical map is one in which the entire genome has been sequenced. By generating cloned fragments such as those described above, high-throughput data can be acquired using automated DNA sequencers capable of reading multiple sequences, often up to 800 bp with high confidence. Using sequence-analysis software the data are screened for overlapping regions, and these are used to build large contigs representing an entire chromosome. The sequence data is then submitted to a database, assigned an accession number and classified into one of four phases of sequence quality. The National Institute for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov) defines these phases as: Phase 0 – single/few pass reads of a single clone (not contigs); Phase 1 – unfinished, might be unordered, unoriented contigs, with gaps; Phase 2 – unfinished, ordered, oriented contigs, with or without gaps; and Phase 3 – finished, no gaps (with or without annotations). The genome sequencing of several organisms has been completed to Phase 3. Currently, the only plant to reach Phase 3 for its entire genome is *Arabidopsis thaliana* [8].

3.2.3 Large-scale Sequencing Projects

In December 2000 it was widely reported that the first plant to have its entire genome

sequenced was the crucifer *Arabidopsis* [8]. The choice of *Arabidopsis* as the model plant for this project might have been a surprise to observers outside the world of plant biology, because it is a weed, not a crop plant. The advantages of using *Arabidopsis* for genome analysis, however, include its relatively small, diploid nuclear genome, a capacity for self-pollination, amenability to out-crossing, a short life cycle, the production of a large number of progeny, and the small physical size of the plant. A wealth of classical genetic information was also already available for *Arabidopsis*. It was predicted that information gleaned from the *Arabidopsis* genome sequence would be helpful in the development of strategies to improve traits of agronomic value in important crop species [8].

When the nearly completed *Arabidopsis* sequence data was organized into contigs representing all five chromosomes, the next major task was to identify genes and assign function to the gene products. This process, referred to as genome annotation, was performed using a combination of algorithms optimized with data based on known *Arabidopsis* gene structures, expressed sequence tags (ESTs), and characterized proteins. Specific programs used to predict introns enabled identification of continuous open reading frames (ORFs), and gene prediction software programs were used to suggest the possible function of each gene product.

Experiments performed to confirm the quality of annotation showed that 93 % of the ESTs matched the predicted ORFs and less than 1 % of the ESTs matched predicted non-coding regions; thus, most putative genes were identified. It was concluded that the 125-Mb genome contains 25,498 genes encoding proteins from 11,000 families. The functions of 69 % of the genes were classified by sequence similarity to proteins of known function from all organisms. The functions of only 9 % of the total compliment of genes have been determined experimentally, however. Approximately 30 % of the predicted gene products, some of which are homologous to proteins found in non-plant species, could not be assigned to a functional category. Predictions based on signal peptides estimated that about 17 % of the gene products are targeted to the secretory pathway, 11 % to mitochondria, and 14 % to the chloroplast. The functional category with the highest proportion of all classified gene products was cellular metabolism (22.5 %), followed by transcription (16.9 %), plant defense (11.3 %), growth (11.7 %), signaling (10.4 %), protein fate, intracellular transport, extracellular transport, and protein synthesis (all <10 % each) [8].

The availability of a nearly complete genome sequence for *Arabidopsis* enables comparisons to be made between plants and other organisms with fully sequenced and annotated genomes. Such comparisons are advancing our understanding of the evolution and biological function of plants and other organisms. Comparisons with diverse organisms such as bacteria, yeast, multicellular eukaryotes, and other plants have been conducted, and the findings are occasionally predictable and at other times surprising. For example, 48–60 % of all genes involved in protein synthesis have counterparts in other eukaryotic genomes, reflecting highly conserved gene function. On the other hand, only 8–23 % of *Arabidopsis* proteins involved in transcription have related genes in other eukaryotic genomes, reflecting more independent evolution. Comparisons of *Arabidopsis* with other plant species are likely to provide the most useful insights toward developing strategies for crop improvement, which is one reason several other plant species are

currently the targets of large-scale sequencing projects.

Crop plants currently selected for large-scale genomic sequencing projects include, soybean, corn, *Medicago truncatula*, and *Oryza sativa* (rice). Of these, the most progress has been achieved with the cereal crop rice. Rice has been cultivated for more than 9000 years and is a major food source for over 50 % of the world population. Rice is considered a model system for plant biology, largely because of its compact genome (430 Mb), relative ease of transformation, and evolutionary relationships with other large-genome cereals such as sorghum (750 Mb), corn (2,500 Mb), barley (5000 Mb), and wheat (15,000 Mb). In 2002, the entire rice genome had been sequenced to Phase 2 quality [5, 13] and in June 2003 chromosomes 1, 4, and 10 (out of 12) had been completed to the Phase 3 level [14–16]. On the basis of the non-overlapping sequences of the 12 chromosomes at Phase 2, a total of 62,435 genes were predicted within a 366-Mb genome (International Rice Genome Sequencing Project – IRGSP 2002; http://rgp.dna.affrc.go.jp/rgp/Dec18_NEWS). The groups that have completed chromosomes 1, 4, and 10 have all made strong arguments supporting the need for Phase 3 quality sequence for all 12 chromosomes. At a February 2004 meeting of the IRGSP it was concluded that the consortium was easily on target to complete the entire genome to Phase 3 quality by December 2004 (http://rgp.dna.affrc.go.jp/IRGSP). A four-year, NSF-funded project to annotate the entire rice genome and assemble the data into a user-friendly format was undertaken in January 2004 by The Institute for Genomics Research (TIGR) (www.tigr.org).

Comparisons showed there is much synteny (i.e. gene order and orientation) and gene homology between rice and other cereal genomes [13]. As expected, synteny with *Arabidopsis* is limited because of to extensive genome reshuffling since the divergence of the monocot and dicot plants [15]. At the level of individual genes, however, the similarities are quite pronounced. Protein-coding genes on chromosomes 1, 4, and 10 with significant homologs in *Arabidopsis* comprise 47, 44, and 67 % of all genes, respectively. A reverse comparison using draft sequences revealed that up to 85 % of predicted *Arabidopsis* genes have homologs in rice [5, 15]. Data obtained from these and other large-scale sequencing projects will undoubtedly create new strategies for improving agronomic traits in cultivated crops that cannot be realized by conventional breeding.

3.3 Expressed Sequence Tags

Expressed sequence tag (EST) databases currently hold the largest amount of nucleotide sequences from plant genomes. ESTs are generated by single-pass sequencing of random cDNA and provide a quick, cost-effective method for generating a large inventory of expressed genes. Because the complete sequencing of most plant genomes is difficult, EST sequencing has emerged as a robust method of rapidly sampling expressed protein-encoding sequences of many plant species. More than 16.1 million ESTs are available in the European Molecular Biology Laboratory (EMBL) sequence database, including over 3.1 million from more than 200 representative plant species [17]. Model plant systems such as *Arabidopsis* and crop plants such as wheat, rice, and soybean are most highly represented in current EST databases. The nearly complete *Arabidopsis* EST collection has been used to refine the annotations of the *Arabidopsis* genome [35], and about

96 % of all ESTs were mapped back to a genomic locus, enabling improved annotation of the genome and verification of EST clusters and sequence quality.

Large EST collections have also been used for comparative genomic analysis among and between different plant families. EST collections from two members of the Solanaceae family, potato and tomato, were found to overlap between 70–80 % of all cDNAs sequenced [19]. Over 70 % of tomato EST sequences overlap with *Arabidopsis*, a member of the Brassicaceae family [20]. Comparison of 11,008 ESTs from the order Asparagales (e.g. onion) to the vast EST collections from plants of the order Poales (e.g. wheat, barley, corn, and rice) revealed inconsistencies in genomic characteristics across the monocots [21]. Although onion and rice share 78 % mean nucleotide similarity across coding regions, there are significant differences between characteristics such as codon usage, mean GC content, GC distribution, and relative GC content at each codon position. This comparison indicates that plants of the order Asparagales are more similar to eudicots than to the Poales for the measured genomic characteristics.

ESTs have been used to compare gene expression in plants under different environmental conditions. Gene expression has been investigated by comparing two EST databases from autumn leaves of field-grown *Populus tremuloides* (aspen) and greenhouse-grown plants with young, fully expanded leaves [22]. Comparison showed the young-leaf EST database contained mostly photosynthetically related genes whereas the autumn leaf library contained an abundance of senescence-induced metallothionnein genes, in addition to cysteine and aspartic proteases. Plastid protein synthesis was also found to be less than 10 % of that in young leaves.

EST analysis has also greatly increased our understanding of the development and physiology of plant tissues. Phloem tissues transport nutrients and other biologically active molecules throughout the plant, yet the molecular mechanisms involved in phloem function are poorly understood. A phloem-specific EST collection has been constructed using *Apium graveolens* (celery) as a model [23]. The main categories of mRNA found in phloem tissues encoded proteins involved in phloem structure, metal homeostasis and distribution, stress responses, and the degradation or turnover of proteins. Moreover, several novel phloem-specific genes were discovered, but require further characterization.

ESTs have also greatly accelerated the rate of gene discovery in plant secondary metabolic pathways. *Mentha* x *piperita* (mint) produces some of the world's most valuable and widely used essential oils. The flavor and scent compounds found in mint are monoterpenes, which are synthesized and stored in specialized secretory structures called peltate glandular trichomes. An EST collection produced specifically from these oil gland secretory cells provided a rich resource of ESTs related directly to essential oil biosynthesis and transport [24]. Over 25 % of the ESTs sequenced represent genes directly involved in essential oil metabolism. Such databases are promising targets for genetic engineering of natural product metabolism in plants. Basil (*Ocimum basilicum)* produces and stores volatile oils containing a limited number of phenylpropanoid compounds within glandular trichomes. EST sequencing of randomly selected glandular trichome cDNAs revealed a relatively high representation (i.e. over 25 %) of known phenylpropanoid biosynthetic genes [25]. Genes related to the biosynthesis of *S*-adenosylmethionine, an important substrate for phenylpropanoid

metabolism, were also common. EST analysis has thus shown that glandular trichomes are highly specialized structures involved in both the biosynthesis and storage of important secondary metabolites. EST sequencing has also been used to further the discovery of novel genes involved in the characteristic floral fragrance of rose [26]. The commercial importance of rose petals depends on compounds involved in their color and scent. A collection of approximately 2100 sequences from rose petals led to the discovery of several critical genes involved in the biosynthesis of volatile sesquiterpenes, the main compounds responsible for the characteristic floral scent of roses.

3.4 Gene Expression Profiling Using DNA Microarrays

There has been unprecedented growth in publicly available plant EST collections over the last few years. For this data to be biologically informative each sequence must be associated with a precise biological function. DNA microarray analysis has been used as a high-throughput method to accelerate the characterization of gene functions and decipher complex gene expression patterns in plant systems. Two types of microarrays are commonly used for expression profiling in plants – oligonucleotide and DNA fragment-based microarrays. *Oligonucleotide-based microarrays* are produced by using selected sequences from existing EST databases. Typically, 25-mer to 70-mer oligonucleotides are synthesized from sequences based on these ESTs and printed on chemically coated glass slides using a photolithographic process. The most widely available and advanced oligonucleotide-based microarrays are available for *Arabidopsis*, and are manufactured by Affymetrix and by Agilent. Recently, a barley microarray (or GeneChip) containing 22,000 different oligonucleotides has become a publicly available new resource for cereal crop genomics research [27]. *DNA fragment-based microarrays* are produced by robotically spotting polymerase chain reaction (PCR)-derived cDNA on to a chemically coated glass slide. The technology used to produce DNA fragment-based microarrays is more widely available and less costly; thus, it remains the method of choice for non-model plant species. For either type of microarray, experiments are performed by isolating mRNA from two different plant or tissue samples; this mRNA is reverse-transcribed to produce cDNA that incorporates nucleotides conjugated to two different fluorescent dyes, each with a unique excitation and emission spectrum. The dye-labeled cDNA will hybridize to the immobilized oligonucleotides or DNA fragments on the microarray slide. The relative transcript abundance between the two samples can be determined by measuring the amount of each label hybridized to each spot. Thus, the expression data of thousands of genes can be deduced from one experiment.

Microarray analysis is an important tool in the deciphering of biotic and abiotic stress-responses in plants. This approach was used to identify cold-, drought-, high-salinity-, and/or absicisic acid (ABA)-inducible genes in rice [28]. By identifying genes inducible under more than one stress condition, a complex pattern of cross-talk among signaling pathways was discovered. Greater cross-talk between signaling pathways coupled to drought, ABA, and high-salinity stresses was observed, whereas less cross-talk occurred between cold and ABA, or cold and high-salinity stresses. Comparison with *Arabidopsis* data showed that 51 out of 73 stress-induced genes have similar

functional annotations in *Arabidopsis*, indicating there are both similarities and differences in stress responses between *Arabidopsis* and rice.

Microarray analysis has been used to investigate transcript level changes in *Nicotiana attenuata* plants challenged with the specialist herbivore *Manduca sexta*, the results provide insight into the signaling and the transcriptional basis of plant defense against herbivores [29]. Simultaneous activation of salicylic acid-, ethylene-, cytokinin-, WRKY-, MYB-, and oxylipin signaling pathways was observed upon herbivore attack. Systemic increases in defense responses were reciprocated by decreases in photosynthetic gene transcript levels, and increases in protein turnover and carbohydrate metabolism gene transcripts. Such widespread changes in gene expression indicate a major metabolomic reconfiguration and a high degree of complexity only detectible using high-throughput genomic approaches.

Microarrays can also be used to analyze changes in the transcriptome of transgenic plants. Several cold-inducible downstream genes of the *Arabidopsis* DREB1A-CBF3 transcription factor were identified by using microarrays [30]. Transgenic plants overexpressing the transcription factor were hybridized along with a wild-type plant to identify differentially expressed genes. Thirty-eight downstream stress-related genes were upregulated in the transformed plants along with 20 previously unreported novel genes. DREB1A-CBF3 was shown to repress the expression of genes involved in plant growth and development, indicating that DREB1A-CBF3 regulates a complex network of genes which could not be easily identified using traditional methodologies.

Transcriptional profiling has also been used to further our understanding of basic plant cell biology and physiology. Microarray analysis has revealed unique characteristics of pollen in *Arabidopsis* [31]. Pollen tubes are used a model to study plant cell growth because of their ability to elongate without cell division. Little is understood, however, about the genetic basis of pollen germination and pollen tube growth. Comparison of the expression profiles of hydrated pollen grains and several vegetative tissues revealed 162 genes most abundantly expressed in pollen that had previously not been associated with this tissue. The discovery of such genes provide new targets to enable understanding of the genetic basis for pollen development and function. Microarray analysis can also be used to identify genes differentially expressed in specific cell types. By use of laser-capture microdissection (LCM) technology, different cell types can be isolated from plant tissues. Total RNA can be extracted from isolated cells (typically 1000 to 10,000 cells) and used to reverse transcribe labeled cDNA, which can be hybridized to a microarray. Several genes were differentially expressed in epidermal cells or vascular tissues of *Zea mays* (maize) [32]. Approximately 250 out of 8800 genes were differentially expressed in both tissues. This study demonstrates the feasibility of using LCM and microarray analysis to conduct high-resolution global transcriptional gene expression analysis in plants.

Microarray analysis has also contributed to our basic understanding of sugar signal transduction pathways in *Arabidopsis* [33]. Sugar and nitrogen are important plant nutrient signals, yet much remains to be understood about the molecular mechanisms underlying these critical signaling components. Microarrays were used to investigate the effects of glucose and inorganic nitrogen on global transcriptional changes. Glucose was found to regulate a broad range of genes involved in carbohy-

drate metabolism, signal transduction, metabolite transport, and even plant stress-responses. Several ethylene biosynthetic and signal transduction genes were repressed by glucose, indicating possible cross-talk between the two signaling pathways. Glucose was found to interact with nitrogen in the regulation of gene expression, indicating that glucose and nitrogen act both as metabolites and as signaling molecules.

3.5 Proteomics

The availability of entire genome sequences, EST databases, and gene expression profiles from a growing number of plant species provides a solid foundation for the application of proteomics to the establishment of a strong functional genomics platform in plants. Also termed "protein profiling", proteomics is the study of the identification, function, and regulation of entire sets of proteins in a tissue, cell, or subcellular compartment. Such information is crucial to understanding how complex biological processes occur at a molecular level and how they differ in various cell types, stages of development, or environmental conditions. Many of the genes identified in genomic and EST databases encode proteins of unknown, hypothetical, or putative function; thus, an examination of protein profiles is essential to fully integrate these data. The definition of proteomics can be subdivided into two specific categories: shotgun or functional approaches [34]. Shotgun proteomics involves a qualitative and/or quantitative survey of all the proteins in a sample. In contrast, functional proteomics uses the same techniques, but focuses on one, or a few, proteins involved in a specific process.

Recent years have seen remarkable advances in the technologies used to conduct proteomics research. The classical technique involves the use of two-dimensional gel electrophoresis (2DE) whereby extracted proteins are separated on polyacrylamide gel on the basis of their isoelectric charge and molecular weight. Individual proteins appear as spots on the gel and can be roughly quantified by measurement of the intensity of the spots. The proteins can be identified by excising the spot from the gel, digesting the polypeptide into smaller peptide fragments using specific proteases, for example trypsin, and sequencing the peptides directly or analyzing them by use of a mass spectrometer (MS). Although this method is still useful and widely used it is limited by sensitivity, resolution, and the abundance range of the different proteins in the sample [18, 34]. For example, abundant proteins dominate the gel whereas less abundant proteins might not be visible. New approaches involve both improved separation methods and advanced detection equipment.

Key developments leading to improved detection of proteins were time-of-flight (TOF) MS and relatively nondestructive methods for converting proteins into volatile ions [18]. Two "soft ionization" methods, matrix-assisted laser-desorption ionization (MALDI) and electrospray ionization (ESI), have made it possible to analyze large molecules such as peptides and proteins. Although MALDI-TOF MS is a relative high-throughput method compared with ESI, the latter is more easily coupled to separation techniques such as liquid chromatography (LC) or high-pressure LC (HPLC) [18]. This has provided an attractive alternative to 2DE, because even low-abundance proteins and insoluble transmembrane proteins can be detected [36, 37]. All MS-based techniques require a substantial and search-

able database of predicted proteins, ideally representing the entire genome. Protein identification is possible by comparing the deduced masses of the resolved peptide fragments with the theoretical masses of predicted peptides in the database.

Mass spectrometers are restricted in the number of ions that can be detected at any point in time. Pre-fractionation of proteins on the basis of isolation of specific cell types or subcellular organelles is often necessary to reduce the complexity of the sample [38]. Another way to fractionate a complex sample is to introduce a combination of chromatographic techniques before MS analysis. This method, referred to as multidimensional protein identification technology (MudPIT) [39], has been used to conduct a shotgun survey of metabolic pathways in the leaves, roots, and developing seeds of rice [36]. Extracted proteins were first digested with a lysine-specific endoprotease and then with trypsin to generate a complex mixture of peptides. The peptides were fractionated on a strong cation exchanger that was positioned immediately upstream of a reversed-phase column. Peptide fractions were selectively eluted from the cation-exchange column using stepwise increases in salt concentration and then separated on a reversed-phase column before mass spectrometric analysis. Results from the MudPIT analysis were compared with those obtained by 2DE–MS. Using 2DE–MS, protein extracts from leaves, roots, and seeds were found to contain, respectively, 348, 199, and 152 unique proteins in each fraction. In contrast, MudPIT was able to identify 867, 1292, and 822 unique proteins in leaves, roots, and seeds, demonstrating the superior detection efficiency of this technique. A total of 165 proteins were, however, detected in 2DE analysis only, which supports the complementary nature of the different proteomic technologies. The proteins were assigned to 16 different functional categories, although most (33 %) were unknown. The second most abundant functional category (21 %) included proteins involved in primary metabolism, which is in agreement with the functional distribution of enzymes predicted from the rice genome project [13]. Of 2528 detected proteins, only 189 were expressed in all three tissues, with the vast majority having tissue-specific expression patterns [36].

Other shotgun studies involved examination of protein profiles in maize chloroplasts during the greening of leaf tissue after de-etiolation [38] and demonstration of the high similarity between the profiles of chloroplast thylakoid lumen proteins of two different species, *Arabidopsis* and spinach [40]. On the basis of the proteins identified strong inferences regarding the signaling motifs required to localize proteins in chloroplasts could be made. Although *in silico* analysis is a valuable tool for predicting the protein profile of an organelle or subcellular compartment, such approaches are error-prone and require experimental verification [34, 40].

Using *Arabidopsis* as a model a functional proteomic approach was used to identify a membrane binding protein, AnnAt1, involved in osmotic stress response [41]. Root microsomal fractions were isolated from plants grown in liquid medium lacking, or supplemented with, salt. The extracts were run on 2DE gels and protein spots with greater than twofold changes in intensity were characterized by MALDI-TOF MS analysis. One protein, AnnAt1 was probably translocated from the cytosol to the membrane on increased salt exposure. The characterization of *AnnAt1* T-DNA mutants coupled to experiments involving treatment with ABA, Ca^{2+}, or EGTA supported the hypothesis that AnnAt1 plays a

role in ABA-mediated stress response in plants.

Several other very new technologies are available for performing proteomic research [18, 42, 43] and some have not been extensively employed in plant biology studies. There are new detection methods, proteomic technologies are being developed in an array format, and there is increasing focus on protein–protein interactions, post-translational modification, and elucidation of three-dimensional protein structure. In terms of plant science, proteomics is a field that is in its infancy and when it matures will certainly provide a wealth of evidence enabling a better understanding of biological systems in general.

3.6 Metabolomics

Metabolites are the products of interrelated biochemical pathways and changes in metabolic profiles can be regarded as the ultimate response of biological systems to genetic or environmental changes [44]. Metabolomics is defined as the systematic survey of all the metabolites present in a plant tissue, cell, or subcellular compartment under defined conditions, and can be further subdivided into three categories based on the type of study being performed. *Targeted metabolomics* involves examination of the effects of a genetic alteration or change in environmental conditions on particular metabolites [45]. Sample preparation is focused on isolating and concentrating the compound of interest to minimize detection interference from other components in the original extract. *Metabolite profiling* refers to a qualitative and quantitative evaluation of metabolite collections, for example those found in a particular pathway, tissue, or cellular compartment [46]. Finally, *metabolic fingerprinting* is a high-throughput method that focuses on collecting and analyzing data from crude extracts to classify whole samples rather than separating individual metabolites [47, 48].

Gas chromatography (GC)–MS or LC–MS are the tools of choice for generating high-throughput data for identification and quantification of small-molecular-weight metabolites [47]. Capillary electrophoresis (CE) is an alternative method which separates particular types of compound more efficiently and can be coupled with MS or other types of detectors. Nuclear magnetic resonance (NMR), infrared (IR), ultraviolet (UV), and fluorescence spectroscopy can be used as alternative means of detection, often in parallel with MS [47]. Time-of-flight MS technology has also been employed in metabolite analysis and provides a means of high sample-throughput. In the end, a combination of methods enables analysis of a broad range of metabolites. The separation and detection strategies used in plant metabolomics have been reviewed in detail [47, 49].

The generation of reproducible and meaningful metabolomic data requires great care in the acquisition, storage, extraction, and preparation of samples [44]. The true metabolic state of samples must be maintained and additional metabolic activity or chemical modification after collection must be prevented. Depending on the type of sample and the analysis performed, this can be achieved in various ways. The most common strategies are freezing in liquid nitrogen, freeze-drying, and heat denaturation to halt enzymatic activity [44]. Metabolomic experiments are typically conducted by comparing experimental plants possessing an expected metabolic modification (i.e. because of introduction of a transgene or exposure to a particular treatment) to control plants. Statistically significant

changes in metabolite levels attributable to perturbations affecting the experimental plants are identified. Natural variability in metabolite levels occurs as part of normal homeostasis in plants; thus, a high number of replicates is typically necessary to establish a statistically significant difference between experimental and control plants, especially if the differences between metabolite levels are subtle [47]. The large datasets and multitude of metabolites require computer-based applications to analyze complex metabolomic experiments. Ideally, such systems compile and compare data from a variety of separation and detection systems (i.e. GC–MS, LC–MS, and NMR) and identify alterations in metabolite levels [49]. Ultimately, gene functions can be predicted or global metabolic profiles associated with particular biological responses can be defined. Although such *in silico* systems have not yet been developed, multivariate data analysis techniques that reduce the complexity of datasets and enable more simplified visualization of metabolomic results are currently available. These include principle-components analysis (PCA), hierarchical clustering analysis (HCA), K-means clustering, and self-organizing maps (SOM) [49].

Considering the natural variability in transcript, protein, and metabolite levels in plants of the same genotype, correlations within complex fluctuating biochemical networks can be revealed using PCA and HCA [47, 50]. Metabolic networks were integrated with gene expression and protein levels using a novel extraction method recently developed whereby RNA, proteins and metabolites were all extracted from a single sample [50]. Metabolites were first extracted in organic and aqueous solvents, after which proteins and RNA were separated using a buffer/phenol extraction. Metabolites were analyzed using GC–TOF MS resulting in the quantification and identification of 652 compounds. Proteins were subjected to tryptic digestion, separation using two-dimensional LC, and detection by tandem MS. This resulted in the quantification and identification of 297 proteins under stringent conditions intended to avoid false positives. The RNA extracted was successfully used in northern-blot analysis with isopropyl malate synthase as a test probe. The overall extraction process was applied to ten independent replicates from two *Arabidopsis* genotypes, Col2 and C24, to see if different biochemical phenotypes could be detected and to evaluate general biochemical patterns. For each genotype, a subset of the 14 most abundant metabolites was integrated with a subset of 22 proteins and the data were subjected to PCA and HCA. The results of PCA showed that the two genotypes were completely separated into distinct clusters, meaning that such analysis could be used for classification purposes. Application of HCA revealed the conservation of some biochemical patterns in both genotypes, including Calvin cycle enzymes and metabolites such as sucrose and fructose. The authors suggest that such datasets coupled to quantitative transcript analyses could reveal more detailed relationships among metabolites and assign their linkage to established functions in biochemical networks [50].

Besides basic research in plant biology, metabolomics is also being developed to assess the safety of genetically modified (GM) foods [51]. One major concern about the use of GM foods is that the molecular alterations designed to produce a beneficial trait could also result in unintentionally hazardous effects. Because society demands that producers demonstrate "substantial equivalence" between transgenic and non-transgenic crop plants, metabolite profiling is expected to provide a reliable

means of detecting differences between metabolite levels and identifying potential problems [51].

3.7 Functional Genomics

The ultimate goal of large-scale sequencing projects is to assign a function to all the genes identified in a genome. On publication of the *Arabidopsis* genome sequence, less than 10 % of the potential genes had been functionally characterized. Most genes were identified on the basis of the similarity of their amino acid or nucleotide sequence to other annotated genes in public databases; thus, unknown genes must be classified as putative until the appropriate physiological experiments have been performed to demonstrate the function of the gene products. In fact, approximately 30 % of possible *Arabidopsis* genes could not even be assigned a putative function because their sequences were distinct from previously characterized genes [8]. Genomic approaches designed to determine the biological function of an unknown gene can be classified into one of two strategies: *forward genetics* and *reverse genetics* (Fig. 3.1). Forward genetics refers to the use of natural or artificially generated mutants with characteristics distinct from wild-type plants to clone the gene responsible for the mutant phenotype. In contrast, reverse genetics starts with a specific sequenced gene of unknown function that is "knocked-out" using directed mutagenesis to evaluate the resulting change in phenotype.

3.7.1 Forward Genetics

Although altered phenotypes can occur naturally, mutants can be generated by treating seeds from wild-type plants with chemicals such as ethylmethanesulfonate (EMS), or by physical treatment, for example with ionizing radiation (UV, X-rays) [52]. When an interesting phenotype has been identified, the gene responsible can be isolated if sufficiently detailed genetic and physical maps are available for the plant species of interest. Greater map densities increase the probability that a gene of interest is close to a molecular marker. The genetic position of the gene is determined by monitoring the co-inheritance of known markers and the phenotype of interest. Markers proximal to the gene of interest are used to identify segments of the genome contained within YAC or BAC libraries. Localization of the gene of interest between two markers enables isolation of the intervening DNA from a wild-type plant and insertion of the entire fragment into the mutant line to test for reversion to the wild-type phenotype (i.e. complementation). If complementation is positive, the genomic fragment can be subdivided until the specific gene responsible for the phenotype is identified [53]. This strategy was used to clone the *ripening-inhibitor* (*rin*) gene from tomato which codes for a transcription factor that regulates the expression of genes involved in fruit ripening [54]. The *rin* mutation arose spontaneously during breeding and results in fruit that do not ripen even when exposed to ethylene. Molecular mapping established RFLP markers positioned close to the *rin* locus on chromosome 5 and facilitated isolation of a 365 kb YAC clone [55]. A library of subclones derived from this YAC was screened further to isolate a smaller fragment shown to map close to the *rin* locus. A modification of the strategy involved using the YAC clone as a probe to screen cDNA libraries from normal and mutant fruit, which identified cDNA with altered transcripts in the mutant plants. Expression of

the wild-type form of one isolated gene in mutant plants led to full complementation (i.e. normal ripening), confirming the identity of the gene as *Rin* [54].

3.7.2 Reverse Genetics

With the availability of EST databases and large-scale genome sequences, reverse genetic approaches to assign functions to uncharacterized genes now predominate over more traditional forward genetic approaches. In mice, yeast, and *E. coli*, homologous recombination can be used to mutate, or effectively turn off (i.e. knock-out), the expression of a specific gene [56]. In flowering plants, homologous recombination occurs at low frequency and is currently not a practical means of producing gene-specific mutants [52]. The moss *Physcomitrella patens* can, however, integrate DNA efficiently by homologous recombination [57, 58]. In the future, mosses might emerge as new model systems in the advancement of plant functional genomics. Moreover, investigating the mechanism of homologous recombination in *Physcomitrella* might lead to a more efficient means of performing this technique in *Arabidopsis* and other model plants.

As illustrated in Fig. 3.3, insertional mutagenesis is an alternative strategy to homologous recombination and involves the random introduction of foreign DNA fragments into a gene, thereby inactivating it. A large group of mutagenized plant lines are created, each carrying one or more random insertions in its genome. The sequence of the inserted DNA is known and acts as a "tag" so that the gene into which it has been inserted can be isolated from a pool of mutagenized plant lines using PCR [56]. One PCR primer is specific for the tag and the other is specific for the gene of interest, enabling amplicons to be produced only when the tag is inserted into the targeted gene. For some species public databases contain DNA-sequence information flanking the insertion sites within mutagenized lines to facilitate this process [59]. One such "sequence-indexed T-DNA insertion-site database" identifies the precise locations of more than 88,000 insertions covering 21,700 of the ~29,454 predicted *Arabidopsis* genes [60]. When a line in which a desired genetic insertion has occurred has been identified, seeds can be obtained and the phenotype of the plants can be characterized. The phenotype of the mutant might not be obvious and might require in-depth analysis such as microscopic examination or molecular and biochemical analyses employing transcript, protein or metabolite profiling. Frequently a phenotype will be detected only under specific environmental conditions [56], or might not occur if the gene is represented by more than one copy in the genome.

The two methods most commonly used for insertional mutagenesis are T-DNA tagging [56] and transposon tagging [59, 61]. The advantages of T-DNA tagging include the introduction of only one or two insertions per line and the stability of the inserts through multiple generations. In contrast, transposon tagging provides a relatively simple means of generating large populations of mutants, often with insertions in multiple members of gene families residing on the same chromosome, because of the tendency of transposition events to occur over relatively short distances from the donor site [59].

Other technologies are used to create reduced or loss-of-function mutants to study the function of a particular gene (or gene family). These include post-transcriptional gene-silencing (PTGS), which includes RNA interference (RNAi) [62],

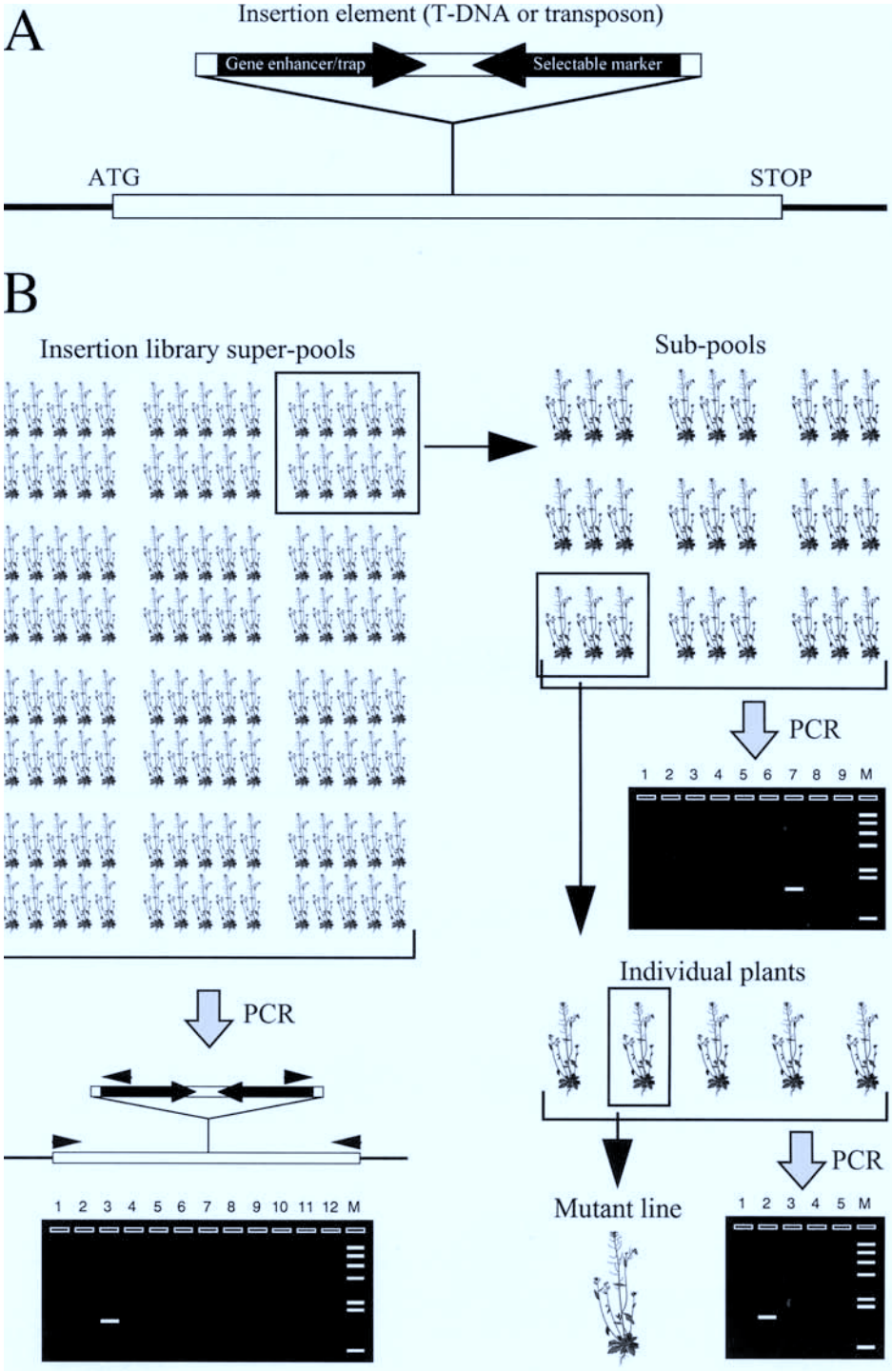

A
Insertion element (T-DNA or transposon)
Gene enhancer/trap
Selectable marker
ATG
STOP
B
Insertion library super-pools
Sub-pools
PCR
1 2 3 4 5 6 7 8 9 M
Individual plants
PCR
1 2 3 4 5 6 7 8 9 10 11 12 M
Mutant line
1 2 3 4 5 M

Fig. 3.3 Insertional mutagenesis is a direct, reverse-genetic means of determining the function of an unknown, sequenced gene. A *loss-of-function* mutation is created using an insertion element (A), usually a T-DNA from *Agrobacterium tumefaciens* or a transposon, and the resulting phenotype is subsequently characterized. Insertion elements often include a selectable marker, which is required to identify transformed plants, and gene enhancers or gene traps consisting of a minimal promoter and reporter gene, which can be activated when inserted near a transcription factor. The technique makes use of the specificity and sensitivity of PCR to screen for specific insertions in a large population of mutagenized plant lines (B). Oligonucleotide primers from the insertional element and from the gene of interest (arrowheads) are used to detect a single insertion event among the entire genome. The sensitivity of PCR enables a single insertion to be detected even in large pools (up to several thousand) of mutagen-treated plants. Finding a single mutant plant with an insertional mutation in a specific gene begins with extraction of genomic DNA from pools of mutagenized plants. PCR is first performed on superpools of DNA extracts, and positive pools are further subdivided until individual plants are identified.

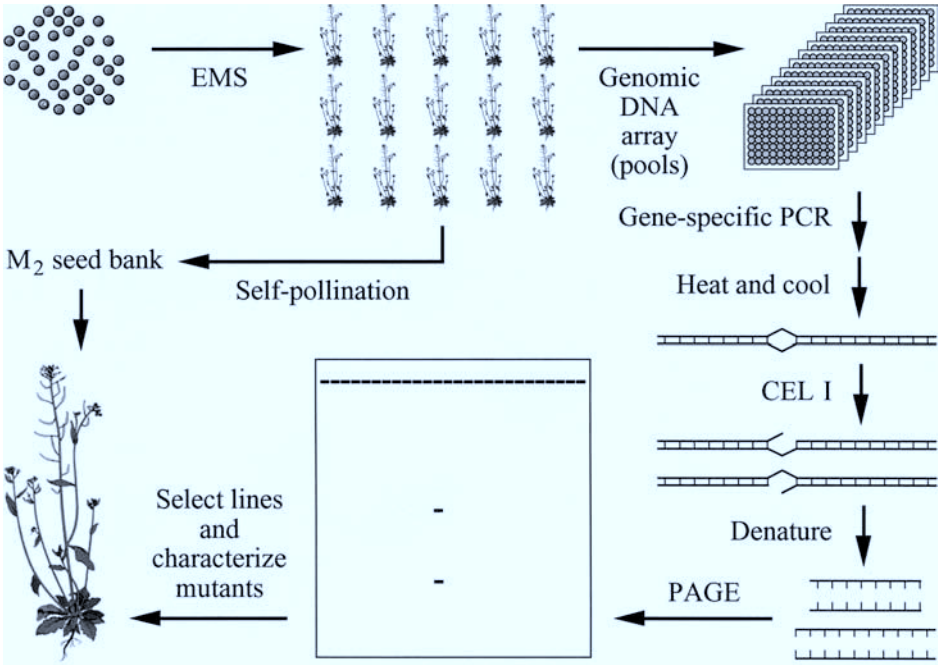


Fig. 3.4 Schematic diagram of TILLING. Seeds are treated with the mutagen ethylmethanesulfonate (EMS) and the resulting plants are catalogued and self-pollinated to create an M_2 seed bank. Leaf disks from the initial population are collected and genomic DNA samples are prepared from each plant. The genomic DNA samples are arrayed and pooled to increase screening throughput. Subsequent screening is performed on smaller pool sizes until an individual plant can be identified. At each stage point mutations are detected using gene-specific primers to amplify the target locus by PCR. Subsequently, samples are heat denatured and reannealed to generate heteroduplexes between mutant amplicons and their wild-type counterparts. Heteroduplexes are cleaved using CEL I endonuclease and visualized by polyacrylamide gel electrophoresis.

antisense RNA expression, co-suppression with sense RNA, and virus-induced gene silencing [52]. In contrast with insertional mutagenesis, one of the major advantages of RNA-mediated gene silencing is its ability to knock out entire gene families, even if they are not closely linked physically. Disadvantages include the lack of stable, inheritable phenotype, variable levels of residual gene activity, and our poor understanding of the mechanisms of some gene-silencing phenomena. In contrast to "knock-out" strategies, *over-expression* of a particular gene might also provide insight toward the determination of its function.

A new strategy with broad potential in the development of high-throughput reverse genetics in plants is TILLING (targeting-induced local lesions in genomes) [63, 64]. TILLING uses traditional chemical mutagenesis followed by high-throughput screening to identify point mutations (Fig. 3.4). Plants derived from mutagenized seeds are catalogued and self-pollinated to create an M_2 seed bank. Genomic DNA samples from each plant are arrayed and pooled to increase screening throughput. Subsequent screening is performed on smaller pool sizes until an individual plant with a point mutation in a specific gene of interest is identified. The target locus is amplified by PCR using gene-specific primers and the amplicons are heat-denatured and reannealed to generate heteroduplexes between mutant and wild-type amplicons. Heteroduplexes are cleaved using the heteroduplex-specific CEL I endonuclease and visualized by polyacrylamide gel electrophoresis [63, 64]. Ultimately, the phenotype of a plant with the desired mutation can be characterized. TILLING enables functional genomic studies to be conducted on plant species that cannot be genetically transformed.

3.8 Concluding Remarks

Considering the vast quantity of bioinformatic data currently available, the pace at which new information is being generated, and the steady stream of new technologies being developed, the future is bright for plant genomics. The development of genomic technologies beginning with the emergence of high-throughput, automated DNA sequencers have been readily applied to the plant biology field [65]. New technologies for integrating genomic, transcriptomic, proteomic, and metabolomic datasets and advancing functional genomic approaches in plants continue to be established. Coupled with the nearly complete sequencing and annotation of the *Arabidopsis* genome, and the availability of genome-wide knock-out mutants, integrated genomic analyses are creating the realistic possibility that unequivocal functions will be assigned to all plant genes in the foreseeable future [1]. Such developments in rice are not far behind, and projects in several other agronomically important plants are in progress. Genomics initiatives will continue to provide plant biologists with a wealth of resources for years to come.

References

1 Chory, J., Ecker, J.R., Briggs, S., Caboche, M., Coruzzi, G.M., et al. (2000), National Science Foundation-sponsored workshop report: "The 2010 Project." Functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them, *Plant Physiol.* **23**, 423–426.

2 Choudhary, M., Mackenzie, C., Nereng, K., Sodergren, E., Weinstock, G. M., et al. (1997), Low-resolution sequencing of *Rhodobacter sphaeroides* 2.4.1(t): Chromosome II is a true chromosome, *Microbiology-Uk* **143**, 3085–3099.

3 Puhler, A., Jording, D., Kalinowski, J., Buttgereit, D., Renkawitz-Pohl, R., et al. (2002), Genome projects of model organisms, in: *Essentials of genomics and bioinformatics.* (Sensen, C. W., Ed.), pp. 5–39. Weinheim: Wiley–VCH.

4 Ferl, R., Paul, A.-L. (2000), Genome organization and expression, in: *Biochemistry and molecular biology of plants.* (Buchannan, B. B., Gruissem, W., Jones, R. L., Eds.), pp. 312–357. Rockville, Maryland: American Society of Plant Physiologists.

5 Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., et al. (2002), A draft sequence of the rice genome (*Oryza sativa* l. Ssp. *Indica*), *Science* **296**, 79–92.

6 Unseld, M., Marienfeld, J. R., Brandt, P., Brennicke, A. (1997), The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides, *Nature Genet.* **15**, 57–61.

7 Sugiura, M., Takeda, Y. (2000), Nucleic acids, in: *Biochemistry and molecular biology of plants.* (Buchannan, B. B., Gruissem, W., Jones, R. L., Eds), pp. 260–310. Rockville, Maryland: American Society of Plant Physiologists.

8 The Arabidopsis Genome Initiative (2000), Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature* **408**, 796–815.

9 Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C., Gelbart, W. M. (1996). *An introduction to gentic analysis.* New York: W. H. Freeman and Company.

10 Koornneef, M., Vaneden, J., Hanhart, C. J., Stam, P., Braaksma, F. J., et al. (1983), Linkage map of *Arabidopsis thaliana*, *J. Heredity* **74**, 265–272.

11 Rick, C. M. (1978), Tomato, *Scientific American* **239**(2), 77.

12 Keim, P., Diers, B. W., Olson, T. C., Shoemaker, R. C. (1990), RFLP mapping in soybean – association between marker loci and variation in quantitative traits, *Genetics* **126**, 735–742.

13 Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R. L., et al. (2002), A draft sequence of the rice genome (*Oryza sativa* l. Ssp. *japonica*), *Science* **296**, 92–100.

14 Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., et al. (2002), The genome sequence and structure of rice chromosome 1, *Nature* **420**, 312–316.

15 Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., et al. (2002), Sequence and analysis of rice chromosome 4, *Nature* **420**, 316–320.

16 The Rice Chromosome 10 Sequencing Consortium (2003), In-depth view of structure, activity, and evolution of rice chromosome 10, *Science* **300**, 1566–1569.

17 Rudd, S. (2003), Expressed sequence tags: Alternative or complement to whole genome sequences?, *Trends Plant Sci.* **8**, 321–328.

18 Zhu, H., Bilgin, M., Snyder, M. (2003), Proteomics, *Annu. Rev. Biochem.* **72**, 783–812.

19 Ronning, C. M., Stegalkina, S. S., Ascenzi, R. A., Bougri, O., Hart, A. L., et al. (2003), Comparative analyses of potato expressed sequence tag libraries, *Plant Physiol.* **131**, 419–429.

20 Van Der Hoeven, R., Ronning, C., Giovannoni, J., Martin, G., Tanksley, S. (2002), Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing, *Plant Cell* **14**, 1441–1456.

21 Kuhl, J. C., Cheung, F., Yuan, Q., Martin, W., Zewdie, Y., et al. (2003), A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders *Asparagales* and *Poales*, *Plant Cell* **16**, 114–125.

22 Bahalerao, R., Keskitalo, J., Sterky, F., Erlandsson, R., Bjorkbacka, H., et al. (2003), Gene expression in autumn leaves, *Plant Physiol.* **131**, 430–442.

23 Vilaine, F., Palauqui, J.-C., Asmselem, J., Kusiak, C., Lemoine, R., et al. (2003), Towards deciphering phloem: A transcriptome analysis of the phloem of *Apium graveolens*, *Plant J.* **36**, 67–81.

24 Lange, B. M., Wildung, M. R., Stauber, E. J., Sanchez, C., Pouchnik, D., et al. (2000), Probing essential oil biosynthesis and secretion by functional evaluation of expressed sequence tags from mint glandular trichomes, *Proc. Nat. Acad. Sci. USA* **97**, 2934–2939.

25 Gang, D. R., Wang, J., Dudareva, N., Nam, K. H., Simon, J. E., et al. (2001), An investigation of the storage and biosynthesis of phenylpropenes in sweet basil, *Plant Physiol.* **125**, 539–555.

26 Guterman, I., Shalit, M., Menda, N., Piestun, D., Dafny-Yelin, M., et al. (2002), Rose scent: Genomics approach to discovering novel floral fragrance-related genes, *Plant Cell* **14**, 2325–2338.

27 Close, T. J., Wanamaker, S., Caldo, R. A., Turner, S. M., Ashlock, D. A., et al. (2004), A new resource for cereal genomics: 22k barley genechip comes of age, *Plant Physiol.* **134**, 960–968.

28 Rabbani, M. A., Maruyama, K., Hiroshi, A., Khan, A., Katsura, K., et al. (2003), Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cDNA microarray and RNA gel-blot analysis, *Plant Physiol.* **133**, 1755–1767.

29 Hui, D., Iqbal, J., Lehmann, K., Gase, K., Saluz, H. P., et al. (2003), Molecular interactions between the specialist herbavore *manduca sexta* (lepidoptera, sphingidae) and its natural host *Nicotiana attenuata* v. Microarray analysis and further characterization of large-scale changes in herbivore-induced mRNAs, *Plant Physiol.* **131**, 1877–1893.

30 Maruyama, K., Sakuma, Y., Kasuga, M., Ito, Y., Seki, M., et al. (2004), Identification of cold-inducible downstream genes of the *Arabidopsis* dreb1a/cfb3 transcription factor using two microarray systems, *Plant J.* **38**, 982–993.

31 Becker, J. D., Boavida, L. C., Carneiro, J., Haury, M., Feijo, J. A. (2003), Transcriptional profiling of *Arabidopsis* tissues reveals the unique characteristics of the pollen transcriptome, *Plant Physiol.* **133**, 713–725.

32 Nakazono, M., Qui, F., Borsuk, L. A., Schnable, P. S. (2003), Laser-capture microdissection, a tool for the global analysis of gene expression in specific plant cell types: Identification of genes expressed differentially in epidermal cells of vascular tissues of maize, *Plant Cell* **15**, 583–596.

33 Price, J., Laxmi, A., St Martin, S. K., Jang, J.-C. (2004), Global transcription profiling reveals multiple sugar signal transduction mechanisms in *Arabidopsis*, *Plant Cell* **16**, 2128–2150.

34 Baginsky, S., Gruissem, W. (2004), Chloroplast proteomics: Potentials and challenges, *J. Exp. Bot.* **55**, 1213–1220.

35 Zhu, W., Schleuter, S. D., Volker, B. (2003), Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping, *Plant Physiol.* **132**, 469–484.

36 Koller, A., Washburn, M. P., Lange, B. M., Andon, N. L., Deciu, C., et al. (2002), Proteomic survey of metabolic pathways in rice, *Proc. Nat. Acad. Sci. USA* **99**, 11969–11974.

37 Ferro, M., Salvi, D., Riviere-Rolland, H., Vermat, T., Seigneurin-Berny, D., et al. (2002), Integral membrane proteins of the chloroplast envelope: Identification and subcellular localization of new transporters, *Proc. Nat. Acad. Sci. USA* **99**, 11487–11492.

38 Lonosky, P. M., Zhang, X., Honavar, V. G., Dobbs, D. L., Fu, A., et al. (2004), A proteomic analysis of maize chloroplast biogenesis, *Plant Physiol.* **134**, 560–574.

39 Whitelegge, J. P. (2002), Plant proteomics: Blasting out of a mudpit, *Proc. Nat. Acad. Sci. USA* **99**, 11564–11566.

40 Schubert, M., Petersson, U. A., Haas, B. J., Funk, C., Schroder, W. P., et al. (2002), Proteome map of the chloroplast lumen of *Arabidopsis thaliana*, *J. Biol. Chem.* **277**, 8354–8365.

41 Lee, S., Lee, E. J., Yang, E. J., Lee, J. E., Park, A. R., et al. (2004), Proteomic identification of annexins, calcium-dependent membrane binding proteins that mediate osmotic stress and abscisic acid signal transduction in *Arabidopsis*, *Plant Cell* **16**, 1378–1391.

42 Kersten, B., Burkle, L., Kuhn, E. J., Giavalisco, P., Konthur, Z., et al. (2002), Large-scale plant proteomics, *Plant Mol. Biol.* **48**, 133–141.

43 De Hoog, C. L., Mann, M. (2004), Proteomics, *Annu. Rev. Genomics Hum. Genet.* **5**, 267–293.

44 Fiehn, O. (2002), Metabolomics – the link between genotypes and phenotypes, *Plant Mol. Biol.* **48**, 155–171.

45 Verdonk, J.C., De Vos, C.H.R., Verhoeven, H.A., Haring, M.A., Van Tunen, A.J., et al. (2003), Regulation of floral scent production in petunia revealed by targeted metabolomics, *Phytochemistry* **62**, 997–1008.

46 Burns, J., Fraser, P.D., and Bramley, P.M. (2003), Identification and quantification of carotenoids, tocopherols and chlorophylls in commonly consumed fruits and vegetables. *Phytochemistry* **62**, 939–947.

47 Weckwerth, W. (2003), Metabolomics in systems biology, *Annu. Rev. Plant Biol.* **54**, 669–689.

48 Johnson, H. E., Broadhurst, D., Goodacre, R., Smith, A. R. (2003), Metabolic fingerprinting of salt-stressed tomatoes, *Phytochem.* **62**, 919–928.

49 Sumner, L. W., Mendes, P., Dixon, R. A. (2003), Plant metabolomics: Large-scale phytochemistry in the functional genomics era, *Phytochem.* **62**, 817–836.

50 Weckwerth, W., Wenzel, K., Fiehn, O. (2004), Process for the integrated extraction identification, and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks, *Proteomics* **4**, 78–83.

51 Kuiper, H. A., Kok, E. J., Engel, K. H. (2003), Exploitation of molecular profiling techniques for GM food safety assessment, *Curr. Opin. Biotech.* **14**, 238–243.

52 Holtorf, H., Guitton, M. C., Reski, R. (2002), Plant functional genomics, *Naturwissenschaften* **89**, 235–249.

53 Somerville, C. R. (1993), New opportunities to dissect and manipulate plant processes, *Phil. Trans. Royal Soc. London Series B-Biol. Sci.* **339**, 199–206.

54 Vrebalov, J., Ruezinsky, D., Padmanabhan, V., White, R., Medrano, D., et al. (2002), A MADS-box gene necessary for fruit ripening at the tomato *ripening-inhibitor* (*rin*) locus, *Science* **296**, 343–346.

55 Giovannoni, J. J., Noensie, E. N., Ruezinsky, D. M., Lu, X. H., Tracy, S. L., et al. (1995), Molecular-genetic analysis of the *ripening-inhibitor* and *non-ripening* loci of tomato – a first step in genetic map-based cloning of fruit ripening genes, *Mol. Gen. Genet.* **248**, 195–206.

56 Krysan, P. J., Young, J. C., Sussman, M. R. (1999), T-DNA as an insertional mutagen in *Arabidopsis*, *Plant Cell* **11**, 2283–2290.

57 Puchta, H. (2002), Gene replacement by homologous recombination in plants. *Plant Mol. Biol.* 48:173–182.

58 Hanin, M., Paszkowsky, J. (2003), Plant genome modification by homologous recombination. *Curr. Opin. Plant Biol.* **6**, 157–162.

59 Parinov, S., Sevugan, M., De Ye, Yang, W.-C., Kumaran, M., et al. (1999), Analysis of flanking sequences from dissociation insertion lines: A database for reverse genetics in *Arabidopsis*, *Plant Cell* **11**, 2263–2270.

60 Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., et al. (2003), Genome-wide insertional mutagenesis of *Arabidopsis thaliana*, *Science* **301**, 653–657

61 Martienssen, R. A. (1998), Functional genomics: Probing plant gene function and expression with transposons, *Proc. Nat. Acad. Sci. USA* **95**, 2021–2026.

62 Waterhouse, P. M., Helliwell, C. A. (2003), Exploring plant genomes by RNA-induced gene silencing. *Nat. Rev. Genet.* 4:29–38.

63 Henikoff, S., Comai, L. (2003), Single-nucleotide mutations for plant functional genomics, *Annu. Rev. Plant Biol.* **54**, 375–401.

64 Henikoff, S., Till, B. J., Comai, L. (2004), TILLING. Traditional mutagenesis meets functional genomics, *Plant Physiol.* **135**, 630–636.

65 Newman, T., De Bruijn, F.J., Green, P., Keegstra, K., Kende, H., et al. (1994), Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. *Plant Physiol.* **106**, 1241–1255.

4
Human Genetic Diseases

Roger C. Green

4.1
Introduction

4.1.1
The Human Genome Project: Where Are We Now and Where Are We Going?

On April 14, 2003 the International Human Genome Sequencing Consortium announced the successful completion of the Human Genome Project (HGP) more than two years ahead of schedule. That month also marked the 50th anniversary of the publication of Watson and Crick's Nobel Prize-winning description of the structure of DNA [1]. This was a modest two-page report that ended with the famous understatement "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material".

The main effort of the Human Genome Project has been the production of the reference sequence of the human genome. Since the first version of the sequence was completed in June 2000, researchers worked to convert the "draft" sequence into a "finished" sequence which has less than one error per 10,000 bases. The finished sequence covers about 99 % of the human genome's gene-containing regions. The only remaining gaps involve small regions that cannot be sequenced with current technology.

4.1.1.1
What Have We Learned?

An unexpected finding to emerge from the sequence of the human genome is that we seem to have far fewer protein-coding genes than previously suspected – approximately 30,000, compared with a figure of approximately 100,000 frequently cited previously. In 2002 the draft sequence of the mouse genome was published [2]. We can learn a lot about ourselves by studying the mouse – we have approximately the same number of genes as the mouse and other mammals, and now know we have only a few thousand more genes than *Arabidopsis*, a roadside weed. This is a humbling discovery [3].

At least 99 % of mouse genes have a corresponding gene in humans. The proteins that are encoded by these genes are often very similar in mice and humans. It is therefore likely that differences in gene or protein structure alone cannot account for interspecies differences (or indeed the differences between individuals of the same

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

species). So what *does* account for the differences between mice and men?

Non-coding RNA One clue to this puzzle comes from a closer examination of that part of the genomes that does *not* code for protein. The vast majority of our DNA (approx. 99 %) lies outside the protein-coding exons (Fig. 4.1 and Box 4.1). The rate at which mutations accumulate in this non-coding DNA is generally much higher that the mutation rate in exons. This is consistent with the hypothesis that this is really "junk" DNA. (This is because mutations that occur in exons are overwhelmingly likely to have a negative rather than positive effect on our ability to procreate and are therefore weeded out by natural selection.) It has been discovered that the sequence of a proportion of this "junk" DNA is actually well conserved between the mouse and human genomes, however [4]. Because the evolutionary paths of mice and humans diverged approximately 75 million years ago, the fact that these DNA regions have changed very little implies there have been

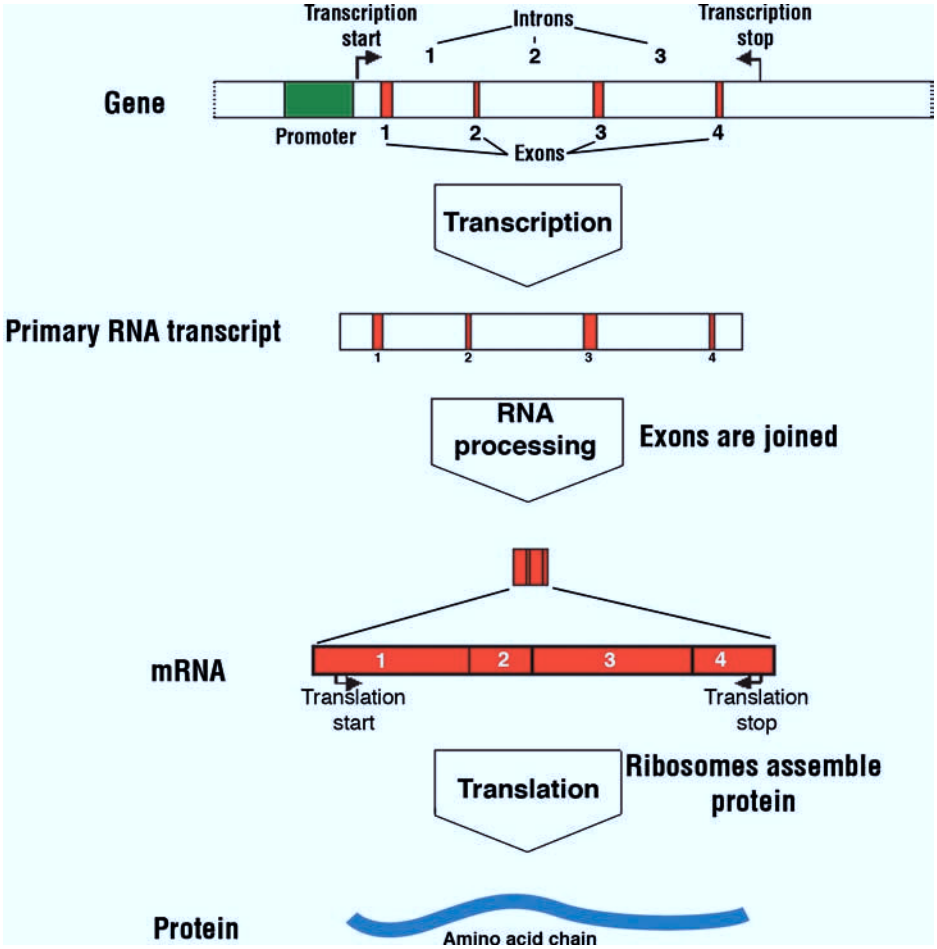


Fig. 4.1 The genome in action.

Box 4.1 What we know about the human genome

The nuclear genome is about 3 billion bp long.

- There are 2 metres of DNA in every nucleus.
- There are 24 distinct chromosomes
  - 22 autosomes (arranged in pairs for a total of 44).
  - X chromosome (2 in females, 1 in males).
- We have about 30,000 protein-coding genes
- Only 1.5 % of our DNA is in the exons of protein-coding genes
- Humans also have tens of thousands of non-coding RNA genes.
  - We have almost no idea of their function.
- The human genome is similar in structure and size to the mouse genome.
- The base sequence of any two humans is about 99.9 % identical.
- The base sequence of chimpanzee DNA is about 99 % identical with human DNA.

Darwinian selection pressures acting to eliminate mutations in these regions. In other words, this is not junk DNA after all – it must play an important role.

The conserved nonprotein-coding DNA is found both within introns and in the intergenic DNA [5]. It has been suggested that there are tens of thousands of expressed regions of this type and that non-coding RNA (ncRNA) has a major role in the regulation of the expression of the protein-coding genes [4, 6–8]. We know almost nothing about the function of these RNA genes but, from a human health perspective, it seems reasonable to assume that if they are required for the normal development and functioning of human tissues then we should expect to find diseases associated with mutations in these genes. Expect to see a major effort over the next few years to characterize these novel genes and to assess their role in maintaining human health.

4.2 Genetic Influences on Human Health

All diseases have a genetic component. The term "heritability" is used to define the proportion of the causation of a disease – or any other characteristic for that matter –

Box 4.2 Examples of single-gene disorders

Autosomal recessive:

- Cystic fibrosis (CF) is quote common in some populations (1 in 2000 North European births). Deficiency of a chloride-transporter protein results in the accumulation of thick mucus in the lungs; this leads to repeated pulmonary infections and eventual pulmonary failure.
- Sickle-cell anemia is a disorder common in individuals of African heritage. Heterozygous carriers are more resistant to malaria but homozygotes have an abnormal hemoglobin which causes the red blood cells to lyse, especially under conditions of reduced oxygen pressure.

Autosomal dominant:

- Huntington disease is a progressive neurodegenerative disorder which appears in mid-life and leads to uncontrolled movements, loss of intellectual capacity, and emotional disturbance. It is caused by an abnormal increase in a triplet repeat sequence on chromosome 4. There is no effective treatment. Presymptomatic DNA testing can show who will develop the disease later in life, or potentially pass it on to children.

X-linked recessive:

- Duchenne muscular dystrophy is a fairly common disorder which, like other X-linked recessive traits, is normally restricted to males (1 out of 3500 male births). The presence of an abnormal muscle protein leads to a gradual loss of muscle function beginning in infancy. Respiratory failure and pulmonary infections usually lead to death before the age of 20.

that has a genetic cause. Traditionally, the study of genetics has been focused on diseases with a heritability close to 1.0. These are the classical inherited (Mendelian) diseases which are caused by a mutation in a single gene that can be vertically transmitted from generation to generation. Traits carried on the autosomes can be dominant – requiring the inheritance of only one copy of the mutant gene – or recessive – when a mutated gene must be inherited from both parents who themselves are usually unaffected "carriers"(Box 4.2). Most single-gene defects are rare conditions, but because there are several thousand known single gene disorders their combined incidence is estimated to be approximately 1 in 200 births.

4.3 Genomics and Single-gene Defects

The past 15 years have seen major advances in our understanding of the precise genetic defects that lead to a large range of single-gene (Mendelian) diseases. The genes and mutations responsible for hundreds of inherited diseases have been discovered. The genetic basis of many of the more common Mendelian traits has now been elucidated – although what is defined as "common" depends very much on the origin of the population being studied. There is now a long list of rare conditions gradually succumbing to the tools of the genomic age.

4.3.1 The Availability of the Genome Sequence Has Changed the Way in which Disease Genes Are Identified

The HGP has provided us with the location and base sequence of most genes, but not all. At this stage however, we still do not know the function of many genes. From the base sequence alone we are usually able to predict the amino acid structure of the resulting protein. By comparing the structure of a novel protein to the structure of known proteins, we can often make an educated guess about the cellular location and functional role of the new protein. We might still be ignorant of any role that a particular gene plays in a disease process, however.

By examining the complement of messenger RNA (mRNA, Fig. 4.1) produced in different tissues, we can determine in which range of tissues each gene is expressed. This can be done by scanning tissue-specific cDNA libraries for specific gene sequences or by using DNA chip technology to examine thousands of gene sequences simultaneously [9]. These approaches have greatly facilitated the linking of genes to diseases. The basic strategy for identifying the responsible genetic change has remained largely unaltered and consists of the steps:

1. Families segregating the disease must be identified and clinically characterized.
2. Statistical methods are used to identify a small region of the genome where the offending gene is located. Such methods include:
   a. linkage analysis
   b. linkage disequilibrium mapping
   c. homozygosity mapping for recessive traits
3. Searching the genes in the candidate region for disease-causing mutations.

Although the first two steps remain time-consuming, the third step has been greatly facilitated by the knowledge gained from the Human Genome Project. For example, the gene responsible for Huntington disease was mapped to the end of chromo-

some 4 in 1983 [10], but it took a further 10 years to complete the third step [11]. Now however, when the genetic locus has been identified, it can take only a matter of weeks to complete the identification of the gene.

In the pre-genomic era, researchers would have to painstakingly construct their own "physical maps" of the candidate region. This entailed separating and cloning hundreds of individual pieces of DNA and could take several years of work. Nowadays the same or better information can be obtained with a few "mouse clicks". As a result, it has been estimated that the cost of cloning (identifying) a single disease-associated gene has dropped from several million dollars in the early days to approximately $100,000 today [12].

4.3.1.1 Positional Candidate Gene Approach

We now have a list – although just how complete is still not clear – of all the genes located within a candidate region and we know their "normal" (reference) sequence. As we gain more insight into the function of these genes, and in which tissues they are expressed, we are able to select those that might plausibly be involved in the disease process we are studying. Eventually the gene from an affected person must be sequenced and compared with the reference sequence of that gene. If a difference which would alter protein function is found, it must be established that every affected person in the family carries the same mutation, and that the mutation is not found in unaffected members of the family nor in the general population. This combination of genetic mapping followed by examination of candidate genes within the region is known as the "positional candidate" method of locating disease genes.

Occasionally we still discover a gene that has not previously been recognized by analysis of the human genome sequence. For instance, a gene responsible for hereditary sensory and autonomic neuropathy was recently mapped to a fairly small region (just over a million base pairs) of chromosome 12 which was thought to contain seven genes. No disease-causing mutation could be found in any of these genes, however. Subsequently a novel single-exon gene, now termed HSN2, was found, hiding within an intron of one of the seven known genes [13]. The protein encoded by this gene bears no resemblance to any other known protein.

4.3.1.2 Direct Analysis of Candidate Genes

The main drawback to the positional candidate method is that it still relies on some type of preliminary genetic mapping using families that have a particular disease. This mapping process is very time-consuming and thus expensive, because it involves collecting data from many family members and accessing hospital records and other information sources. As we develop more information about the human genome, and as technological advances in sequence-identification continue, we might be able to jump past the mapping step altogether in many cases. We now know the DNA sequence of most genes. We will gradually gain an understanding of the functional significance of most of these genes. We will have information about variations in the DNA sequences of the genes within defined populations. We will know in which tissues, and possibly under what circumstances, these genes are expressed to produce proteins. With this additional information, and with some knowledge of the underlying pathophysiology of the disease, it will be possible to identify the candidate genes most likely to be involved in a genetic disease. The DNA of candidate genes from

affected patients will then be sequenced to look for evidence of mutations. With further improvements in DNA sequencing technology, including use of gene-chip microarrays, it will be possible to scan hundreds of candidate genes in a relatively short time. Thus from a number of candidates, one or more disease-associated genes will be identified. Successes using this approach have already been published [14, 15].

4.3.2 Applications in Human Health

The discovery of the genetic basis of a disease has several consequences. One that has immediate application is the provision of DNA-based diagnosis and risk assessment.

4.3.2.1 Genetic Testing

As soon as a causative mutation has been discovered it is possible to design a simple test that can detect the presence of this mutation in DNA obtained from a blood sample or from other tissue. The purpose of genetic testing is usually to identify carriers of genetic defects that could predispose the carrier, or the carrier's children, to an inherited disease. (Less commonly, DNA testing is used to help make a clinical diagnosis, for instance in sickle-cell disease, hemochromatosis, or cystic fibrosis.) There are many possible reasons for conducting a genetic test. For dominant traits it might be to predict who might be susceptible to a disease later in life, so that the appropriate interventions can take place to prevent or ameliorate the condition. Such interventions might take the form of clinical screening programs to detect the disease at an early stage [16], of prophylactic surgery, or of pharmaceutical intervention [17, 18]. For recessive traits, carrier screening is conducted to determine if a couple is at risk of bearing an affected child, or to determine prenatally the genotype of the fetus. For conditions in which there is no effective intervention, for example Huntington disease, the demand for genetic testing is comparatively low.

A major obstacle to implementation of genetic testing is the occurrence of genetic heterogeneity. An inherited disease can often be caused by a mutation in any of several different genes (locus heterogeneity) and usually there will be a large number of different possible mutations in each gene (allelic heterogeneity). There are also several difficult ethical, social, and legal issues surrounding the subject of genetic testing [19, 20]. Further complications arise if considering testing large groups of the population (population screening) [21].

Locus Heterogeneity The extreme example of locus heterogeneity is retinitis pigmentosa, which is a progressive eye disease that leads to loss of vision. The pattern of inheritance can be dominant, recessive, or sex-linked. More that 35 chromosomal regions have been linked to this disease and 30 genes have already been identified [22].

Locus heterogeneity often results from a defect in a biochemical pathway that requires several different proteins to function properly. Thus a mutation in any one of several genes has the potential to disrupt the pathway and produce the same disease phenotype. For example, the most common inherited form of cancer, hereditary non polyposis colon cancer (HNPCC), results from a defect in one of the DNA-repair systems within the cell. Several different proteins are required for this specific repair mechanism to work correctly. It is known that HNPCC can result from a mutation in at least five of these proteins, yet the clinical features are similar in all families [23, 24], so the particular gene involved in each fam-

ily must be identified before testing is possible.

Allelic Heterogeneity Sickle-cell anemia was one of the first inherited diseases in which the molecular basis was discovered, a mutation in the β subunit of hemoglobin [25] (Box 4.2). It was subsequently found that a single DNA mutation was responsible for virtually all cases, and that detection of this mutation could be used to identify carriers of the sickle-cell trait. Sickle-cell disease seems exceptional in this regard, however, and there will usually be many different mutations found in each gene (allelic heterogeneity). In practice, this means that each family could potentially have a different mutation and that the precise mutation in the family must be determined before DNA-based testing can be implemented. Sometimes a "founder effect" can be observed in which several affected families within a particular region, or possibly in diverse regions, have the same mutation by virtue of sharing a common ancestor [26–28].

4.3.3 Gene Therapy

As soon as scientists realized that a defective gene could cause a disease, they began to speculate about the possibility of fixing the broken gene. About 15 years ago, when we actually began to identify and sequence disease-associated genes, predictions were made that we would see some kind of gene therapy within several years. This proved to be wishful thinking. Despite choosing diseases where we had a good understanding of the underlying pathophysiology and where we could access the critical cell or tissue, there have been few real advances in gene therapy of inherited diseases. After a decade of unsuccessful attempts to cure one form of severe combined immunodeficiency disease by using gene therapy, researchers apparently achieved success in treating another genetic type of the same disease [29, 30]. Sadly however, within a few years, two of the ten children treated developed leukemia-like disease [31]. Inserting the new gene into the patients' own genomes had activated a cancer-causing gene that was located nearby. This setback and the earlier death of another patient in a gene-therapy trial [32] has led to a decline in gene-therapy research. Thus treatment of defective genes still lies some indeterminable time in the future. More promising however is the prospect of treating non-inherited diseases using genetic material as the therapeutic agent [33, 34].

4.4 Genomics and Polygenic Diseases

The terms "polygenic", "multifactorial" and "complex" are used somewhat interchangeably to describe diseases that involve changes to more than one gene and that, in most cases, also involve non-genetic (environmental) factors. Most of the common diseases that afflict the human species can be so classified. The success of the HGP will have a much greater impact on the elucidation of the etiology of common multifactorial diseases than it will on that of the monogenic diseases. Sorting out the genetic factors involved in common complex disorders such as schizophrenia, obesity, or diabetes, remains a challenging proposition. For any multifactorial disease we are faced with the need to identify multiple genes, each possibly having only a small effect on the outcome. It is the combination of mutations in several genes – coupled with environmental factors – that is responsible for the onset of the disease. Indeed even the term "muta-

tion" might not be appropriate to describe the genetic variations involved. Whereas the mutation in single-gene diseases often has a profound effect on the function of the encoded protein and is rare in the general population, it is likely that the "mutations" in the polygenic diseases will be found to result in much more subtle changes in protein function and to be much more common. We should, perhaps, use instead the term "polymorphisms" to describe these variations in gene structure. The effects of these polymorphic variants on protein function will usually be quantitative rather than qualitative – for example the binding coefficient of an enzyme–substrate complex or the affinity of a growth factor for a receptor might be affected.

Offshoots of the HGP might also begin to clarify the role which the variable penetrance of genetic variants seems to play in the elusive relationship between genotype and phenotype in multifactorial diseases. The penetrance of a genetic variant can be defined as the proportion of individuals who carry such a variant for which a change in phenotype is actually observed. For the mutations involved in Mendelian traits the penetrance is usually high, sometimes approaching 100 %. In complex diseases, however, the penetrance of any one variant, in isolation, might be low. It has become apparent that linkage studies have low power to map genes with low penetrance, even for Mendelian traits, and this problem would be much more serious if attempting to map genes involved in complex traits by using linkage analysis.

An alternative method of detecting specific genetic risk factors is by means of association studies. These studies measure the occurrence of specific genetic variants in affected patients and in an unaffected control group. Association studies can be performed in several ways.

4.4.1 Candidate Genes and their Variants

The conceptually simplest approach to conducting association studies (the "direct approach") entails testing affected and control populations for the frequency of a specific gene variant which is postulated to be directly involved in the disease process. If the variant occurs in the patient population at a significantly higher frequency one can infer that this particular genetic variant contributes to the onset of the disease. The opposite may also be true – some variants might be protective.

The genome sequences from two individuals selected at random from the worldwide population differ by about 0.1 % (1 variant per 1000 bases on average) [35]. The most common type of variation is substitution of a single base with a different one. It is estimated there are more than 10 million such variant sites (called single nucleotide polymorphisms, SNPs) within the human genome where the rarer allele (variant) occurs at a frequency of at least 1 % in the general population [36, 37]. There will be very many more alleles that are rarer (less than 1 % in the population) but because of their rarity, it is unlikely these will have a significant role in the etiology of common diseases. It has, however, been estimated that out of these ten million SNPs only about 50,000 are likely to result in altered protein structure [37]. These are referred to as cSNPs (coding SNPs) and are found in only the 1.5 % of the genome that forms the exons (Fig. 4.1). The vast majority of SNPs occur in introns or intergenic regions of the genome where they are, for the most part, not subject to evolutionary selective pressure. There will still be SNPs that occur in regulatory regions that do not alter protein structure but which might have a quantitative effect on protein expression. These

Table 4.1 Examples of genetic variants identified for common diseases by use of association studies.

Disease	*Gene*	*References*
Alzheimer's disease	APOE	38
Asthma	ADAM33	39
Deep vein thrombosis	Factor V	40
Hypertriglyceridemia	APOAV	41
Inflammatory bowel disease	NOD2	42
Myocardial infarction	LTA	43
Schizophrenia	Neuregulin 1	44
Stroke	PDE4D	45
Type 1 diabetes	HLA	46
	Insulin	47
	CTLA4	48
Type 2 diabetes	PPARG	49, 50

might occur in the promoter regions of protein-coding genes or within the ncRNA genes (Sect. 4.1.1.1).

Direct association studies so far completed have tended to use only one or a few of these cSNPs, selected on the basis of the biological significance of the polymorphic candidate gene to the disease under investigation (Tab. 4.1). Sometimes regions of the genome associated with a particular disease have been discovered not by genotyping a cSNP directly involved in the disease process but by analyzing polymorphic loci which are in "linkage disequilibrium" with a nearby disease-causing allele.

4.4.2 Linkage Disequilibrium Mapping

Linkage disequilibrium (LD) can be thought of as the co-occurrence of specific alleles at two loci in the genome at a higher frequency in a population than would be predicted by random chance [51]. Thus if a particular SNP allele is found to be associated with a disease it might not be that this first SNP is directly involved in the disease process but that is in linkage disequilibrium with another allele in a second nearby gene. Linkage disequilibrium extends to more than just two adjacent loci. Alleles of several SNPs that are close together tend to be inherited together in blocks. A linear set of particular SNPs alleles in one region of a chromosome is called a haplotype. It has been found that most chromosome regions have only a few common haplotypes (defined as haplotypes with a frequency of at least 5 % in the population) rather than the much larger number which represents the theoretical number of possible combinations. These few common haplotypes seem to account for most of the variation from person to person in a population [52]. Even though a chromosome region might contain many SNPs, if there is strong linkage disequilibrium in the region the analysis of only a few "tag" SNPs provides most of the information on the genetic variation in that particular region by enabling the identification of the common haplotypes. It has also been found that regions of strong linkage disequilibrium within the genome are quite long, with half of such blocks being 44 kbases or larger in European and Asian populations [52].

4.4.2.1 The Hapmap Project

The International HapMap Project was organized in 2002 and its objective, as one

might expect, is to develop a haplotype map of the human genome. The HapMap will describe the common patterns of genetic variation in humans by characterizing the haplotypes within chromosome regions with sets of strongly associated SNPs. It will develop the key tag SNPs which will define these regions and will also note the chromosome regions where linkage disequilibrium is weak.

Chromosome regions in which a group of patients and a population-matched group of controls differ in their haplotype frequencies will identify places to look for candidate disease-associated genes. To complete a genome-wide analysis satisfactorily it is estimated that 200,000 to one million SNPs must be analyzed, which is at least an order of magnitude less than would be required in the absence of knowledge from the HapMap project [53]. Recent evidence obtained from re-sequencing of more than 100 genes from 23 Africans and 24 people of European ancestry indicates, however, that the complexity of the haplotype architecture might have been underestimated and that this might have important implications for genome-wide association studies [54].

4.4.3
Whole-genome Resequencing

Another approach to discovering the genetic basis of a common disease is to re-sequence all the genes – or indeed the entire genomes – from sets of patients with complex diseases [55]. This powerful method will have the advantage of identifying rare variants and commonly occurring polymorphisms. The technology to attempt this task is not yet within our grasp. Current rates of DNA re-sequencing are too slow to gather enough data cost-effectively. In addition, the computing power to analyze and compare whole genomes in many individuals simultaneously has also not yet been developed. We must not forget, however, that just 20 years ago any thought of sequencing the entire human genome even once seemed to be wishful thinking.

The rate at which we have so far uncovered the genetic basis of common disease is perhaps disappointing. Although it is clear that no single approach will identify all the genetic variants that contribute to human disease, we can be sure our increasing technological capabilities coupled with enhanced knowledge of genome structure will provide more rapid advances in the years to come. The next section discusses what we have learned so far about two of the most significant classes of human illness – cancer and cardiovascular disease.

4.5
The Genetic Basis of Cancer

We know what causes cancer – all cancers are caused by genetic mutations. An accumulation of DNA mutations within a single cell is required to turn that cell into one that lacks the normal controls on growth, division, and the capacity to invade other tissues. It takes time to accumulate the required number of genetic changes within a single cell, which is why cancer is usually a disease of the aged. Gain-of-function mutations are needed in proto-oncogenes, which cause an up-regulation of proteins whose normal function is to stimulate cell growth and division. In addition loss-of-function mutations are required in tumor-suppressor genes which normally act as "brakes" on cell division. Often there will also be mutations occurring in cell-adhesion molecules, genes involved in DNA repair and maintaining genetic stability, and mutations that inactivate telomerase function.

Although all cancers are caused by mutations, the predisposition to cancer is not usually inherited as a Mendelian trait. Instead, the cancer-causing mutations are a result of somatic events (that is events occurring in the normal cells of the body as opposed to the germline cells) such as spontaneous mutations, action of carcinogens, radiation damage, and the action of some viruses. Although most mutations are somatic, there are also inherited DNA mutations that can increase the risk of developing cancer. The nature of these inherited variants in the predisposition to common cancers is described in the next section. The commonest sites of cancer in many populations are the lung, colon, and breast. Because lung cancer is overwhelmingly related to tobacco smoking any inherited genetic factors have been difficult to study and little is known about them. In contrast, study of inherited forms of breast and colorectal cancer has greatly enhanced our understanding of the disease process.

4.5.1 Breast Cancer

It has long been recognized that a positive family history of breast cancer is an important risk factor for developing the disease. Highly-penetrant mutations in at least five genes are known to confer a high risk of developing breast cancer as an autosomal dominant trait [56] (Tab. 4.2). The two major genes are BRCA1 [57] and BRCA2 [58]. Women who have inherited a mutation in one of these genes are at high risk of developing breast cancer and ovarian cancer. (Men rarely develop breast cancer, but are more likely to do so if they carry a BRCA mutation.) Inherited cancer genes identified previously, such as that for the childhood eye cancer retinoblastoma [59] or for the polyposis coli form of colon cancer [60], have shed considerable light on cases of cancer in patients who had not inherited the mutated gene. Geneticists therefore expected that BRCA would do the same for the 95 % of breast cancer patients who are not BRCA carriers, but these hopes were dashed. By examining the DNA from many breast tumors it was found that the BRCA1 or BRCA2 genes were rarely mutated in sporadic (non-inherited) cases [61, 62]. Since then, however, we have learned that both BRCA genes interact with other genes and proteins to regulate cell behavior. Although BRCA gene mutations do not seem to be directly connected to sporadic cancers, evidence now suggests that defects in other parts of the BRCA pathway could be important in triggering breast cancer and other cancers. For example the CHEK2

Table 4.2 Highly penetrant breast cancer susceptibility genes.

Gene	*Function*	*Syndrome*	*References*
BRCA1	Transcription, cell-cycle control, DNA-damage repair	Familial breast/ ovarian cancer	57, 65
BRCA2	Transcription, cell-cycle control, DNA-damage repair	Familial breast/ ovarian cancer	58, 66
p53	Transcription, cell-cycle checkpoint (DNA integrity)	Li–Fraumeni	67
STL11/LKB1	Serine-threonine kinase	Peutz–Jeghers	68
PTEN	Phosphatase, tumor-suppressor	Cowden	69

gene encodes a cell-cycle checkpoint kinase that is involved in the BRCA1-p53 pathway. It has been reported that in those not carrying a BRCA mutation the CHEK2 1100delC mutation confers a twofold increased risk of breast cancer in women and a tenfold increased risk in men [63] although more recent studies question the role of CHEK2 mutations in male breast cancer at the population level [64].

Although the BRCA1 and BRCA2 proteins are not similar, they both seem to be involved in the repair of double-strand DNA breaks, chromatin remodeling, and the regulation of gene transcription [70]. More recently Hughes-Davies and colleagues have identified a protein (EMSY) that binds to the BRCA2 protein and, like BRCA2 itself, has functions associated with DNA repair and transcriptional regulation. They demonstrated that EMSY is a repressor protein which binds within the BRCA2 transcriptional activation domain and silences this function [71]. They further showed that the EMSY gene was amplified in 13 % of sporadic breast cancer and 17 % of higher-grade ovarian cancer but was rarely amplified in other tumor types. EMSY amplification was associated with worse survival, suggesting that it might be of prognostic value. The clinical overlap between sporadic EMSY amplification and familial BRCA2 deletion implies that a BRCA2-dependent pathway is involved in sporadic breast and ovarian cancer.

4.5.1.1

Cancer Risk in Carriers of BRCA Mutations

Initial estimates by the Breast Cancer Linkage Consortium put the risk of developing breast cancer at up to 85 % by 70 years of age and the risk of ovarian cancer at 42 % [72]. These estimates were obtained by studying families with very high rates of breast and ovarian cancer. Later studies, which used analysis of carriers unselected on the basis of family history, identified a significantly lower risk for breast cancer (26–60 %), with the risk in BRCA2 carriers lower than that in BRCA1 carriers [73]. Similar reductions in risk estimates for ovarian cancer were also reported. A substantial discussion ensued about which risk estimates were "correct". It now seems that both might be right. In BRCA families with multiple women affected at young ages there are probably other genetic risk factors (modifying genes) co-segregating in the families. These serve to increases the overall risk of breast cancer in these families above that attributable to BRCA mutations alone. In the absence of a strong family history however, the lower risk estimates are likely to be more useful for risk prediction.

The search for a high-penetrance putative "BRCA3" locus has not been successful [74]. The search illustrate the difficulties in identifying genes for a disease with high population prevalence. It is likely there are multiple genes still to be identified among non-BRCA1/2 families, with any one novel gene accounting only for a small proportion of such families. Modeling algorithms suggest that several common low-penetrance genes with multiplicative effects on risk might account for the residual non-BRCA1/2 familial aggregation of breast cancer [75]. The modifying effect of these gene variants might explain the previously reported differences between population-based estimates for BRCA1/2 penetrance and estimates based on high-risk families.

Nevertheless, only a few percent of all breast cancers are associated with inherited mutations in BRCA1 or BRCA2, and only about 0.1–0.4 % of women in outbred Western populations are thought to carry such mutations [76, 77]. There are a few other genes in which inherited mutations predispose to breast cancer (Tab. 4.2), but

these are much rarer still than families with BRCA mutations. But even if BRCA carriers are excluded, family history still seems a very significant risk factor. This means that the bulk of genetic variation responsible for breast cancer remains to be explained.

4.5.2 Colon Cancer

There are two well-characterized inherited colon cancer syndromes that have made significant contributions to our understanding of colorectal carcinogenesis – familial adenomatous polyposis (FAP), linked to mutations in the APC gene, and hereditary non-polyposis colorectal cancer (HNPCC), which is caused by mutations in one of a number of DNA mismatch repair genes.

4.5.2.1 Familial Adenomatous Polyposis

Familial adenomatous polyposis (also known as adenomatous polyposis coli) is a syndrome characterized by the occurrence of multiple – typically more than one hundred – adenomatous polyps in the distal region of the large bowel. These adenomas are pre-malignant neoplasms, one or more of which invariably progresses into a carcinoma. The trait is inherited in a dominant pattern and the average age at diagnosis of colon cancer is in the forties. The polyposis phenotype is easily recognized by sigmoidoscopy and the penetrance is close to 100 %. An atypical or attenuated phenotype (AFAP) is characterized by a smaller and variable number of polyps, later age at diagnosis, proximal location in the colon, and reduced penetrance. The attenuated type is more difficult to diagnose. The APC gene was identified in 1991 [78] and encodes a tumor-suppressor protein. Although familial adenomatous polyposis accounts for less than 1 % of all colon tumors, it quickly became apparent that the APC gene had a central role to play in carcinogenesis in general and in the development of colorectal carcinoma in particular. Mutations in the APC gene are found in about 80 % of all colorectal tumors [79], in about 60 % of adenomatous polyps [80] and are early events in tumorigenesis [80]. Mutations of APC are also found in tumors occurring at many other anatomic sites, for example stomach [81] and breast [82], but generally at a lower frequency than found in colon tumors. The APC protein is involved in cytoskeletal remodeling, cell–cell adhesion and cell migration [24]. It acts to regulate levels of β-catenin through the *Wnt*-signalling pathway [83]. The occurrence of APC mutations is a key element in one of the major models for tumorigenesis, the chromosomal instability (CIN) pathway. CIN tumors account for about 85 % of all colon cancer and are characterized by abnormalities in chromosome number and structure. It is believed the APC protein is important in the function of the mitotic spindle, which is derived from elements of the cytoskeleton. Any derangement of chromosome-spindle interaction has the potential to lead to mis-segregation of the chromosomes and hence to aneuploidy.

4.5.2.2 Hereditary Non-polyposis Colon Cancer

Hereditary non-polyposis colon cancer (HNPCC) is considerably more common than FAP but probably accounts for less than 5 % of colon tumors. It is probably under-diagnosed, because there is no pathognomonic feature and diagnosis depends on having a significant multi-generation history of colon or other tumors [23]. Other anatomic sites frequently involved include the uterine endometri-

um, stomach, ovary, the renal pelvis, and ureter [84]. HNPCC is caused by inherited mutations in one of a number of mismatch repair (MMR) genes which code for proteins whose function is to repair base-mismatches in DNA, especially those occurring in regions of simple repeats (mono- or dinucleotide repeats) [85]. Loss of MMR function leads to a higher rate of accumulation of mutations in other genes which then leads to carcinogenesis. Tumors from HNPCC patients are characterized by instability of their DNA in simple repeats. This is referred to as microsatellite instability (MSI). MSI is also observed in about 15 % of sporadic colon tumors. In most of these cases MMR function is lost because of epigenetic (not inherited) methylation of the promoter region of MLH1 which leads to silencing of the MLH1 gene [86]. In tumors exhibiting MSI a number of genes have been identified as key targets for mutation in their microsatellite regions. These include the transforming growth factor receptor type 2 (*TGFβRII*) [87], insulin-like growth factor II receptor (*IGFIIR*) [88], the antiapoptotic gene *Bax* [89], the cell-cycle regulator *E2F2* [88], and others [24, 90]. Mutations in these genes are involved in the progression of tumorigenesis initiated by defects in the mismatch repair pathway.

The microsatellite instability (MSI) pathway characterizes approximately 15 % of all colon tumors and is distinct from the chromosomal instability (CIN) pathway which involves APC mutations and is involved in most colon cancer.

4.5.2.3 Modifier Genes in Colorectal Cancer

Analysis of cohorts of twins reveals a relatively large effect of heritability for several forms of cancer (prostate, colorectal, and breast) suggesting that our current knowledge of cancer genetics is limited [76]. This effect is probably because of a combination of low-penetrance tumor susceptibility genes which are relatively common in the population and might confer a much higher attributable risk in the general population than rare mutations in high-penetrance cancer-susceptibility genes such as BRCA1/2, APC, and the HNPCC genes. The search for these genes is ongoing but significant progress might not be achieved until large-scale genome-wide association studies are conducted as described in Sect. 4.4. An example of the results achievable can be seen in the combined results of several studies into the role of variants of the type 1 TGFβ receptor (TGFBR1). As indicated in Sect. 4.5.2.2, the TGFβ pathway has been implicated in the development of colon cancer. Several studies have demonstrated a variant of the TGFBR1 gene which is seen at increased frequency in a variety cancer patients compared with healthy controls. Combined analysis of 12 studies shows that carriers of the TGFBR1*6A variant have a 38 % increased risk of breast cancer, a 41 % increased risk of ovarian cancer, and a 20 % increased risk of colorectal cancer [91, 92].

4.6 Genetics of Cardiovascular Disease

Cardiovascular disease is the leading cause of morbidity and mortality in most western countries. In the United States cardiovascular disease accounts for 39 % of all deaths [93]. Although environmental and lifestyle factors are of major importance in the development of cardiovascular disease there is also a wealth of evidence in support of a significant genetic component. Understanding the basis whereby a mutation in a single gene can cause disease has appreciably increased our understanding of

the pathophysiology of the more common complex cardiovacular diseases.

4.6.1 Monogenic Disorders

4.6.1.1 Hypercholesterolemia

Elevated low-density lipoprotein (LDL) cholesterol is a well-established risk factor for coronary heart disease. Five monogenic diseases leading to hypercholesterolemia have been described [94]. The most common type of familial hypercholesterolemia is caused by mutation of the LDL receptor gene (LDLR) [95]. The LDL receptor is located on the surface of a number of cell types and is important in removing LDL from the circulation. Individuals inheriting one mutant LDLR gene have total plasma cholesterol in the range of 8.7–10 mmol L^{-1} (desirable level is <5 mmol L^{-1}) and are at significantly increased risk of heart disease. Those homozygous for a mutation (two mutant copies) can have cholesterol levels of 17–30 mmol L^{-1} and usually have severe atherosclerosis and die in late childhood [96].

"Familial ligand-defective apolipoprotein B-100 disease" is clinically similar to LDLR-associated familial hypercholesterolemia. Apolipoprotein B-100 is the major protein found in LDL and specific mutations in the apo-B protein prevent LDL from binding to LDL receptors and thus elevate LDL cholesterol [97].

4.6.1.2 Hypertension

Hypertension is another major risk factor in heart disease and stroke. Elevated blood pressure is highly prevalent in the general population. Multiple environmental and genetic factors are involved in the etiology of hypertension but rare inherited forms of the disease have been informative. Genetic studies have identified mutations in eight genes that cause Mendelian forms of hypertension. Given the diverse physiological mechanisms for regulating blood pressure it is surprising that all the genes identified so far are involved in the renal salt-reabsorption pathway, including genes for glucocorticoid metabolism, glucocorticoid receptors, and renal ion channels [98].

4.6.1.3 Clotting Factors

Genetic factors affecting blood clotting have been studied and a number of prothrombic gene polymorphisms have been described [99, 100]. The significance of some of these variants is still under debate, but a prothrombin variant and a variant of clotting factor V (factor V Leiden) are risk factors for venous thrombosis. Factor V Leiden is the most common hereditary blood coagulation disorder in the United States and is present in 5 % of the Caucasian population [101]. Factor V Leiden increases the risk of venous thrombosis 3–8-fold for heterozygous and 30–140-fold for homozygous individuals [102].

4.6.1.4 Hypertrophic Cardiomyopathy

Hypertrophic cardiomyopathy (HCM) is a relatively common genetic disease, with a prevalence of about 1 in 500 [103], and is an important cause of disability and death in patients of all ages, especially sudden cardiac death in young adults. It is inherited as a Mendelian autosomal dominant trait and is caused by mutations in any one of eleven genes, each encoding a protein of the cardiac sarcomere involved in muscle contraction. Three genes are the most commonly affected: myosin heavy chain, cardiac troponin T and myosin-binding protein C [104].

4.6.1.5 Familial Dilated Cardiomyopathy

Dilated cardiomyopathy (DCM) is a major cause of morbidity and mortality and is the most common cause of congestive heart failure and reason for heart transplant. Familial DCM accounts for more than a third of DCM cases and is clinically and genetically heterogeneous with a prevalence of approximately 40 per 100,000 [105, 106]. Autosomal dominant DCM is the most common form. Numerous genetic loci have been implicated and it seems that DCM results from mutations that affect elements of the cell structure that connect the extracellular matrix to the nucleus through the sarcolemma, the dystrophin complex, the cytoskeleton, the contractile apparatus, and the intermediate filaments [105–107].

4.6.1.6 Familial Arrhythmias

Arrhythmias are a common feature of the hypertrophic and dilated cardiomyopathies described in the previous section. Cardiac arrhythmias are also the primary feature of another group of heart diseases, however, including the "ion channelopathies" (primary electrical disease without underlying structural pathology) and arrhythmogenic right ventricular cardiomyopathy (ARVC).

Brugada Syndrome Brugada Syndrome is now known to be the same condition as "sudden unexplained death syndrome" (SUDS). There is a characteristic ECG pattern and the disease is transmitted as an autosomal dominant trait. So far only one gene has been identified, the *SCN5A* gene that codes for a subunit of the sodium channel. Only about 20 % of patients have mutations in *SCN5A* and other genetic loci must also be involved [106, 108].

Long QT Syndrome Long QT syndrome (LQTS) is characterized by the appearance of a long QT interval on the ECG and an atypical ventricular tachycardia (Torsades de pointes). The prevalence is about 1 in 5000 and mutations in six ion-channel genes and a structural protein have been described [109, 110].

Arrhythmogenic Right Ventricular Cardiomyopathy Arrhythmogenic right ventricular cardiomyopathy (ARVC), also called arrhythmogenic right ventricular dysplasia (ARVD), involves progressive fatty infiltration of the wall of the right ventricle and is associated with increasing severity of arrhythmia and sudden cardiac death. Eight chromosomal loci have been implicated with three genes being identified so far [105, 111].

4.6.2 Multifactorial Cardiovascular Disease

Compared with the number of genes identified in the monogenic cardiac diseases described above, few genes underlying the common forms of cardiovascular disease have been identified. The preferred approach to this problem is the association study in which genotypes in cardiac patients are compared to those of a group of matched controls (Sect. 4.4.1). Yamada et al. [112] typed 112 polymorphisms of 71 candidate genes in 2819 unrelated myocardial infarction (MI) patients and 2242 controls. Genes were selected that were potentially associated with coronary atherosclerosis or vasospasm, hypertension, diabetes mellitus, or hyperlipidemia. They found significant associations with polymorphisms in the connexin 37, plasminogen-activator inhibitor and stromelysin-1 genes. They suggest that determination of the genotypes of the three might be used to predict risk, especially for the stromelysin polymor-

phism that had an odds ratio of 4.7 in women.

Topol et al. [113] typed 72 SNPs from 62 candidate genes in 352 patients with familial premature myocardial infarction and in 418 controls. Candidate genes were chosen for their acknowledged role in endothelial cell biology, vascular biology, lipid metabolism, and the coagulation cascade. Variants in three members of the thrombospondin protein family could be associated statistically with premature coronary artery disease.

In a recent study Tobin et al. [114] studied 58 SNPs in 35 genes previously implicated in cardiovascular disease. They also determined 14 six-SNP haplotypes in 16 genes. Two SNPs, in α-adducin and cholesteryl ester transfer, were each associated with a significant protective effect on MI (odds ratio 0.73 and 0.82). A specific haplotype in the paraoxonase 1/paraoxonase 2 genes was also protective (odds ratio 0.52). Two apolipoprotein C III haplotypes were associated with an increased risk of MI (odds ratios of 1.41 and 1.71).

These studies did not use Bonferroni corrections to correct for multiple tests. As with any association study of this type, firm conclusions about the significance of the results must await confirmation from other researchers using different populations. Nevertheless, these studies point to the future. We should expect larger scale studies, in terms of both numbers of subjects and numbers of SNPs genotyped. Ultimately genome-wide association studies will be conducted using thousands of SNPs arrayed on DNA chips. These future studies will identify cSNPs or chromosomal regions in linkage disequilibrium with genes that are associated with increased or decreased risk. Candidate cSNPs must then be tested in large groups of patients and controls in different populations of different ethnicities.

4.7 Conclusions

Knowledge gained from the human genome project combined with new genomic and proteomic technologies has the capacity to transform the practice of medicine. We will eventually be able to characterize and treat diseases on a biological and molecular basis, rather than by empiricism. The impact on human health might surpass that of antibiotics in reducing morbidity and increasing life expectancy. It is difficult to predict the time it will take to fully implement these benefits (remembering the lesson of gene therapy), but it will certainly take longer than it did for antibiotics. Meanwhile, it is vital to educate the existing cadre of physicians and other health-care workers about the implications of the new discoveries, so they might take the best advantage of subsequent developments.

It is not unreasonable to imagine that, at some time in the not too distant future, the standard practice will be to obtain the sequence of all of the coding DNA in every individual's genome. An individual profile of all known cSNPs and other coding variants could possibly substitute for the complete sequence. This information, carried on a microchip on our health card, will define each of us in molecular terms with regard to our disease susceptibilities and our likely response to drug therapy. As one might imagine, the social and ethical considerations of such a capability will be the subject of considerable discussion and the public as a whole will rightly be wary of a move in this direction. A development in this direction is, however, the logical extension of the human genome project and other technological advances. Thus we have an obligation to educate a wider audience about the science and the implications.

It is to be hoped that the philosophical discussions will be able to keep pace with the speed of our scientific discoveries.

References

1 Watson JD, Crick FHC (1953) Molecular structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 171:737–738

2 Waterston RH, Lindblad-Toh K, Birney E et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562

3 Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408:796–815

4 Dermitzakis ET, Reymond A, Scamuffa N et al. (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* 302:1033–1035

5 Mattick JS (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 25:930–939

6 Carrington JC, Ambros V (2003) Role of microRNAs in plant and animal development. *Science* 301:336–338

7 Szymanski M, Barciszewska MZ, Zywicki M et al. (2003) Noncoding RNA transcripts. *J. Appl. Genet.* 44:1–19

8 Kampa D, Cheng J, Kapranov P et al. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* 14:331–342

9 The Tumor Analysis Best Practices Working Group (2004) Guidelines: Expression profiling – best practices for data generation and interpretation in clinical trials. *Nat. Rev. Genet.* 5:229–237

10 Gusella JF, Wexler NS, Conneally PM et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306:234–238

11 The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72:971–983

12 Halford SE, Rowan AJ, Lipton L et al. (2003) Germline mutations but not somatic changes at the MYH locus contribute to the pathogenesis of unselected colorectal cancers. *Am. J. Pathol.* 162:1545–1548

13 Lafreniere RG, MacDonald ML, Dube MP et al. (2004) Identification of a Novel Gene (HSN2) Causing Hereditary Sensory and Autonomic Neuropathy Type II through the Study of Canadian Genetic Isolates. *Am. J. Hum. Genet.* 74:1064–1073

14 Leach FS, Nicolaides NC, Papadopoulos N et al. (1993) Mutations of a *mutS* homolog in hereditary nonpolyposis colorectal cancer. *Cell* 75:1215–1225

15 Murphy RT, Mogensen J, Shaw A et al. (2004) Novel mutation in cardiac troponin I in recessive idiopathic dilated cardiomyopathy. *Lancet* 363:371–372

16 Bradley BA, Evers BM (1997) Molecular advances in the etiology and treatment of colorectal cancer. *Surg. Oncol.* 6:143–156

17 Ruschoff J, Wallinger S, Dietmaier W et al. (1998) Aspirin suppresses the mutator phenotype associated with hereditary nonpolyposis colorectal cancer by genetic selection. *Proc. Natl. Acad. Sci. USA* 95:11301–11306

18 Grann VR, Jacobson JS, Whang W et al. (2000) Prevention with tamoxifen or other hormones versus prophylactic surgery in BRCA1/2-positive women: a decision analysis. *Cancer J. Sci. Am.* 6:13–20

19 Lucassen A, Parker M (2004) Confidentiality and serious harm in genetics – preserving the confidentiality of one patient and preventing harm to relatives. *Eur. J. Hum. Genet.* 12:93–97

20 Wang C, Gonzalez R, Merajver SD (2004) Assessment of genetic testing and related counseling services: current research and future directions. *Soc. Sci. Med.* 58:1427–1442

21 Godard B, ten Kate L, Evers-Kiebooms G et al. (2003) Population genetic screening programmes: principles, techniques, practices, and policies. *Eur. J. Hum. Genet.* 11 Suppl. 2:S49–S87

22 Hims MM, Diager SP, Inglehearn CF (2003) Retinitis pigmentosa: genes, proteins and prospects. *Dev. Ophthalmol.* 37:109–125

23 Lynch HT, Lynch JF (2000) Hereditary nonpolyposis colorectal cancer. *Semin. Surg. Oncol.* 18:305–313

24 Narayan S, Roy D (2003) Role of APC and DNA mismatch repair genes in the development of colorectal cancers. *Mol. Cancer* 2:41

25 Ingram VM (1959) Abnormal human haemoglobin. III. The chemical difference between normal and sickle cell haemoglobins. *Biochim. Biophys. Acta* 36:402–411

26 Abeliovich D, Kaduri L, Lerer I et al. (1997) The founder mutations 185delAG and 5382insC in BRCA1 and 6174delT in BRCA2 appear in 60% of ovarian cancer and 30% of early-onset breast cancer patients among Ashkenazi women. *Am. J. Hum. Genet.* 60:505–514

27 Petty EM, Green JS, Marx SJ et al. (1994) Mapping the gene for hereditary hyperparathyroidism and prolactinoma (MEN1Burin) to chromosome 11q: evidence for a founder effect in patients from Newfoundland. *Am. J. Hum. Genet.* 54:1060–1066

28 Lynch HT, Coronel SM, Okimoto R et al. (2004) A founder mutation of the MSH2 gene and hereditary nonpolyposis colorectal cancer in the United States. *JAMA* 291:718–724

29 Buckley RH (2000) Gene therapy for human SCID: Dreams become reality. *Nat. Med.* 6:623–624

30 Cavazzana-Calvo M, Hacein-Bey S, de Saint BG et al. (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease . *Science* 288:669–672

31 Cavazzana-Calvo M, Thrasher A, Mavilio F (2004) The future of gene therapy. *Nature* 427:779–781

32 Marshall E (1999) Gene therapy death prompts review of adenovirus vector. *Science* 286:2244–2245

33 Gottesman MM (2003) Cancer gene therapy: an awkward adolescence. *Cancer Gene Ther.* 10:501–508

34 Boillee S, Cleveland DW (2004) Gene therapy for ALS delivers. *Trends Neurosci.* 27:235–238

35 Cargill M, Altshuler D, Ireland J et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22:231–238

36 Reich DE, Gabriel SB, Altshuler D (2003) Quality and completeness of SNP databases. *Nat. Genet.* 33:457–458

37 Kruglyak L, Nickerson DA (2001) Variation is the spice of life. *Nat. Genet.* 27:234–236

38 Strittmatter WJ, Roses AD (1996) Apolipoprotein E and Alzheimer's disease. *Annu. Rev. Neurosci.* 19:53–77

39 Van Eerdewegh P, Little RD, Dupuis J et al. (2002) Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* 418:426–430

40 Dahlback B (1997) Resistance to activated protein C caused by the factor VR506Q mutation is a common risk factor for venous thrombosis. *Thromb. Haemost.* 78:483–488

41 Pennacchio LA, Olivier M, Hubacek JA et al. (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* 294:169–173

42 Hugot JP, Chamaillard M, Zouali H et al. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411:599–603

43 Ozaki K, Ohnishi Y, Iida A et al. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* 32:650–654

44 Stefansson H, Sigurdsson E, Steinthorsdottir V et al. (2002) Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet.* 71:877–892

45 Gretarsdottir S, Thorleifsson G, Reynisdottir ST et al. (2003) The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat. Genet.* 35:131–138

46 Dorman JS, LaPorte RE, Stone RA et al. (1990) Worldwide differences in the incidence of type I diabetes are associated with amino acid variation at position 57 of the HLA-DQ beta chain. *Proc. Natl. Acad. Sci. USA* 87:7370–7374

47 Bell GI, Horita S, Karam JH (1984) A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 33:176–183

48 Nistico L, Buzzetti R, Pritchard LE et al. (1996) The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. *Hum. Mol. Genet.* 5:1075–1080

49 Deeb SS, Fajas L, Nemoto M et al. (1998) A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nat. Genet.* 20:284–287

50 Altshuler D, Hirschhorn JN, Klannemark M et al. (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* 26:76–80

51 Goldstein DB, Weale ME (2001) Population genomics: linkage disequilibrium holds the key. *Curr. Biol.* 11:R576–R579

52 Gabriel SB, Schaffner SF, Nguyen H et al. (2002) The structure of haplotype blocks in the human genome. *Science* 296:2225–2229

53 Goldstein DB, Ahmadi KR, Weale ME et al. (2003) Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.* 19:615–622

54 Crawford DC, Carlson CS, Rieder MJ et al. (2004) Haplotype Diversity across 100 Candidate Genes for Inflammation, Lipid Metabolism, and Blood Pressure Regulation in Two Populations. *Am. J. Hum. Genet.* 74:610–622

55 Dean M (2003) Approaches to identify genes for complex human diseases: lessons from Mendelian disorders. *Hum. Mutat.* 22: 261–274

56 Martin AM, Weber BL (2000) Genetic and hormonal risk factors in breast cancer. *J. Natl. Cancer Inst.* 92:1126–1135

57 Miki Y, Swensen J, Shattuck-Eidens D et al. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* 266:66–71

58 Wooster R, Bignell G, Lancaster J et al. (1995) Identification of the breast cancer susceptibility gene BRCA2. *Nature* 378:789–792

59 Ponder BAJ (1990) Inherited predisposition to cancer. *Trends Genet.* 6:213–220

60 Nakamura Y (1993) The role of the adenomatous polyposis coli (*APC*) gene in human cancers. *Adv. Cancer Res.* 62:65–87

61 Futreal PA, Liu Q, Shattuck-Eidens D et al. (1994) *BRCA1* mutations in primary breast and ovarian carcinomas. *Science* 266:120–122

62 Lancaster JM, Wooster R, Mangion J et al. (1996) *BRCA2* mutations in primary breast and ovarian cancers. *Nat. Genet.* 13:238–240

63 Meijers-Heijboer H, van den OA, Klijn J et al. (2002) Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in non-carriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* 31:55–59

64 Syrjakoski K, Kuukasjarvi T, Auvinen A et al. (2004) CHEK2 1100delC is not a risk factor for male breast cancer population. *Int. J. Cancer* 108:475–476

65 Jhanwar-Uniyal M (2003) BRCA1 in cancer, cell cycle and genomic stability. *Front. Biosci.* 8:s1107–s1117

66 Powell SN, Kachnic LA (2003) Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation. *Oncogene* 22:5784–5791

67 Yang Y, Li CC, Weissman AM (2004) Regulating the p53 system through ubiquitination. *Oncogene* 23:2096–2106

68 Bignell GR, Barfoot R, Seal S et al. (1998) Low frequency of somatic mutations in the LKB1/Peutz–Jeghers syndrome gene in sporadic breast cancer. *Cancer Res.* 58:1384–1386

69 Hanssen AM, Fryns JP (1995) Cowden syndrome. *J. Med. Genet.* 32:117–119

70 Venkitaraman AR (2002) Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* 108:171–182

71 Hughes-Davies L, Huntsman D, Ruas M et al. (2003) EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. *Cell* 115:523–535

72 Ford D, Easton DF, Stratton M et al. (1998) Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* 62:676–689

73 Gradia S, Acharya S, Fishel R (1997) The human mismatch recognition complex hMSH2-hMSH6 functions as a novel molecular switch. *Cell* 91:995–1005

74 Thompson D, Szabo CI, Mangion J et al. (2002) Evaluation of linkage of breast cancer to the putative BRCA3 locus on chromosome 13q21 in 128 multiple case families from the Breast Cancer Linkage Consortium. *Proc. Natl. Acad. Sci. USA* 99:827–831

75 Antoniou AC, Pharoah PD, McMullan G et al. (2002) A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br. J. Cancer* 86:76–83

76 Lichtenstein P, Holm NV, Verkasalo PK et al. (2000) Environmental and heritable factors in the causation of cancer – analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* 343:78–85

77 Peto J, Collins N, Barfoot R et al. (1999) Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer [see comments]. *J. Natl. Cancer Inst.* 91:943–949

78 Groden J, Thliveris A, Samowitz W et al. (1991) Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* 66:589–600

79 Rowan AJ, Lamlum H, Ilyas M et al. (2000) APC mutations in sporadic colorectal tumors: A mutational "hotspot" and interdependence of the "two hits". *Proc. Natl. Acad. Sci. USA* 97:3352–3357

80 Powell SM, Zilz N, Beazer-Barclay Y et al. (1992) *APC* mutations occur early during colorectal tumorigenesis. *Nature* 359:235–237

81 Fang DC, Luo YH, Yang SM et al. (2002) Mutation analysis of APC gene in gastric cancer with microsatellite instability. *World J. Gastroenterol.* 8:787–791

82 Furuuchi K, Tada M, Yamada H et al. (2000) Somatic mutations of the APC gene in primary breast cancers. *Am. J. Pathol.* 156:1997–2005

83 Morin PJ, Sparks AB, Korinek V et al. (1997) Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science* 275:1787–1790

84 Peltomaki P (2003) Role of DNA mismatch repair defects in the pathogenesis of human cancer. *J. Clin. Oncol.* 21:1174–1179

85 Peltomäki P, Lothe RA, Aaltonen LA et al. (1993) Microsatellite instability is associated with tumors that characterize the hereditary non-polyposis colorectal carcinoma syndrome. *Cancer Res.* 53:5853–5855

86 Menigatti M, Di Gregorio C, Borghi F et al. (2001) Methylation pattern of different regions of the MLH1 promoter and silencing of gene expression in hereditary and sporadic colorectal cancer. *Genes Chromosomes Cancer* 31:357–361

87 Lu SL, et al. (2003) Genomic Structure of the Transforming Growth Factor Beta Type II Receptor Gene and Its Mutations in Hereditary Nonpolyposis Colorectal Cancers. *Cancer Res.* 56:4595–4598

88 Miyaki M, Iijima T, Shiba K et al. (2001) Alterations of repeated sequences in 5' upstream and coding regions in colorectal tumors from patients with hereditary nonpolyposis colorectal cancer and Turcot syndrome. *Oncogene* 20:5215–5218

89 Yagi OK, Akiyama Y, Nomizu T et al. (1998) Proapoptotic gene BAX is frequently mutated in hereditary nonpolyposis colorectal cancers but not in adenomas. *Gastroenterology* 114:268–274

90 Jeong SY, Shin KH, Shin JH et al. (2003) Microsatellite instability and mutations in DNA mismatch repair genes in sporadic colorectal cancers. *Dis. Colon Rectum* 46:1069–1077

91 Kaklamani VG, Hou N, Bian Y et al. (2003) TGFBR1*6A and cancer risk: a meta-analysis of seven case-control studies. *J. Clin. Oncol.* 21:3236–3243

92 Pasche B, Kaklamani V, Hou N et al. (2004) TGFBR1*6A and cancer: a meta-analysis of 12 case-control studies. *J. Clin. Oncol.* 22:756–758

93 (2003) *NHLBI fact book, fiscal year 2003.* National Heart, Lung, and Blood Institute, Bethesda, Md.

94 Pullinger CR, Kane JP, Malloy MJ (2003) Primary hypercholesterolemia: genetic causes and treatment of five monogenic disorders. *Expert Rev. Cardiovasc. Ther.* 1:107–119

95 Heath KE, Gahan M, Whittall RA et al. (2001) Low-density lipoprotein receptor gene (LDLR) world-wide website in familial hypercholesterolaemia: update, new features and mutation analysis. *Atherosclerosis* 154:243–246

96 Soutar AK, Naoumova RP, Traub LM (2003) Genetics, clinical phenotype, and molecular cell biology of autosomal recessive hypercholesterolemia. *Arterioscler. Thromb. Vasc. Biol.* 23:1963–1970

97 Ceska R, Vrablik M, Horinek A (2000) Familial defective apolipoprotein B-100: a lesson from homozygous and heterozygous patients. *Physiol Res.* 49 Suppl 1:S125–S130

98 Lifton RP, Gharavi AG, Geller DS (2001) Molecular mechanisms of human hypertension. *Cell* 104:545–556

99 Sykes TC, Fegan C, Mosquera D (2000) Thrombophilia, polymorphisms, and vascular disease. *Mol. Pathol*. 53:300–306

100 Kottke-Marchant K (2002) Genetic polymorphisms associated with venous and arterial thrombosis: an overview. *Arch. Pathol. Lab. Med*. 126:295–304

101 Zivelin A, Griffin JH, Xu X et al. (1997) A single genetic origin for a common Caucasian risk factor for venous thrombosis. *Blood* 89:397–402

102 Price DT, Ridker PM (1997) Factor V Leiden mutation and the risks for thromboembolic disease: a clinical perspective. *Ann. Intern. Med*. 127:895–903

103 Maron BJ, Gardin JM, Flack JM et al. (1995) Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. *Circulation* 92:785–789

104 Seidman JG, Seidman C (2001) The genetic basis for cardiomyopathy: from mutation identification to mechanistic paradigms. *Cell* 104:557–567

105 Franz WM, Muller OJ, Katus HA (2001) Cardiomyopathies: from genetics to the prospect of treatment. *Lancet* 358:1627–1637

106 Keller DI, Carrier L, Schwartz K (2002) Genetics of familial cardiomyopathies and arrhythmias. *Swiss. Med. Wkly*. 132:401–407

107 Vatta M, Mohapatra B, Jimenez S et al. (2003) Mutations in Cypher/ZASP in patients with dilated cardiomyopathy and left ventricular non-compaction. *J. Am. Coll. Cardiol*. 42:2014–2027

108 Priori SG, Napolitano C, Gasparini M et al. (2002) Natural history of Brugada syndrome: insights for risk stratification and management. *Circulation* 105:1342–1347

109 Paulussen AD, Gilissen RA, Armstrong M et al. (2004) Genetic variations of KCNQ1, KCNH2, SCN5A, KCNE1, and KCNE2 in drug-induced long QT syndrome patients. *J. Mol. Med*. 82:182–188

110 Antzelevitch C (2003) Molecular genetics of arrhythmias and cardiovascular conditions associated with arrhythmias. *J. Cardiovasc. Electrophysiol*. 14:1259–1272

111 Ahmad F (2003) The molecular genetics of arrhythmogenic right ventricular dysplasia-cardiomyopathy. *Clin. Invest. Med*. 26:167–178

112 Yamada Y, Izawa H, Ichihara S et al. (2002) Prediction of the risk of myocardial infarction from polymorphisms in candidate genes. *N. Engl. J. Med*. 347:1916–1923

113 Topol EJ, McCarthy J, Gabriel S et al. (2001) Single nucleotide polymorphisms in multiple novel thrombospondin genes may be associated with familial premature myocardial infarction. *Circulation* 104:2641–2644

114 Tobin MD, Braund PS, Burton PR et al. (2004) Genotypes and haplotypes predisposing to myocardial infarction: a multilocus case-control study. *Eur. Heart J*. 25:459–467

Part II
Genomic and Proteomic Technologies

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

5
Genomic Mapping and Positional Cloning, with Emphasis on Plant Science

Apichart Vanavichit, Somvong Tragoonrung, and Theerayut Toojinda

5.1
Introduction

Most genetic traits important to agriculture or human diseases are manifested as observable phenotypes. In many instances, the complexity of the phenotype/genotype interaction and the general lack of clearly identifiable gene products render the direct molecular cloning approach ineffective, so additional strategies, for example genome mapping, are more effective means of identifying the gene(s) in question. In the pre-genomic era, use of genome-mapping approaches required probability statistics to identify the gene positions, followed by positional cloning to identify the underlying genes. In the post-genomic era, to completely characterize genes of interest, the initially mapped region of a trait must to be narrowed to a size suitable for positional cloning and data mining. Strategies for gene identification within the critical region have to be applied after the sequencing of a potentially large clone or set of clones that contains this gene(s). Limited successes of positional cloning have been reported for cloning many genes responsible for human diseases, including cystic fibrosis and muscular dystrophy, and plant disease-resistance genes [1–3].

5.2
Genome Mapping

Genome mapping is used to identify the genetic location of mutants, or qualitative and quantitative trait loci (QTL). Linking the traits to markers using genetic and family information of a recombinant population can identify the gene location. Through mapping we can discover how many loci are involved, where the loci are positioned in the genome, and what contribution each allele might make to the trait. To determine relationships between marker loci and the target trait, mapping requires:

- segregating populations (genetic stocks),
- marker data set(s), and
- a phenotypic data set.

5.2.1
Mapping Populations

The crucial requirement for successful genome mapping in plant science is the choice of a suitable segregating population.

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

Doubled haploids (DH), recombinant inbred lines (RI), backcrosses (BC), and F_2 populations are the primary types commonly used for plant-genome mapping. Each of these populations has unique strengths and weaknesses. In plant science the populations are developed from biparental crosses or backcrosses between parents that are genetically different. The resulting F_2 population is developed by selfing the F_1 individual. The BC can be developed further by crossing the F_1 with one of the two parents used in the initial cross. The DH population is usually derived by doubling chromosomes of haploid cells obtained from the F_1 generation. A variety of methods have been used to produce doubled haploids in plants, for example ovary culture, anther culture, microspore culture, or chromosome elimination. The RI populations are produced by random selfing or sib-mating of individuals of the F_2 or BC_1 population until they become virtually homozygous lines. Most RI populations have been developed by a single-seed descent method.

The population type and the number of progeny determine the resolution of the linkage map. The map in turn affects the precision and accuracy of the number, location, and effect of gene/QTL (quantitative trait loci) which can be detected. Because of their high heterozygosity, F_2 and backcross populations, although easy to develop, cannot indefinitely supply resources for DNA studies, and multiple replicated experiments are not possible. The advantage of F_2 populations over other population types is the large amount of genetic information per progeny when codominant markers are used. The DH and RI populations, on the other hand, are renewable and more permanent resources, multiple-replicated experiments for QTL analysis are feasible. Because RI populations have undergone additional cycles of recombination during selfing whereas F_1-derived DH populations have undergone meiosis only once, RI populations are expected to support the generation of higher resolution maps than the DH populations. The DH, RI, and backcross populations, however, provide half of the genetic information per progeny compared with the F_2 populations.

5.2.2 Molecular Markers: The Key Mapping Reagents

Two classes of marker can be used to detect genetic variation – phenotypic and molecular markers. Phenotypic markers are expression-dependent. Morphological markers are phenotypic markers frequently used for genetic mapping. Phenotypic markers are not an ideal type of marker for genome mapping because of several drawbacks. First, they are limited in number, distribution, and the degree to which the loci can be used to detect polymorphisms. Second, other genes (epistasis and pleiotropism) frequently modify the expression of phenotypic markers. Sequence polymorphisms are especially useful for development of molecular markers, because they are usually stable, numerous, and informative, and detection is more reproducible. Desirable molecular markers for genome mapping must have additional useful properties, for example:

- they must be highly abundant and evenly distributed throughout the genome,
- a highly polymorphic information content (PIC),
- a high multiplex ratio,
- they must be codominant, and
- they must be neutral.

In addition, methods developed for the detection of these markers must have:

- low start-up costs,
- robustness and high reproducibility, and
- a guarantee of transfer of the detection procedure among laboratories.

Basic causes of polymorphisms in stretches of DNA are length polymorphisms because of large insertions, deletions, or rearrangements; single nucleotide polymorphisms; and effects caused by sequence repeats, especially during meiosis. Many molecular markers are now available, each with different advantages and disadvantages. Among these, five commonly used types are RFLP (restriction fragment length polymorphism), RAPD (random amplified polymorphic DNA), AFLP (amplified fragment length polymorphism), SSR (simple sequence repeat), and SSCP (single strand conformational polymorphism) markers.

5.2.2.1

RFLP

RFLP, the first molecular markers [4], are used to detect polymorphisms based on differences in restriction fragment lengths. These differences are caused by mutations or insertion–deletions, which create or delete restriction endonuclease recognition sites. RFLP assays are performed by hybridizing a chemically or radioactively labeled DNA probe to a Southern blot. RFLP are usually specific to a single clone-restriction enzyme combination and most are codominant, highly locus specific, and often multiallellic. RFLP analyses are, however, tedious and inefficient, because of the low multiplex ratio, low genotyping throughput, high labor intensity, and the requirement for large amounts of high-quality template DNA.

5.2.2.2

RAPD

RAPD is a PCR-based technique. A single, arbitrary oligonucleotide primer (typically 10-mer) is used to amplify genomic fragments flanked by two complementary primer-binding sites in an inverted orientation [5]. At low stringency, numbers different of PCR products are generated. Polymorphisms result from either base changes that alter the primer binding site, rearrangements, or insertion–deletions at or between oligonucleotide primer binding sites in the genome. The primary advantages of RAPD are the simplicity of the experimental setup, the low overhead and experiment costs, and the high multiplex ratio. Polymorphic DNA can be isolated and cloned as probes for hybridization or sequencing [6]. The initial cycles of amplification probably involve extensive mismatch, however, and rigorous standardization of the reaction conditions is required for reliable, repeatable results. RAPD do not, furthermore, have defined locus identity, so it can be difficult to relate RAPD loci between different experimental populations of the same species. RAPD can be used to identify dominant mutations.

5.2.2.3

AFLP

AFLP assays are based on a combination of restriction digestion and PCR amplification. AFLP are caused by mutations or insertion–deletions in a restriction site that create or abolish restriction endonuclease recognition sites. They are visually dominant, biallelic, and high-throughput. A principal drawback of AFLP is their time-consuming assay, because the method is based on DNA sequencing procedures. The issue of locus identity needs to be established on a case-by-case basis. An additional drawback of AFLP is that they are reported to cause map expansion and often densely cluster in centromeric regions of the chromosome in species with large genomes. Although linkage map expansion is usually attributed to poor data quality, AFLP technology can be used for genomic analysis in any organism without the need for formal marker development [7–9]. Reproducibility

and high multiplex ratio make AFLP one of today's standard methods for the characterization of markers for genome mapping.

5.2.2.4 SSR

SSR or microsatellites are tandemly repeated mono-, di-, tri-, tetra-, penta, and hexanucleotide motifs. The SSR assay is based on the PCR amplification of tandem repeats using unique flanking DNA sequences as oligonucleotide primers. The polymorphism among individuals is because of the variation in the number of repetitive units. SSR are codominant and often multiallellic, which helps to achieve unambiguous identification of alleles. They are also highly abundant and randomly dispersed throughout most genomes. SSR provide an excellent framework for markers with locus identity. They can be multiplexed to achieve high throughput. A drawback of SSR technology is that the development of SSR is labor intensive and costly. Although SSR are specific for the species they were developed for, the method has now replaced traditional RFLP for generation of many linkage maps, largely because they are technically simple and cheap, consume minute amounts of DNA, and can be delivered with a rapid turn-around time and high PIC (polymorphic information content) [2, 10].

5.2.2.5 SSCP

The SSCP assay is based on changes in the conformation of single-stranded DNA of a specific sequence containing mutations or insertion–deletions under non-denaturing conditions [10–13]. The conformation of the folded DNA molecule is dependant on intra-molecular interactions. SSCP is one of the most sensitive methods for detecting changes in nucleotide sequences of an entire fragment much larger than 1000 bp. SSCP assays are usually performed using heat-denatured DNA on non-denaturing sequencing gels. The strength of this method for genome mapping is its simplicity, multiallellicity, codominance, and locus identity [14–16]. The development of markers is, however, labor-intensive and costly. SSCP have not yet been automated.

5.2.3 Construction of a Linkage Map

Differences between genetic information in progenitors can be visualized by using markers based on morphological, protein, or DNA data. Many potential DNA markers suitable for developing high-density linkage maps have been established for a variety of organisms [17–19]. To construct the linkage map, polymorphic markers are scored on the random segregating populations. The distances and orders of those markers are determined on the basis of the frequency of genetic recombinations occurring in the population. Because two linked markers tend to be inherited together from generation to generation, the distance between markers can be estimated from the observed fraction of recombinations. Map construction basically involves five steps. First, a single-locus analysis, which is a statistical approach used to identify the data quality using a single-locus genetic model. The χ^2 method is widely used to test the marker segregation according to its expected ratio of the randomly segregating population. For example, in a BC progeny, each marker locus will segregate with a 1:1 ratio for Aa and aa, a 1:1 ratio for AA and aa in DH and RIL, and a 1:2:1 ratio for AA, Aa, and aa in the F_2 generation. A significant departure (segregation ratio distortion) from the expected segregation ratio can be a sign of a wrong genetic model, low data quality, or non-random sampling [5]. If the

segregation ratio of each marker does not deviate from the expected ratio, the analysis can proceed to the next step [20]. In the second step of the map-construction process a two-locus analysis is used to test for association or non-independence among the marker alleles located on the same chromosome. Linkage is usually established by testing the independence of the two loci in segregating populations. A goodness of fit or a log likelihood ratio has been used to test for independence of two loci. Recombination fraction, lod score (base-10 log likelihood ratio), and significant *P*-values are used as criteria to infer whether each pair of loci belongs to same linkage group [21, 22].

During the third step, a three-locus analysis is used to determine the ordering of the loci or the linear arrangement of markers in a linkage group. Two methods are used to find the best locus ordering among the potential orders in each linkage group, double crossing-over and two-locus recombination fraction. The order of the three loci can be determined by finding the least-occurring double recombinants. When the double recombinant classes are identified, the order of the loci can be determined. The two-locus recombination fraction approach is used to determine the locus order by comparing the likelihood of the three possible orders. The order associated with the highest likelihood values is the most likely order. The ordering of more than three loci can be determined using the maximum likelihood approach [21, 23, 24].

During the fourth step, a map distance is calculated. In the process of the linkage map construction, the recombination fractions are mathematically calculated from data obtained by mapping of the population [25]. On the basis of the data obtained, mapping methods are subsequently used to convert the recombination fraction into the map distance. Different kinds of mapping function have been proposed. Mapping functions work only for specific conditions. There is no universal mapping function. The differences among the commonly used mapping functions are because of the assumptions about distribution of crossovers on the genome, crossover interference, and length of the chromosome segment. Genes or genetic markers are organized in a linear fashion on a map, thus their relative positions on the map can be quantified additively. If the expected number of crossovers is one per genome segment, the map distance between two genes or genetic markers flanking the segment is defined as 1 Morgan (M) or 100 centi-Morgans (cM). The commonly used mapping methods are Morgan's, Haldane's, and Kosambi's. Morgan's mapping method can be appropriately applied for a small genome, which most probably has a small expected number of multiple-crossovers compared with a large genome [26]. As the size of the genome increases, the expected number of multiple-crossovers becomes larger and the map distance has to be adjusted for multiple-crossovers by use of Haldane's mapping method, by assuming that crossovers occur uniformly (randomly) along the length of the chromosome (in the absence of crossover interference) [27]. Experimental evidence has been found that crossover interference exists and crossovers occur non-randomly in larger genomes. Therefore, Kosambi's mapping method was invented to take into account the crossover interference. The rationale for Kosambi's method is that crossover interference depends on the size of a genome segment. The interference is absent when a segment is sufficiently large and increases as segment size decreases [28].

The final step in the mapping procedure is linkage map construction. During recent decades, many approaches have been devel-

oped for building a multi-locus model. The least-squares method was implemented in a computer package "Joinmap" for genomic mapping. The EM algorithm uses a set of procedures for obtaining a maximum likelihood estimate. The Lander–Green algorithm greatly reduces the computational complexity of obtaining multi-locus recombination fractions from traditional approaches. This method has been widely implemented in computer packages such as Mapmaker [29], Gmendel [30], and PGRI. The joint maximum likelihood is a method to estimate recombination fractions and crossover interference by simultaneously using a multiplicative model. This algorithm has been implemented in the computer package "GLIM" which can be used to apply generalized linear regression. The simulation approach involves the comparison of multilocus likelihoods of the data using different mapping functions. This approach can be used to identify a mapping method which fits the data well. It was implemented in the computer program Linkage [31].

Although most linkage maps are based on population sizes ranging from 100 to 200 individuals, the study of larger progenies can help enhance map resolution and estimation of map distances [32].

5.3 Positional Cloning

Information about map positions of genes is used to conduct chromosome walking during positional cloning. In the initial step, flanking markers which are tightly linked to the target gene must be identified. These tightly linked markers are then used as initial points for the development of the high-resolution map around the target region, using highly polymorphic content markers. When the flanking markers are narrowed down, the next step is to construct a physical map around the target region. The candidate region can subsequently be narrowed down further, sometimes to a region being covered by a single large insert clone. After characterization of the genes by sequencing, functional analysis by complementation in transformed plants is the most important piece of evidence for the successful identification and cloning [33].

5.3.1 Successful Positional Cloning

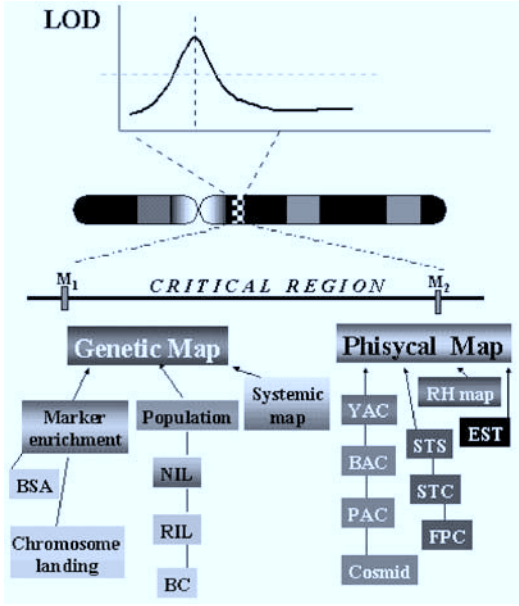
More than 100 inherited disease genes in humans have been isolated (http://genome.nhgri.nih.gov/clone/). Significant progress in positional cloning in plants was achieved, however, mostly because of the development of high-density maps and large insert libraries in major crops such as rice [19], *Arabidopsis* [34], tomato [35], and barley [36]. Two classic examples of gene identification and location were the cloning of the *Pto* gene in tomato [2] and the *Xa21* gene in rice [3]. These genes are responsible for resistance against bacterial pathogens. *Arabidopsis* has become a model plant for map-based cloning, because of the simplicity of identification of mutations, comprehensive genetic and physical maps, and the ease of gene transformation. Examples of disease-resistance genes from *Arabidopsis* that have been cloned include the *RPM1* gene against *P. syringae* [37], *RPS2* against a different strain of *P. syringae* [38], *RPP13* against downy mildew fungus [39], *Mlo*, against broad-spectrum fungal attack in barley [40], and *I2* against fusarium wilt in tomato [41]. Other disease-resistance genes that have been identified are *Tm2a* against *TMV* in tomato [42], *Asc* against alternaria stem cancer in tomato [43], *Pi-b* against rice blast [44], *Pi-ta2* also against rice blast [44],

and *Rar-1* against powdery mildew in barley [45, 46]. Successful positional cloning has been achieved for the gene responsible for the resistance to beet cyst nematode, $Hs1^{pro\text{-}1}$, [47]. The *Br* gene responsible for resistance to bruchid, the grain weevil that destroys mungbean seeds, was cloned by Kaga and Ishimoto [48].

Attempts to clone the plant genes responsible for resistance to the stress of abiotic factors such as cold, drought, flooding, etc., have faced more challenges, because of their complex interaction with several genetic and non-genetic factors. This is particularly true for genes related to phytohormone activity. Many *Arabidopsis* mutants with defects in signal-transduction pathways have been used for map-based cloning. Successful gene isolation by map-based cloning was reported for genes responsible for insensitivity to abscisic acid, *ABI1* [49], *ABI2* [50], *ABI3* [51], and *ABI4* [52], elongation-regulating hormone, auxin, *AXR1 & 2* [53], senescence-promoting ethylene, *ETR* [54], *EIN2*, and *CTR1* [55], gibberellic acid, *GAI* [56] and *GA1* [57]. Positional cloning has been successfully used to isolate genes involved in development that are important to agriculture, for example *d1*, dwarfism in rice [58], and a *MADBOX* gene controlling fruit dehiscence (*jointless*) in tomato [59].

5.3.2 Defining the Critical Region

The identification and isolation of genes by positional cloning strategies can be conceptually divided into a series of steps. The starting point is collection of families or



BSA = Bulk Segregant Analysis; NL = New Isogenic Line; RIL = RecombinantInbred Line; BC = Back Cross, YAC = Yeast Artificial Chromosome, BAC = Bacterial Artificial Chromosome; PAC = P1 Artificial Chromosome; STS = Sequence Tagged Site; STC = Sequence-Tagged-Connector; FPC = FingerprintContig; EST = Expression Sequence Tag; M = Flanking Marker

Fig. 5.1 Genetic and physical approaches to refining the critical region.

germplasms with defects or special traits of interest. After the critical region is identified by flanking markers, usually spanning tens of million base pairs, more refinement is achieved by using additional genetic markers that map in the vicinity. The extent to which the map region can be determined by genetic means depends on the size of the population and the number and informativeness of the genetic markers in the region. In the ideal case the defective genes were caused by cytological abnormalities, which immediately establishes the critical region containing the defective genes. Of critical importance for phenotyping screening limits is the number of progeny that can be evaluated in the study.

Genetic mapping can be narrowed down to 1–3 cM in optimum examples of human diseases or as tightly as 0 cM in plants [42, 60]. The corresponding physical size, however, can vary widely, because of genome size and regional and sex-specific differences in recombination rates. In humans, 1 cM typically corresponds to approximately 1–3 Mb. In plants, the physical-to-genetic distance per 1 cM varies with genome size (e.g. 100 kb in *Arabidopsis*, 250 kb in rice, 1000 kb in maize). In recombination hot spots the physical-to-genetic distance can be particularly small and such regions have frequently been associated with gene richness. Choices of strategies for positional cloning, as illustrated in Fig. 5.1, depend on the tools available for particular organisms.

5.3.3 Refining the Critical Region: Genetic Approaches

Flanking markers identified in the preliminary map normally are too far to reach the target QTL by most large-DNA insert-cloning technologies. The generation of more polymorphic markers within a specific region can be achieved by genetic or physical means. Recombination events near the target gene at a resolution of 0.1 cM can help facilitate gene identification.

Positional cloning is laborious and costly for organisms for which high-resolution maps are not available. An ideal method would be to directly isolate region-specific markers at high density to identify overlapping genomic clones covering the genes of interest without generating a genetic map or performing a chromosome walking procedure. Chromosome landing and pooled progeny techniques are based on the identity of their chromosomal region of interest [61]. In bulk segregant analysis (BSA) pooled progenies are based on their phenotypic identity. BSA has been successfully used to isolate markers surrounding the major loci [43]. BSA has the same advantage as chromosome landing, because there is no requirement for the construction of a high-density map. Combined with high-multiplex ratio markers such as AFLP, SSR, and RAPD, BSA and chromosome landing can reduce the number of DNA samples necessary to score thousands of markers [62]. Genetically directed representational difference analysis (GDRDA) uses phenotypic pooling, combined with a subtractive method, to specifically isolate markers from a locus of interest [20]. This method was often insufficient to locate specific markers in a specific region. RFLP subtraction has been used to isolate large numbers of randomly located RFLP spanning the mouse genome [63] and *Volvox carteri* [64]. Three RFLP markers linked to *recA* at 0 cM have been isolated by Corrette-Bennett et al. [64], but it is not clear if their methodology can be applied to plant or animal genomes in which repetitive sequences and retrotransposons are abundant.

5.3.4 Refining the Critical Region: Physical Approaches

The decision to start the physical mapping process depends on the mapping tools available for a particular organism. In humans, typically a 1–3 Mb interval can be reached by using YAC physical mapping. In other instances, when the physical map is not well refined, the critical region must be narrowed down to as small a distance as possible, using the wealth of polymorphic markers in the particular region. Typically, mapping the closest genetic markers is used to initiate clone isolation (e.g. YAC, BAC, or PAC). If necessary, new markers can be generated from the ends of the clone which can then be used to screen the next adjacent overlapping clones, a strategy called "chromosome walking". Additional markers such as STS and EST, which were identified in the critical region, can be used to assist clone isolation. In an ideal case, such as that illustrated in Fig. 5.2, the entire critical genomic interval between the flanking markers can be isolated in YAC or BAC clones. Fine mapping with such specific markers can significantly narrow the critical region, and gene isolation can be achieved with less effort than physical mapping.

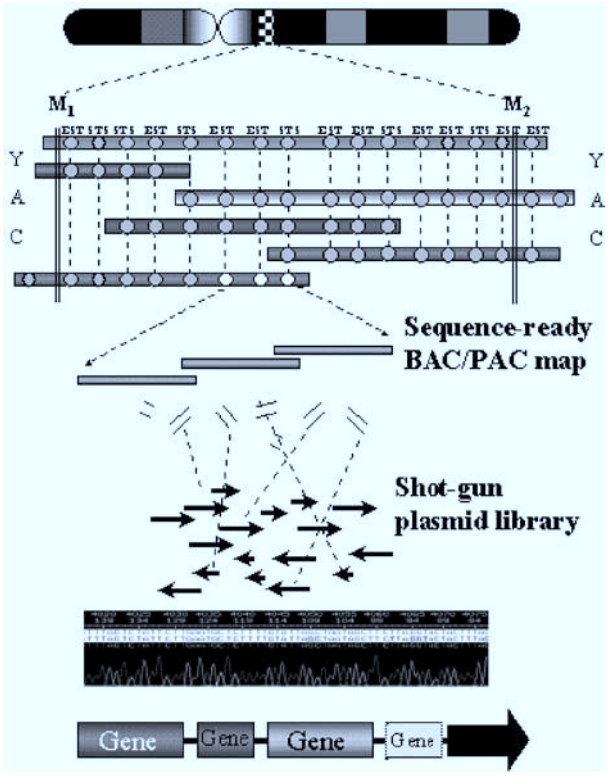


Fig. 5.2 Refinement of the critical region by physical mapping and sequencing.

5.3.5 Cloning Large Genomic Inserts

Cloning large genomic fragments facilitates access to genes known by map position, physical mapping, and large-scale genome sequencing. The two most powerful cloning vectors are yeast artificial chromosomes (YAC) and bacterial artificial chromosomes (BAC). The development of YAC enables cloning of DNA inserts of more than 500 kb [65]. However, cloning in YAC normally generates a large amount of chimerism and rearrangements, which limits the usability of this method in physical mapping and map-based cloning [39, 65–67]. On the other hand, BAC, which utilize the *E. coli* single-copy fertility plasmid (F plasmid) can maintain up to 350 kb inserts with little or no rearrangement or chimerism [68–71]. An alternative to cloning in BAC, the PAC system is an artificial chromosome vector developed on the basis of bacteriophage P1 [66]. Both BAC and PAC, as compared with YAC, have much higher cloning efficiencies, improved fidelity, and are easier to handle. Because of its high stability and ease of use, the BAC cloning system has emerged as the vector of choice for the construction of large insert libraries such human [72], bovine [47], *Arabidopsis* [73], rice [74], and sorghum [75]. To add even more capabilities to the BAC system, the T-DNA locus and origin of replication from *Agrobacterium tumefaciens* were engineered into a binary bacterial artificial chromosome (BIBAC) vector to make it an ideal plant transformation vector with the capacity of replicating in both *E. coli* and *A. tumefaciens* [27]. The original vectors have subsequently been modified to improve the transformation efficiency by electroporation, cloning features, and selection specificity [76]. Rice and wheat genomic libraries were recently constructed using the improved BIBAC vector [76]. Using the improved BIBAC vector, a gene causing filamentous flower was isolated [77]. The entire 150-kb human genomic fragment was transferred into tobacco plants [78]. Recently, an additional P-lox site has been inserted into the pBACwich vector to enable site-specific recombination when using particle bombardment for plant transformation [79]. The cre-lox system was successfully used to mediate recombination between *Arabidopsis* and *Nicotiana tabacum* chromosomal regions [80]. Such improvements can now be used, with the help of positional cloning, to understand the function of genes.

5.3.6 Radiation Hybrid Map

A radiation hybrid map (RH map) can be constructed on the basis of radiation-induced breakage between loci. The map distance between markers, estimated by the frequency of chromosome breakage, is proportional to the physical distance between the markers. The map resolution is, depending on the amount of radiation used, on average between 100–1000 kb. Unlike physical maps, RH mapping is more random, because chromosome breakage does not depend on the frequency of restriction enzymes or a cloning bias. Nonetheless, monomorphic markers such as EST and STS can be assigned to a RH map. With the high resolution achieved by RH mapping and the high density markers, RH mapping is a very powerful tool for positional cloning. Positional cloning of hyperekplexia, a human autosomal dominant neurological disorder was achieved within 6 weeks after the critical region was identified and unrelated candidate genes were eliminated using RH mapping [5]. RH mapping is extremely useful in any chromosomal

region where recombination is suppressed. Publicly available RH mapping panels have been maintained at Genethon (http://www.genethon.fr/genethon_en.html), the Whitehead Institute (http://www.genome.wi.mit.edu/), and Stanford University (http://shgc-www.stanford.edu/ Mapping/rh).

5.3.7 Identification of Genes Within the Refined Critical Region

The most challenging step in positional cloning is identification of genes in the critical region. There are strategies for the identification of genes in large genomic clones such as YAC, BAC, or PAC clones. Ideally the trait is cosegregated with the marker or an EST derived from high-resolution mapping. When the critical region cannot be narrowed down further and no other known marker is available, small DNA fragments from a YAC or BAC clones containing the gene can be used to hybridize in the orthologous region in such systems as the human/mouse/rat system, *Arabidopsis*–tomato–*Brassica*, or the rice/maize/wheat system. Detection of cross-hybridization reveals conserved sequences between these species and, therefore, genes or EST identified in such a syntenic region are likely to have biological function. The presence of CpG islands nearby often marks the 5′ ends of genes and can be subsequently used for gene isolation. CpG islands, exon trapping, and direct cDNA selection are complementary approaches that can be used to identify exons in genomic sequences.

5.3.7.1 Gene Detection by CpG Island

Its unusual G + C-rich DNA first distinguished CpG islands from other genomic sequences [38]. More than 60% of human genes contain CpG islands, both in promoters and at least a part of one exon [81].

5.3.7.2 Exon Trapping

The presence of consensus sequences at "splice junctions" enables isolation of adjacent exons by "exon amplification" or "exon trapping" [82]. Occasionally the entire internal exon can be captured [83]. In a similar fashion, several genes causing human diseases have been isolated by trapping their terminal exons using the poly-A tail signal [84].

5.3.7.3 Direct cDNA Selection

The candidate genomic clones can be used either templates to screen cDNA libraries constructed from the target tissues or subtractive hybridization. Using a method called "direct cDNA selection", the target genomic clones are fixed on an affinity matrix to capture cDNA by homology-based hybridization. The success of the method relies on the source and quality of the cDNA that are about to be captured. The genes can only be captured if they are expressed in the tissue from which mRNA or cDNA libraries were isolated. Some genes with low levels of expression or which are absent in the target tissue might be difficult or impossible to isolate by direct cDNA selection.

5.4 Comparative Mapping and Positional Cloning

Comparative mapping combines genetic information accumulated from related species. Usually, linkage maps of each species have been constructed with various kinds of molecular marker and used independently

for particular genetic purposes. Comparisons among maps of related species, using low copy-number sequences as a probe to hybridize with genomic DNA, indicated the substantially conserved orders of the DNA sequences among their genomes. This is well documented in the grass family, which diverged ~60 million years ago [85–87]. The mammals including human and mouse evolved from a common progenitor ~70 million years ago and are also documented as showing conserved orders of DNA sequences [88]. The existence of conserved gene orders (colinearity) and contents in related species indicates that new genes are rarely created within evolutionary time frames of at least ten million years. Most new genes probably arise from gene duplication and/or gene modification of currently existing genes [89]. Colinearity can therefore enable gene prediction across families and the extrapolation of mapping data from one organism to another. Candidate genes can thus also be easily isolated and predicted from species for which well established linkage maps are not available.

5.4.1 Synteny, Colinearity, and Positional Cloning

Rice (430 Mb), maize (3500 Mb), and wheat (16,000 Mb) separated 50–60 million years ago [90]. Only genes which are extensively conservative can be found in orthologous regions which diverged several million years ago [58]. Comparative maps have been constructed between rice and maize [85], between oat and maize [91], among rice strains, between wheat and maize based on a common set of cDNA [92], and between rice and barley [72, 93]. Being phylogenetically 60 million years apart, members of the grass family still share extensive syntenies in a number of regions. Therefore, the idea that the map position of one species can be used to identify and compare orthologous alleles across species is feasible for grasses [94, 95].

Comparative mapping in cereals, including rice, foxtail millet, sugar cane, sorghum, maize, *Triticeae* and oats, revealed substantial colinearity when RFLP and EST markers were used. Rice chromosomal segments reconstructed in the context of its relative genome are termed rice linkage segments (RLS) [96]. Among dicots, *Arabidopsis* and *Brassica* species, tomato, and soybean, were co-linear (also called macro-syntheny) to a lesser extent than cereals. The synteny between monocots and dicots was limited.

In positional cloning experiments, microcolinearity can be extremely useful for cloning genes from species with large genomes, such as wheat and maize, using information from small genome species like rice based on their synteny. When a 6.5 cM region in barley's chromosome 1 containing the barley rust resistance gene *Rpg1* was compared with the 2.5 cM syntenic region in rice chromosome 6, the order of RFLP markers was conserved [72]. In the case of the *Adh1* locus, composition and arrangement of genomic DNA fragments were compared between maize YAC and sorghum BAC clones, containing the orthologous loci [3, 58]. Because of a 75 kb stretch of highly repetitive elements in maize, chromosome walking to the *Adh1* gene was made possible by cross-referencing to the sorghum BAC. In a similar case, synteny was reported in the *sh2-a1* homologous regions among maize, sorghum, and rice, where the distance between the two genes was 140 kb in maize but only 19 kb in rice and sorghum [97]. These studies reveal that small rearrangements, including frequent insertions of transposons or retrotransposons can occur without significant rearrangement of the orthologous region [89].

Another striking demonstration of colinearity in the grass family is the gene which causes dwarfism. The so-called "green revolution" genes were found in wheat (*Rht*, reduced height), maize (*d8*, dwarf), and rice (*d*, dwarf). Sequence comparison revealed that the *Rht* and maize *d8* genes were orthologs of the *Arabidopsis GAI*, which is the gibberellin-insensitive mutant [98]. Comparative mapping between wheat and rice using RFLP markers linked to *Rht-D1b* and between rice and maize using RFLP markers linked to D8 clearly showed colinearity among wheat chromosome D4, rice chromosome 3, and the maize chromosome 1. Additionally, one of the spontaneous rice dwarf mutants, *d1*, was isolated by positional cloning and was found to be an ortholog or the alpha subunit of G-protein, which is related to the *GAI* mutant found in *Arabidopsis* [60]. The *d1* mutant, however, was mapped to chromosome 5. Because at least 54 dwarf mutants have been identified in rice, at least one might actually be located in the synthenic region of chromosome 3. Although comparative mapping is a powerful tool for finding genes in large syntenic regions, extensive gene or segmental duplication in the reference genome can obstruct such comparison and lead to the false assignment of a syntenic region.

5.4.2 Bridging Model Organisms

Genomes of model organisms have played a critical role in the positional cloning of human diseases. Humans and mice are the most extensively mapped mammals. The map location of more than 3000 genes and syntenic regions has been identified and displayed at http://www3.ncbi.nlm.nih.gov/Homology/. These maps are also linked to the MGD (Mouse Genome Database) and OMIM (Online Mendelian Inheritance in Man; http://www3.ncbi.nlm.nih.gov/omim), a catalog of human genes and genetic disorders. To integrate maps of the two different species, type I markers such as EST have been the most useful tool for anchoring loci during comparative mapping [99].

The complete genomic sequence and annotated set of genes for budding yeast (http://www.stanford.edu/Saccharomyces) has opened another possibility for the cross-referencing of human genes [100]. Several human disease genes have been cloned in this way, examples include *MDR1*, the multidrug-resistance gene in humans, which encodes a protein required for the pheromone factor involved in yeast mating [101]. The human neurofibromatosis type 1 gene can complement the function of defective IRA in yeast [102]. Tremendous conservation of genes still exists between the human and the fly [103].

Completion of the whole genome sequence of *Arabidopsis* and rice opens an opportunity to compare more closely two genomes that diverged approximately 200 million years ago. Alignment between 189.5 Mb of rice and *Arabidopsis* revealed 4–22 rice orthologs covering 3.2 cM [104]. One *Arabidopsis* contig aligned with two distinct rice physical regions, showing these are actually duplicated. The concept of using genome-cross referencing can be illustrated for genes controlling flowering time. In *Arabidopsis* the genes for flowering time have been cloned and completely characterized. In cereals, photoperiod sensitivity can be related to genes for flowering time in *Arabidopsis*. The barley *Ppd-H1* plays a major role in regulating flowering time in barley [105]. Located on chromosome 2H, the barley flowering gene was homologous with the wheat Ppd gene series in the wheat and barley RFLP map. The barley Ppd-H1,

the homolog of the wheat Ppd genes [106], was located on barley chromosome 2H or in the vicinity of the junction between the RLS7 and RLS4a [87] flanked by two cDNA at the distal end of rice chromosome 7 [107] where the rice *Hd2* was identified [108]. Analysis of the *Arabidopsis* genomic sequence did not identify any region equivalent to Hd2 and Ppd-H1 regions of rice and barley. The cDNA sequence was used to search for homologs in the *Arabidopsis* genome, but failed to provide candidate genes. This has led to the conclusion that significant colinearity exists among species and has been maintained over a long evolutionary time.

5.4.3 Predicting Candidate Genes in the Critical Region

The most direct approach to gene identification in a genomic region involves analysis of DNA sequences. Such a high-quality sequence will soon be available for human, *C. elegans*, and *Arabidopsis* [109], among others. Taking the human genome project as an example, the availability of the comprehensive high-resolution map and physical map that assembles EST, STC, STS, and STR into large genomic contigs enables human disease genes to be assigned precisely and rapidly to the critical region. When genomic sequences are available genes can be predicted more accurately using modern computational tools including GenScan, GeneMarkHMM, Xgrail, and Glimmer, as illustrated in Fig. 5.3. All the genes identified by homology or by prediction in the critical region can become candidates. It is essential to understand how those candidate genes function. To prove that one of those candidate genes is the responsible gene, it must be demonstrated that the mutation in the gene is genetically associated with the phenotype. At this stage the availability of single nucleotide polymorphisms (SNP) will substantially improve the rate of mutation discovery in the candidate genes. Ultimate proof of the candidate gene requires evidence of a complementation test, however.

5.4.4 EST: Key to Gene Identification in the Critical Region

As sequence-based markers, EST play a crucial role in both gene-based physical map construction and candidate gene identification. In recent years, EST production on a massive scale has been conducted at the Institute for Genomic Research, TIGR (http://www.tigr.org/), and over 100,000 EST have been released by TIGR and others to be maintained in the "dbEST" database (http://www.ncbi.nlm.nih.gov/dbEST/) at NCBI. Among the largest publicly available EST collections, more than 500,000 EST have been produced at Washington University in St Louis, supported by Merck and Company (http://genome.wustl.edu/est/esthmpg.html). Because the current EST are only 97 % accurate and short, unedited, single-pass reads, they are clustered into "tentative consensus sequences" at TIGR (http://www.tigr.org/) or "uniquegene clusters" at NCBI (http://www.ncbi.nlm.nih.gov/UniGene/).

EST have been also produced from organisms other than the human. One of the largest collections is the 300,000 mouse EST funded by the Howard Hughes Medical Institute (http://genome.wustl.edu/est/mouse_esthmpg.html). Other model organisms are *C. elegans* (http://ddbj.nig.ac.jp/htmls/c-elegans/html/), *Arabidopsis thaliana* (http://genome-www.stanford.edu/Arabidopsis/), rice (http://ww.staff.or.jp/), and *Drosophila melanogaster* (http://fly2.

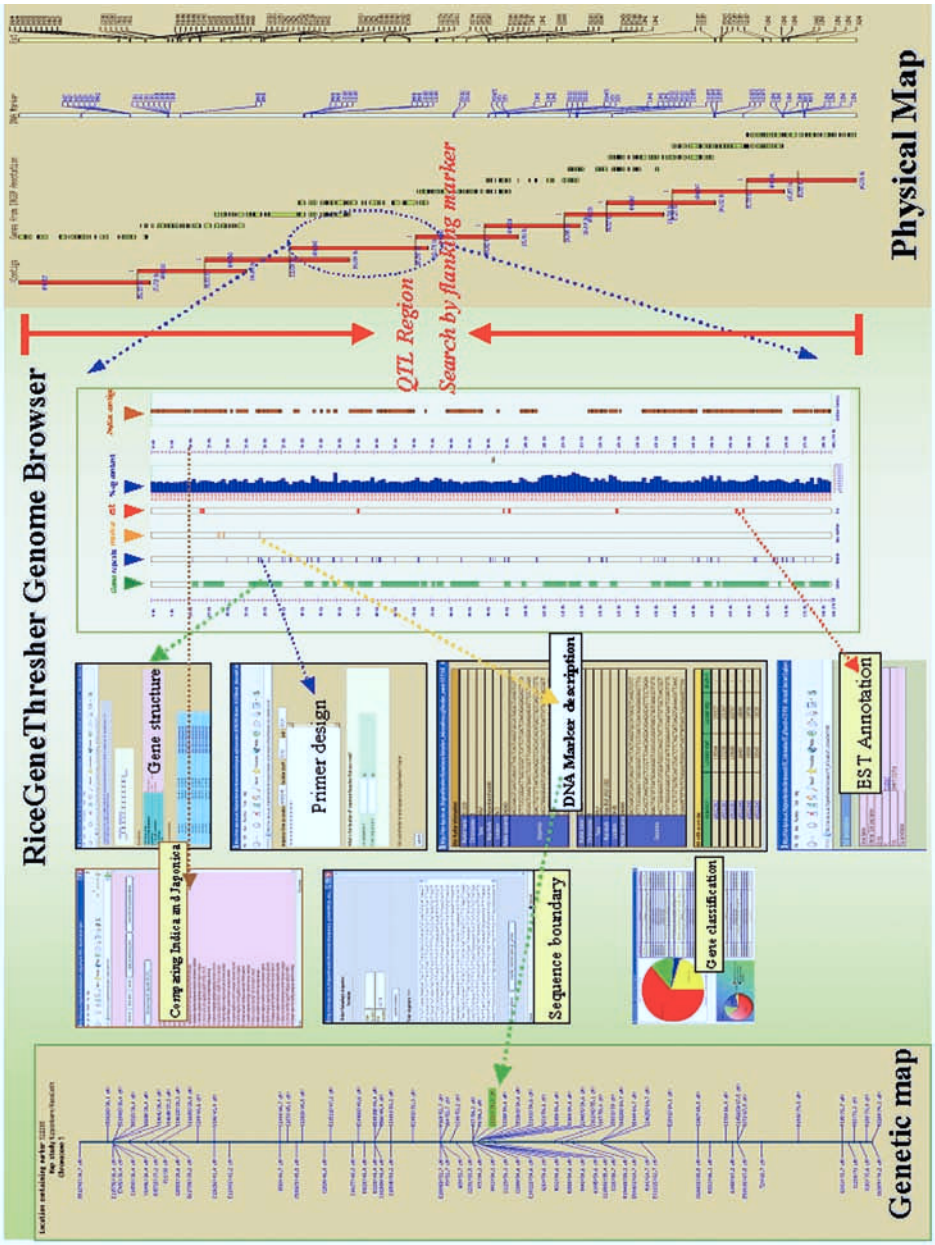


Fig. 5.3 A genome browser facilitates positional cloning (http://dna.kps.ku.ac.th).

berkeley.edu/). These homologous or orthologous resources are curated at "HomoloGene" at NCBI (http://www.ncbi.nim.nih.gov/HomoloGene/).

Identification of EST on physical and RH maps of a critical region can aid immediate identification of candidate genes and simplify the positional cloning of particular genes by leaping across taxonomic boundaries, because of the conserved protein sequences.

5.4.5 Linkage Disequilibrium Mapping

Genetic variation within coding sequences is highly conserved and can rarely be detected by length polymorphism. Polymorphisms corresponding to differences at a single nucleotide level, which are caused either by deletion, insertion, or substitution, are biallelic in diploids, occurring frequently and uniformly in most genomes at roughly one in every 500–1000 bp. Because genomic and EST sequences have increased at an exponential rate in the past few years, SNP discovery in coding regions is projected for at least 100,000 markers with the aim of identifying and cataloging all human genes and creating more or less complete human SNP maps. Because SNP residing within coding regions are rare, these mutations might correspond to defects that are associated with polymorphism at the protein level, diseases, or other phenotypes. To prove that SNP are associated with a particular disease, haplotype analysis of SNP in the candidate genes must be conducted among affected and unaffected individuals from the same family. In this case, the sample size must be large enough to reveal statistical differences between the affected and unaffected pools and show a linkage disequilibrium. Positional cloning can be more effective and less time-consuming if SNP are used not only to refine the critical region but also to confirm the position of the real candidate genes.

5.5 Genetic Mapping in the Post-genomics Era

Genome projects can dramatically simplify the long, tedious process of positional cloning. Physical mapping can also be tedious and costly. For most genome projects, BAC-end sequences or sequence tag connectors (STC) from a 10–20 X BAC library can substantially simplify physical map construction. "End walking" can be performed *in silico*, initiated by several rounds of BLAST searching of the sequences of flanking markers or the next BAC end sequences against the respective BAC-end sequence database until the critical region completely overlaps with a BAC contig. In addition, a BAC fingerprint database, in which contigs are being assembled, can be used to confirm and anchor the genetic map by using well-mapped molecular information, as illustrated in Fig. 5.4.

Although there have been several successes in map-based cloning in many species, cloning QTL has remained a formidable task for their small allelic difference and large environmental effects. Genomic tools are available to make genome mapping more efficient. Abundant molecular markers can be developed from available genomic sequence, enabling first-pass genome mapping to be done quickly. Fine-scale mapping can be developed to narrow down QTL regions to as small as 50 kb.

Recent advances in oligonucleotide array technologies have added new opportunities to genotype several thousand of genes in a single array or GeneChip [110]. By hybridizing labeled total genomic DNA to the

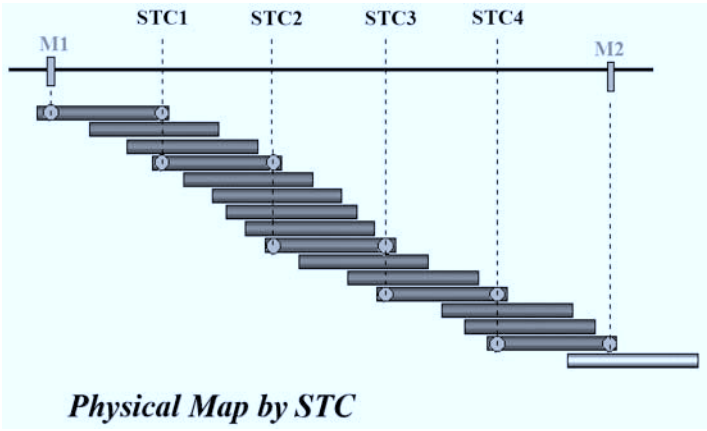


Fig. 5.4 *In silico* physical mapping by bridging sequence tag connectors.

GeneChip, allelic variants like indels and SNP that differentially hybridize to the array features can be detected. This approach is useful in LD mapping, dissecting QTL, and in assessing species population structure in yeast [111] and, recently, in *Arabidopsis* [112]. In *Arabidopsis* the ATH1 GeneChip, comprising 8000 genes and 103,860 features with unique positions in the genome, detected nearly 4000 polymorphisms between Columbia and Lansberg *erecta* [113]. About 700 polymorphisms were used as informative markers to construct a linkage map in a RIL population. The map had very high resolution of 0.5 cM. Those gene-specific markers in conjunction with expression and metabolic profiling can be used to dissect QTLs in *Arabidopsis*. This suggests that LD mapping in *Arabidopsis* is now feasible by gene-array genotyping. The creation of an LD map using GeneChip array is, however, very expensive compared with the more conventional ways. The GeneChip was used to locate the *ERECTA* locus to within a 12 cM interval using bulk segregant analysis of 15 Columbia or Landberg F2s [112]. Another application was the identification and mapping of mutant genes in a population treated with a mutagen.

5.5.1 eQTL

Traditional genome mapping and map-based cloning have been very successfully used for dissecting simple Mendelian traits. They have been much less successful for positional cloning of QTL, however. Interplay of many loci and environmental factors affect the phenotypic variation and weaken the statistical association between allelic variation and phenotypic variability. Advances in new genomic tools, for example transcriptomics, proteomics, and metabolomics, make it possible to monitor several thousand genes, proteins, and metabolites on the genomic scale simultaneously. By combining expression data with the QTL, identification of genes underlying the QTL might be possible without fine scale mapping.

Using PQL or "protein quantity loci" was one of the first attempts to link the variability of protein density induced by drought and the drought QTL in maize [113]. Protein profiling using two-dimensional gel electrophoresis from 140 RIL generated 200 quantified proteins, 35 were induced by stress and genotypic effects were observed for 10 proteins. Several PQL coincided with

QTL for drought and ABA responses. These were narrowed down to small regions apparently involved in regulating the gene expression of a large number of other genes throughout the genome when combined with the phenotypic data.

By measuring the relative abundance of mRNA produced by many genes simultaneously, new technologies such as microarrays have been used to shed light on regulation of gene transcription. cDNA microarrays have been the prominent method for gene expression analysis in rice [114], barley [115], poplar [116], and strawberry [117]. Several reports recently shown that a significant portion of gene expression levels are under genetic control in different organisms. Determination of gene expression levels by exploiting segregating populations is termed "eQTL". Compared with "PQL", eQTL can lead to higher throughput and genome coverage. Schadt et al. [118] used a mouse gene oligonucleotide microarray to monitor the gene expression levels of 23,574 genes in liver tissues from 111 F2 mice constructed from two standard inbred strains. They found hotspot regions in genomic DNA that are involved in regulating the expression of many other genes. Increasing numbers of genetic factors underlying QTL identified so far were noncoding. Nucleotide variation in the promoter or upstream sequence might be responsible for eQTL. If this is so, searching only for coding sequences in the candidate genes might fail to identify the causal factors responsible for genetic variation in the QTL. Interpreting noncoding regulatory variants poses special problems. Noncoding regulatory sequences might be involved in splicing variation. For example, sequence variation at a splice junction in the waxy locus was responsible for a QTL responsible for variation of the amylose content of rice [119].

By combining physical maps and global transcription data, genome transcription maps can be constructed for Drosophila. Advances in microarray technology, proteomics, and single-nucleotide polymorphism genotyping have made it attractive to correlate these molecular data with eQTL. Gene-expression levels can also be used to identify objectively the most promising candidate genes for complex traits. Because the number of candidate genes that physically reside in linked regions is often large, it seems appropriate to restrict attention to those genes in which the expression levels also map to the region and show genetic correlation with the phenotypes. Studies that combine gene product data with genetic marker data are the new frontier in dissecting the genetic basis of QTL.

References

1 Bent AF, Kunke BN, Dahlbeck D, Brown KL, Schinidt R et al. (1994) RPS2 of *Arabidopsis thaliana* a leucine-rich repeat class of plant disease resistance gene, *Science* 265:1856–1860

2 Martin GB, Brommonschenkel SH, Chunwongse J, Frary A, Ganal MW et al. (1993) Map-based cloning of a protein kinase gene conferring disease resistance in tomato, *Science* 262:1432–1436

3 Song WY, Wang GL, Chen LL, Kim HSPi, LY et al. (1995) A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa-21, Science* 270:1804–1806

4 Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in human using restriction fragment length polymorphisms, *Am J Hum Genet* 32:314–331

5 Warrington JA, Bengtsson U (1994) High-resolution physical mapping of human 5q31-q33 using three methods: Radiation hybrid mapping, interphase fluorescence *in situ* hybridization, and pulse-field gel electrophoresis, *Genomics* 24:395–398

6 Laroche A., Demeke T, Gaudet DA, Puchalski B, Frick M, McKenzie R (2000) Development of a PCR marker for rapid identification of the Bt-10 gene for common bunt resistance in wheat, *Genome* 43:217–223

7 Huys G, Rigouts L, Chemlal K, Portaels F, Swings J (2000) Evaluation of amplified fragment length polymorphism analysis for inter- and intraspecific differentiation of *Mycobacterium bovis*, *M. tuberculosis*, and *M. ulcerans*. *J Clin Microbiol* 38:3675–3680

8 Mackill DJ, Zhang Z, Redona ED, Colowit PM (1996) Level of polymorphism and genetic mapping of AFLP markers in rice, *Genome* 39:969–977

9 Ranamukhaarachchi DG, Kane ME, Guy CL, Li QB (2000) Modified AFLP technique for rapid genetic characterization in plants, *Biotechniques* 29:858–859, 862–866

10 Orita M, Iwahara H, Kanazava H, Hayashi K, Sekiya T (1989) Detection of polymorphism of human DNA by gel electrophoresis as single-strand conformation polymorphisms, *Proc Natl Acad Sci USA* 86:2766–2770

11 Fujita K, Silver J (1994) Single-strand conformational polymorphism, *PCR Methods Appl* 4:S137–S139

12 Hayashi K (1991) PCR-SSCP A simple and sensitive method for detection of mutation in the genomic DNA, *PCR Methods Appl* 1:34–38

13 Hayash K (1992) PCR-SSCP A method for detection of mutations, *GATA* 9:73–79

14 Bodenes C, Laigret F, Kremer A (1996) Inheritance and molecular variation of PCR-SSCP fragments in pedunculate oak (*Quercus robur* L.), *Theor Appl Genet* 993:348–354

15 Fukuoka S, Inoue T, Miyao A, Zhong HS, Saki T, Minobe Y (1994) Mapping sequence-tagged sites in rice by single strand conformation polymorphisms, *DNA Res* 1:271–274

16 Urquhart BG, Williams JL (1996) Sequencing of a novel cDNA and mapping to bovine chromosome 3 by single-strand conformation polymorphism (SSCP), *Anim Genet* 27:438

17 Cause MA, Fulton TM, Cho YG, et al. (1994) Saturated molecular map of the rice genome based on an interspecific backcross population, *Genetics* 138:1251–1274

18 Heun M, Keneedy AE, Anderson JA et al. (1991) Construction of an RFLP map of barley (*Hordeum vulgare* L.), *Genome* 34:437–447

19 Kurata N, Nagamura Y, Yamamoto K, Harushima Y, Sue N et al. (1994) A 300

kilobase interval genetic map of rice including 880 expressed sequences, *Nature Genet* 8:365–372

20 Lisitsyn N (1995) Representational difference analysis: finding the differences between genomes, *Trends Genet* 11:303–307

21 Liu BH (1998) Statistical Genomics: Linkage, Mapping and QTL Analysis. New York: CRC press

22 Larsen SO (1979) A general program for estimation of haplotype frequencies from population diploid data, *Comput Programs Biomed* 10:48–54

23 Lathrop GM, Lalouel JM, White RL (1986) Construction of human linkage maps: likelihood calculations for multilocus linkage analysis, *Genet Epidemiol* 3:39–52

24 Springer PS, Edwards KJ, Bennetzen JL (1994) DNA class organization on maize *Adh1* yeast artificial chromosomes, *Proc Natl Acad Sci USA* 87:103–107

25 Huh M (2000) Maximum likelihood vs. minimum chi-square – a general comparison with applications to the estimation of recombination fractions in two-point linkage analysis, *Genome* 43:853–856

26 Morgan TH (1928) The Theory of Genes. New Heaven, CT: Yale University Press

27 Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors, *J Genet* 8:299–309

28 Kosambi DD (1944) The estimation of map distances from recombination values, *Ann Eugen* 12:172–175

29 Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ et al. (1987) MAP-MAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations, *Genomics* 1:174–181

30 Liu BH, Knapp SJ (1990) GMENDEL: a program for Mendelian segregation and linkage analysis of individual or multiple progeny populations using log–likelihood ratios, *J Heredity* 81:407

31 Lathrop GM, Lalouel JM (1984) Easy calculation of lod scores and genetic risks on small computers, *Am J Hum Genet* 36:460–465

32 Gessler DD, Xu S (1999) Multipoint genetic mapping of quantitative trait loci with dominant markers in outbred populations, *Genetica* 105:281–291

33 Gibson S, Somerville C (1994) Isolating plant genes, *Trends Biotechnol* 12:306–313

34 Kornneef M (1994) *Arabidopsis* genetics. In: Meyerowitz EM, Somerville CR (Eds) *Arabidopsis*, Cold Spring Harbor, NY: Cold Spring Harbor Press, pp. 5.89–5.120

35 Tanksley SD, Ganal MW, Prince JP, de Vicente MC, Bonierbale MW et al. (1992) High density molecular linkage maps of the tomato and potato genome, *Genetics* 132:1141–1160

36 Sherman JK, Fenwick AL, Namuth DM, Lapitan NLV (1995) A barley RFLP map, alignment of three barley maps and comparisons to *Gramineae* species, *Theor Appl Genet* 86:705–712

37 Grant MR, Godiard L, Straube E, Ashfield T, Lewald J et al. (1995) Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance, *Science* 296:843–846

38 Bird AP (1987) CpG islands as gene markers in the vertebrate nucleus, *Trends Genet* 3:342–347; Bittner EP, Can C, Gunn N, Pinel M, Tor M et al. (1999) Genetic and physical mapping of the *RPP13* locus in *Arabidopsis*, responsible for specific recognition of several *Peronospora parasitica* (downy mildew) isolates *Mol Plant Microb Interact* 12:792–802

39 Burke DT (1990) YAC cloning options and problems, *GATA* 7:94–99

40 Buschges R, Hollricher K, Panstruga R, Simons G, Wolter M et al. (1997) The barley *Mlo* gene A novel control element of plant pathogen resistance, *Cell* 88:695–705

41 Simons G, Groenendijk J, Wijbrandi J, Reijans M, Groenen J et al. (1998)Dissection of the *Fusarium I2* gene cluster in tomato reveals six homologs and one active gene copy, *Plant Cell* 10:1055–1068

42 Pillen K, Ganal MW, Tanksley SD (1996) Construction of a high-resolution genetic map and YAC-contigs in the tomato *Tm-2a* region, *Theor Appl Genet* 93:228–233

43 Mesbah LA, Kneppers TJA, Takken FLW, Laurent P, Hille J, Nijkamp HJJ (1999) Genetic and physical analysis of a YAC contig spanning the fungal disease resistance locus *Asc* of tomato (*Lycopersicon esculentum*), *Mol Gen Genet* 261:50–57

44 Monna L, Miyao A, Zhong HS, Yano M, Iwamoto M et al. (1997) Saturation mapping with subclones of YACs: DNA marker production targeting the rice blast disease resistance gene, *Pi-b*, *Theor Appl Genet* 94:170–176

45 Dietrich RA, Richberg MH, Schmidt R, Dean C, Dang JL (1997) A novel zinc finger protein is encoded by the *Arabidopsis LSD1* gene and functions as a negative regulator of plant cell death, *Cell* 88:685–694

46 Lahaya T, Shirasu K, Schulze-Lefert P (1998) Chromosome landing at the barley *Rar1* locus, *Mol Gen Genet* 260:92–101

47 Cai D, Kleine M, Kifle S, Harloff HJ, Sandal NN et al. (1997) Positional cloning of a gene for nematode resistance in sugar beet, *Science* 275:832–834

48 Kaga A., Ishimoto M (1998) Genetic localization of a bruchid resistance gene and its relationship to insecticidal cyclopeptide alkaloids, the vignatic acids, in mungbean (*Vigna radiata* L. Wilczek), *Mol Gen Genet* 258:378–384

49 Leung J, Bouvier-Durand M, Morris PC, Guerrier D, Chefdor F, Giraudat J (1994) *Arabidopsis* ABA response gene *ABI1*: features of a calcium-modulated protein phosphatase, *Science* 264:1448–1451

50 Leung J, Merlot S, Giraudat J (1997) The *Arabidopsis ABSCISIC ACID-INSENSITIVE 2 (ABI2)* and *(ABI1)* genes encode redundant protein phosphatases 2C involved in abscisic acid signal transduction, *Plant Cell* 9:759–771

51 Giraudat J, Hauge BM, Valon C, Smalle J, Pacy F, Goodman HM (1992) Isolation of the *Arabidopsis ABI3* gene by positional cloning. *Plant Cell* 4:1251–1261

52 Finkelstein RR, Wang ML, Lynch T, Rao S, Goodman HM (1998) The *Arabidopsis* abscissic acid response locus *ABI4* encodes an APETALA2 domain protein, *Plant Cell* 10:1043–1054

53 Leyser HMO, Lincoln CA, Timpte C, Lammer D, Turner J, Estelle M (1993) *Arabidopsis* auxin-resistance gene *AXR1* encodes a protein related to ubiquitin-activating enzyme E1, *Nature* 364:161–164

54 Chang C, Kwok S, Bleecker A, Meyerowitz E (1993) *Arabidopsis* ethylene-response gene ETR1: Similarity of product to two-component regulators, *Science* 262:539–544

55 Kieber JJ, Rothenberg M, Roman G, Feldmann KA, Ecker JR (1993) *CTR1*, a negative regulator of the ethylene response pathway in *Arabidopsis*, encodes a member of the raf family of protein kinases, *Cell* 72:427–441

56 Peng J, Carol P, Richards DE, King KE, The arabidopsis GAI gene difines a signal pathway that negatively regulates gibberellin responses, *Genes Dev* 11:3194–3205

57 Sun TP, Goodman HM, Ausubel FM (1992) Cloning the *Arabidopsis GA1* locus by genomic subtraction, *Plant Cell* 4:119–128

58 Avramova Z, Tikhonov A, SanMiguel P, Jin YK, Liu C, et al. (1996) Gene identification in a complex chromosomal continuum by local genomic cross-referencing, *Plant J* 10: 1163–1168

59 Mao L, Begum D, Chuang HW, Budiman MA, Szymkowiak EJ et al. (2000) *Jointless* is a *MADS-box* gene controlling tomato flower abscission zone development, *Nature* 406:910–913

60 Ashikari M, Wu J, Yano M, Sasaki T, Yoshimura A (1999) Rice gibberellin insensitive dwarf mutant gene Dwarf 1 encodes the alpha subunit of GTP-binding protein Proc Natl Acad Sci USA 96:10284–10289

61 Tanksley SD, Ganal MW, Martin GB (1995) Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes, *Trends Genet* 11:63–68

62 Thomas CM, Vos P, Zabeau M, Jones DA, Norcott KA et al. (1995) Identification of amplified restriction length polymorphism (AFLP) markers tightly linked to the tomato *Cf-9* gene for resistance to *Cladosporium fulvum*, *Plant J* 8:785–794

63 Rosemberg M, Przybylska M, Straus D (1994) "RFLP Subtraction": A method for making libraries of polymorphic markers, *Proc Natl Acad Sci USA* 91:6113–6117

64 Corrette-Bennett J, Rosenberg M, Przybylska M, Ananiev E, Straus D (1998) Positional cloning without a genome map using 'Targeted RFLP Subtraction' to isolate dense markers tightly linked to the reg A locus of *Volvox carteri*, *Nucleic Acids Res* 26:1812–1818

65 Burke DT, Carle GF, Olson MV (1987) Cloning of large segments of exogenous DNA into yeast using artificial-chromosome vectors, *Science* 236:806–812

66 Green ED, Riethman HC, Dutchik JE, Olson MV (1991) Detection and characterization of chimeric yeast artificial-chromosome clones, *Genomics* 11:658–669

67 Neil DL, Villasante A, Fisher RB, Vetrie D, Cox B, Tyler-Smith C (1990) Structural instability of human tandemly repeated DNA sequences cloned in yeast artificial chromosome vectors, *Nucleic Acids Res* 18:1421–1428

68 Cai L, Taylor JF, Wing RA, Gallagher DS, Woo SS, Davis SK (1995) Construction and characterization of a bovine bacterial artificial chromosome library, *Genomics* 29:413–425

69 Jiang J, Gill BS, Wang GL, Ronald PC, Ward DC (1995) Metaphase and interphase fluorescence *in situ* hybridization mapping of the rice genome with bacterial artificial chromosomes, *Proc Natl Acad Sci USA* 92:4487–4491

70 Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T et al. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using F-factor-based vector, *Proc Natl Acad Sci USA* 89:8794–8797

71 Woo SS, Jiang J, Gill BS, Paterson AH, Wing RA (1994) Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*, *Nucleic Acids Res* 22:4922–4931

72 Kilian A, Kudrna DA, Kleinhofs A, Yano M, Kurata N et al. (1995) Rice/barley synteny and its application to saturation mapping of the barley Rpg1 region, *Nucleic Acids Res* 23:2729–2733

73 Choi S, Creelman RA, Mullet JE, Wing RA (1995) Construction and characterization of bacterial artificial chromosome library of *Arabidopsis thaliana*, *Plant Mol Biol Rep* 13:124–128

74 Zhang HB, Choi S, Woo SS, Li Z, Wing RA (1996) Construction and characterization of two rice bacterial artificial chromosome libraries from the parents of a permanent recombinant inbred mapping population, *Mol Breeding* 2:11–24

75 Williams JGK, Kubelik AR, Livak KJ, Rafalshi JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers, *Nucleic Acids Res* 18: 6531–6535

76 Shibata D, Liu YG (2000) *Agrobacterium*-mediated plant transformation with large DNA fragments, *Trends Plant Sci* 5:354–357

77 Sawa S, Watanabe K, Goto K, Kanaya E, Hayato M, Okada K (1999) *FILAMENTOUS FLOWER*, a meristem and organ identity gene of *Arabidopsis*, encodes a protein with a zinc finger and HMG-related domains, *Genes Dev* 13:1079–1088

78 Hamilton CM, Frary A, Lewis C, Tanksley SD (1996) Stable transfer of intact high molecular weight DNA into plant chromosomes, *Proc Natl Acad Sci USA* 93:9975–9979

79 Choi S, Begum D, Koshinsky HODW, Wing RA (2000) A new approach for the identification and cloning of genes the pBACwich system using Cre/lox site-specific recombination, *Nucleic Acids Res* 28:19e

80 Koshinsky HA, Lee E, Ow DW (2000) Cre-lox site-specific recombination between *Arabidopsis* and tobacco chromosomes, *Plant J* 23: 715–722

81 Larsen F, Gunderson G, Lopez R, Prydz H (1992) CpG islands are gene markers in human genome, *Genomics* 13:1095–1107

82 Duyk GM, Kim S, Myers RM, Cox DR (1990) Exon trapping A genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA, *Proc Natl Acad Sci USA* 87:8995–8999

83 Nehls M, Pfeifer D, Boehm T (1994) Exon amplification from complete libraries of genomic DNA using a novel phage vector with automatic plasmid excision facility: Application to mouse neurofibromatosis-1 locus, *Oncogene* 9:2169–2175

84 Krizman DB, Berget SM (1993) Efficient selection of 3′ terminal exons from vertebrate DNA, *Nucleic Acids Res* 21:5198–5202

85 Ahn S, Tanksley SD (1993) Comparative map of the rice and maize genomes. *Proc Natl Acad Sci USA* 90:7980–7984

86 Guimaraes CT, Sills GR, Sobral BWS (1997) Comparative mapping of Andropogoneae: *Saccharum* L. (sugarcane) and its relation to sorghum and maize, *Proc Natl Acad Sci USA* 94:14261–14266

87 Van Deynze AE, Nelson JC, Yglesias ES, Harrington SE, Braga DP et al. (1995) Comparative mapping in grasses: Wheat relationships, *Mol Gen Genet* 284:744–754

88 Carver EA, Stubbs L (1997) Zooming in on the human/mouse comparative map: Genome conservation re-examined on a high-resolution scale, *Genome Res* 7:1123–1137

89 Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes microcolinearity and its many exceptions, *Plant Cell* 12:1021–1029

90 Moore G, Gale MD, Kurata N, Fravell RB (1993) Molecular analysis of small grain cereal genomes: current status and prospects, *Biotechnology* 11:584–589

91 Ananiev EV, Riera-Lizarazu O, Rines HW, Phillips RL (1997) Oat/maize chromosome addition lines A new system for mapping the maize genome, *Proc Natl Acad Sci USA* 94:3524–3529

92 Ahn S, Anderson JA, Sorrells ME, Tanksley SD (1993) Homologous relationships of rice wheat and maize chromosomes, *Mol Gen Genet* 241:481–490

93 Dunford RP, Kurata N, Laurie DA, Money TA, Minobe Y, Moore G (1995)Conservation of fine-scale DNA marker order in the genome of rice and the Triticeae, *Nucleic Acids Res* 115:133–138

94 Bennetzen JL, Freeling M (1993) Grasses as a single genetic system genome composition colinearity and compatibility, *Trends Genet* 9:259–261

95 Paterson AH, Lin YR, Li Z, Schertz KF, Doebley JF et al. (1995) Convergent domestication of cereal crops by independent mutations at corresponding genetic loci, *Science* 269:1714–1718

96 Moore G, Devos KM, Wang Z, Gale MD (1995) Grasses, line up and form a circle, *Curr Biol* 5:737–739

97 Chen M, SanMiguel P, Liu CN, de Oliveira AC, Tikhonov A et al. (1997) Microcolinearity in the maize, rice, and sorghum genomes, *Proc Natl Acad Sci USA* 94:3431–3435

98 Peng J, Richards DE, Hertley NM, Murphy GP, Devos KM et al. (1999) *Nature* 400:256–261

99 Debry RW, Seldin MF (1996) Human mouse homology relationships, *Genomics* 33:337–351

100 Bassett Jr DE, Boguski MS, Hieter P (1996) Yeast genes and human disease, *Nature* 379:589

101 Raymond M, Gros P, Whiteway M, Thomas DY (1992) Functional complementation of yeast *ste6* by a mammalian multidrug resistance *mdr* gene, *Science* 256:232–235

102 Ballester R, Marchuk D, Boguski M, Saulino A, Letcher R et al. (1990) The NF1 locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins, *Cell* 63:851

103 Mounkes LC, Jones RS, Liang BC, Gelbart W, Fuller MT (1992) A *Drosophila* model for xeroderma pigmentosum and Cockayne's syndrome: haywire encodes the fly homolog of ERCC3, a human excision repair gene, *Cell* 71:925

104 Salse J, Piegu B, Cooke R, Delseny M (2002) Synteny between *Arabidopsis* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project, *Nucleic Acids Res* 30:2316–2328

105 Laurie DA, Pratchett N, Bezant JH, Snape JW (1995) RFLP mapping of five major genes and eight quantitative trait loci controlling flowering time in a winter x spring barley (*Hordeum vulgare* L.) cross, *Genome* 38:575–585

106 Borner A, Korzun V, Worland AJ (1998) Comparative genetic mapping of loci affecting plant height and development in cereals, *Euphytical* 100:245–248

107 Laurie DA (1997) Comparative genetics of lowering time in cereals, *Plant Mol Biol* 35:167–177

108 Yamamoto T, Kuboki Y, Lin Y, Sasaki T, Yano M (1998) Fine mapping of quantitative trait loci *Hd1, Hd2, and Hd3,* controlling heading date of rice, as single Mendelian factors, *Theor Appl Genet* 97:37–44

109 Rounsley S, Lin X, Ketchum KA (1998) Large-scale sequencing of plant genomes, *Curr Opin Plant Biol* 1:136–141

110 Hazen SP, Kay SA (2003) Gene arrays are not just for measuring gene expression, *Trends Plant Sci* 8:413–416

111 Winzeler EA et al. (2003) Genetic diversity in yeast assessed with whole-genome oligo-nucleotide arrays, *Genetics* 163:79–89

112 Borevitz JO et al. (2003) Large-scale identification of single-feature polymorphisms in complex genomes, *Genome Res* 13:513–523

113 Zivy M, Prioul JL, Cornic G, Westhoff P, Lacroix B, de Vienne D (1995) Characterizing proteins induced by drought in maize and search for their QTLs, *Proc First Int Congr Integrated Studies on Drought Tolerance of Higher Plants*, August 31–September 2, Montpellier, France

114 Kawasaki S, Borcher C, Beyholos M, Wang H, Brazille S, Kawai K, Galbraith D, Bohnert HJ (2001) Gene expression profiles during the initial phase of salt stress in rice, *Plant Cell* 13:889–905

115 Negishi T, Nakanishi H, Yazaki J, Kishimoto N, Fujii F, Shimbo K, Yamamoto K, Sakata K, Sasaki T, Kikuchi S et al. (2002) cDNA microarray analysis of gene expression during Fe-deficiency stress in barley suggests that polar transport of vesicles is implicated in phytosiderophore secretion in Fe-deficient barley roots, *Plant J* 30:83–94

116 Hertzberg M, Aspeborg H, Scharder J, Anderson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlen M, Teeri TT, Lundeberg J et al. (2001) A transcriptional roadmap to wood formation, *Proc Natl Acad Sci USA* 98:14732–14737

117 Aharoni A, Keizer LC, Bouwmeester HJ, Sun Z, Alvarez-Huerta M, Verhoeven HA, Blaas J, van Houwelingen AM, De Vos RC, van der Voet H et al. (2000) Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays, *Plant Cell* 12:647–662

118 Schadt EE et al. (2003) The tranginics of gene expression surveyed in maize, moues and man, *Nature* 422:297–302

119 Wang ZY, Zheng FQ, Shen GZ, Gao JP, Snustad DP, Li MG, Zhang JL, Hong MM (1995) The anymose content in rice endosperm is related to the post-transcriptional regulation of the *Waxy* gene, *Plant J* 7:613–622

6
DNA Sequencing Technology

Lyle R. Middendorf, Patrick G. Humphrey, Narasimhachari Narayanan, and Stephen C. Roemer

6.1
Introduction

DNA sequencing technology is a major component of the genomics discovery pipeline. The technology is rooted in the late 1960s and early 1970s when efforts were made to sequence RNA. Nucleotide sequences of 5S-ribosomal RNA from *Escherichia coli* [1], 16S- and 23S-ribosomal RNA [2], and R17 bacteriophage RNA coding for coat protein [3] are some early examples of RNA sequencing. A few years later Sanger reported on the sequencing of bacteriophage f1 DNA by primed synthesis with DNA polymerase [4, 5]. At the same time Gilbert and Maxam [6] reported the DNA nucleotide sequence of the *lac* operator.

This pioneering work led to the plus/minus method reported by Sanger and Coulson [7] which determined nucleotide sequence on the basis of two approaches:

1. a "minus" system in which four separate samples of partially double-stranded DNA fragments (containing a "full length" template "–" strand and random chain extension of an oligonucleotide primer for the "+" strand) were further incubated with DNA polymerase in the presence of only three deoxyribonucleoside triphosphates such that synthesis proceeded as far as it could until the polymerase needed to incorporate the missing nucleotide of the particular sample; and
2. a "plus" system in which the four separate samples of partially double-stranded DNA fragments were further incubated in the presence of only one of the four triphosphates and then subjected to exonuclease activity which degraded the single-stranded overhang of the "-" strand and any double-stranded DNA from its 3′ end until it stopped at a residue corresponding to the one triphosphate present.

The DNA fragments for both approaches were then subjected to gel electrophoresis for length (and thus sequence) determination. One limitation of the plus/minus method was that incomplete representation of all possible starting lengths within the initial population of the partially double-stranded DNA fragments, perhaps because of the sequence context dependency of the

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

polymerase kinetics, would result in missing products. Another limitation of the plus/minus method was its inability to provide accurate sequence assessment of multiple runs of a given base type.

In 1977 Sanger reported the use of modified nucleoside triphosphates (containing dideoxyribose sugar) in combination with natural deoxyribonucleotides to terminate chain elongation, thus overcoming the limitations of the plus/minus method. In that same year Maxam and Gilbert [8] disclosed a method for sequencing DNA that utilized chemical cleavage of DNA preferentially at guanines, at adenines, at cytosines and thymines equally, and at cytosines alone. These two methods accelerated manual sequencing based on electrophoretic separation of DNA fragments labeled with radioactive markers and subsequent detection via autoradiography.

The first reports of automation of DNA sequencing occurred in the mid-1980s, because of novel techniques to fluorescently label DNA [9–15]. This automation, in conjunction with the commencement of the human genome initiative [16], spurred the explosion in genomics research that is in existence today. DNA sequencing technology is now only one tool, albeit a very important and dynamic one, in the genomics toolbox, along with other tools such as DNA array and lab-on-a-chip technologies and automated protein analysis.

This chapter illustrates the multi-disciplinary nature of DNA sequencing technology in that its organization is delineated into chemistry, biology, instrumentation, and software components. It is intended to provide an exhaustive reference structure to enable further in-depth investigation of each of these components, and the reader is invited to take advantage of the reference list to capture the full essence of sequencing technologies.

6.2 Overview of Sanger Dideoxy Sequencing

DNA sequencing is the determination of the nucleotide sequence of a specific deoxyribonucleic acid (DNA) molecule. Knowing the sequence of a DNA molecule is pivotal for making predictions about its function and facilitating manipulation of the molecule. Originally, DNA was sequenced using one of two methods. Maxam and Gilbert [8] devised a method that chemically cleaved DNA selectively between specific bases. Sanger et al. [17] developed an enzymatic method based on the use of chain-terminating dideoxynucleotides.

The Sanger dideoxy method is now by far the most widely used technique for sequencing DNA. Informative texts by Alphey [18] and Ansorge et al. [19] review many variations made to this sequencing technique, but the principle remains the same. The method depends on the synthesis of a new strand of DNA starting from a specific priming site and ending with the incorporation of a chain-terminating nucleotide.

Specifically, a DNA polymerase extends an oligonucleotide primer annealed to a unique location on a DNA template by incorporating deoxynucleotides complementary to the template. Synthesis of the new DNA strand continues until the reaction is randomly terminated by inclusion of a dideoxynucleotide. These nucleotide analogs are incapable of supporting further chain elongation because the ribose moiety of the incorporated dideoxynucleotide lacks the 3′-hydroxyl necessary for forming a phosphodiester bond with the next incoming deoxynucleotide. This results in a population of truncated sequencing fragments of different length.

Typically, the identity of the chain-terminating nucleotide at each position is specified by running four separate base-specific

reactions, each of which contains a different dideoxynucleotide (ddATP, ddCTP, ddGTP, or ddTTP). The four such fragment sets are loaded in adjacent lanes of a polyacrylamide gel and separated by electrophoresis according to fragment size (Fig. 6.1). Remarkably, DNA fragments differing in length by just one nucleotide can be resolved. If a radioactive label is introduced into the sequencing reaction products, autoradiographic imaging of the DNA band pattern in the gel can be used to deduce the DNA sequence [17, 20]. If the reaction products are labeled with an appropriate fluorescent dye, an automated DNA sequencing system is used for real-time detection of DNA fragments as they move through a portion of the electrophoresis gel irradiated by a laser. The fluorescence emission is collected by a detector and the resulting signal produces a band or trace pattern which correlates to a DNA sequence.

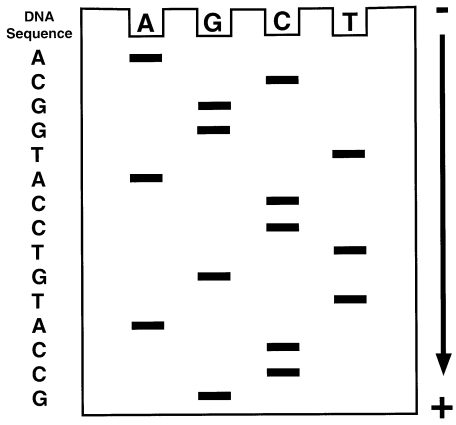


Fig. 6.1 DNA sequencing electrophoresis. The DNA fragments are prepared to terminate at one of four base types (A, G, C, T). A-type fragments of different length are loaded in the "A" loading well at the top of the gel, and so forth for the G-, C-, and T-type fragments. Over time the shorter fragments in each lane migrate farther down the gel (toward the positive electrode). The DNA sequence is determined by noting the particular lane in which each succeeding band is spatially located in the vertical dimension. (Taken from Ref. [21].)

6.3 Fluorescence Dye Chemistry

The original methods of DNA sequencing [8, 17] were implemented through the use of radioactive labels. High sensitivity and ease of labeling initially made radioactive methods popular in thousands of biology laboratories around the world that practiced manual radioactive DNA sequencing. The dangers associated with radioactivity such as health hazards and waste-disposal regulations, along with the lack of automation, however, paved the way for the emergence of alternative non-radioactive labels [22]. Most prominent among the sensitive, non-radioactive detection techniques are chemiluminescence and fluorescence. Despite excellent sensitivity, chemiluminescence methodology is not viable for DNA sequencing because of its indirect detection limitation. Fluorescent detection [23], on the other hand, employs direct detection methodology that is simple, sensitive, and easy to automate. Fluorescence methods and fluorescent dye labels have set a new standard in today's DNA sequencing community.

Several methods have been developed for sequencing DNA by using fluorescent labels [9, 10, 12, 13]. Commercial instruments employ one or more of the following methods for automated sequencing:

- four distinct dye-labeled primers with non-fluorescent terminators per DNA sample;
- one dye labeled primer with non-fluorescent terminators per DNA sample; and
- one non-fluorescent primer with four distinct fluorescent terminators per DNA sample (Sect. 6.4).

This section provides a brief summary of important aspects of the advancement of the chemistry of fluorescent dyes for DNA sequencing.

6.3.1 Fluorophore Characteristics

Fluorescence is the emission of light from electronically excited fluorophores. An electron of the fluorophore is energized into an excited orbital by absorption of a photon where it is paired with a second electron that is in the ground-state orbital [23]. The excited orbital is one of several vibrational energy levels associated with one or more electronic energy states. The fluorophore is usually excited into a higher vibrational level of either the first or second electronic energy state. In a very fast process known as internal conversion the excited molecule first relaxes to the lowest vibrational level of the first electronic energy state. This is followed by relaxation to a higher excited vibrational ground-state level with emission of a photon. Because of the multiplicity of vibrational levels and electronic levels, the spectra of both absorption and emission are polychromatic and are usually mirror images of each other.

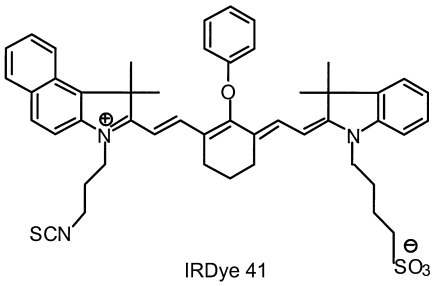
Both the absorption and emission spectra of the fluorophore depend on its chemical structure and the environment (solvent, pH, temperature, etc.) of the fluorophore. The spectral wavelength of fluorescence emission is generally independent of the excitation wavelength of the absorbed photons. Because of the rapid initial non-radiative decay associated with internal conversion and the final decay to higher vibrational levels of the ground state, however, the energy of the emitted photon is less than that of the absorbed photon. This shifts the fluorescence spectra to longer wavelengths relative to the absorption spectra and is known as the Stokes Shift [24].

6.3.2 Commercial Dye Fluorophores

The physiological response of the human eye qualitatively defines the visible wavelength region (in nanometers or nm) of the electromagnetic spectrum. Wavelengths shorter than, but adjacent to, that of the visible region, are denoted ultraviolet. Wavelengths longer than, but adjacent to, that of the visible region, are denoted near-infrared. The commercialized fluorescent labels currently in use in automated DNA sequencing are either visible dyes (450–650 nm absorption and fluorescence range) or near-infrared dyes (650–860 nm absorption and fluorescence range).

The first commercialized near-infrared dyes introduced for automated DNA sequencing were IRDye 41 and IRDye 40 [25–26], (Fig. 6.2). These dyes are from the heptamethine carbocyanine dye family and nominally absorb and fluoresce near 800 nm. IRDye41 was attached to a DNA primer via a stable thiourea linkage formed by conjugating the dye to an amino linker located at the 5′ end of the primer. A phosphoramidite version (IRDye800, Fig. 6.3) [28] enables direct labeling of DNA primers using an automated DNA synthesizer. For dye-labeled terminator chemistry, the IRDye 800 is attached to bases which are linked to a triphosphate through an acyclo bridge (Fig. 6.4). The incorporation of this substrate terminates DNA chain elongation in a manner similar to that achieved by using dideoxynucleotides (Sect. 6.4). Dye properties for IRDye 40, IRDye 41 and IRDye800 are listed in Tab. 6.1.

Commercialized near-infrared dyes that absorb and fluoresce around 650–700 nm



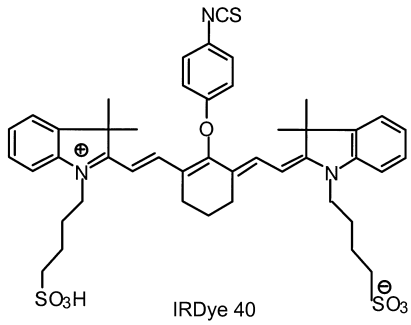


Fig. 6.2 Structures of IRDye41 and IRDye40. Both dyes are members of the polymethine carbocyanine dye family which is characterized by two heteroaromatic residues connected by a conjugation bridge of polyethylene units. The length of the conjugating bridge affects the absorbance and fluorescence maxima [29]. IRDye41 and IRDye40 are heptamethine carbocyanine dyes which contain seven carbons in their conjugating bridge. The isothiocyanate (NCS) reactive functionality is used to couple the dye to a primary amine which results in a thiourea linkage.

are from the pentamethine carbocyanine dye family. They include IRDye700 [28], Cy5 [30–32] and Cy5.5 [33]. Dye properties for IRDye700, Cy5, and Cy5.5 are listed in Tab. 6.1 and the structures are shown in Fig. 6.5.

Figure 6.6 shows two fluorescein dye derivatives (FAM, JOE) and two rhodamine dye derivatives (TAMRA ROX) first used for four visible dye primer-based DNA sequencing. Fluorescein dye has also been used in single dye sequencers (ALF DNA Sequencer, Amersham Biosciences). Figure 6.7 shows two rhodamine dyes (R110, R6G) which are combined with TAMRA and ROX for use in four visible color dye terminator-based DNA sequencing. Dye properties for FAM, JOE, TAMRA, ROX, R110, and R6G are listed in Tab. 6.1.

To furnish more even and narrower peak heights than the rhodamine dye terminators and to reduce spectral overlap among the dyes, a family of dichlororhodamine (dRhodamine) dyes has been designed [34]. These dyes (dR110, dR6G, dTAMRA, dROX) are distinguished from R110, R6G, TAMRA, and ROX by addition of two chlorides to the phenyl ring of the rhodamine [35]. Figure 6.8 shows the 4,7-dichloro-substituted R110 (dR110). Dye properties for dR110, dR6G, dTAMRA, and dROX are listed in Tab. 6.1.

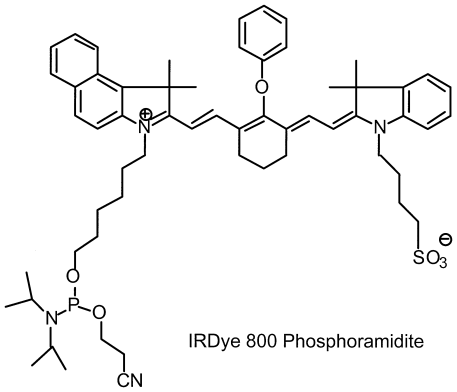


Fig. 6.3 Structure of IRDye800 phosphoramidite. The amidite functionality is used to couple the dye to the 5′-OH of the 5′ terminus nucleotide of an oligonucleotide via automated DNA synthesis. Further information is given in the legend to Fig. 6.2.

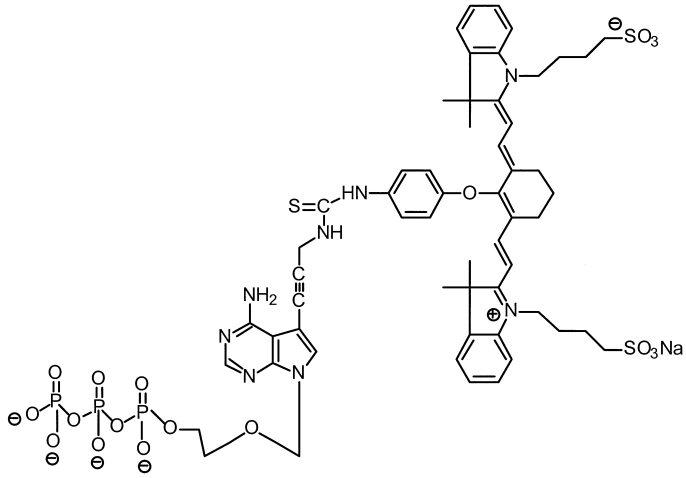
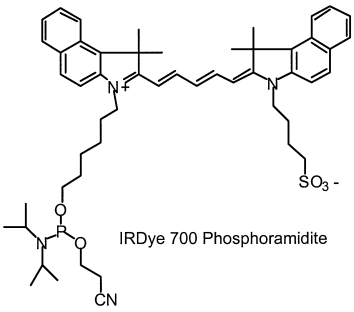


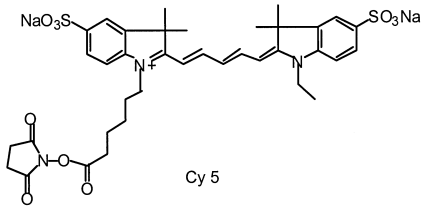
Fig. 6.4 Structure of IRDye800-acyclo-ATP. The dye is linked to an adenine base, which in turn is linked to a triphosphate. There is no ribose sugar. IRDye800-acyclo-CTP, IRDye800-acyclo-GTP, and IRDye800-acyclo-UTP are similarly synthesized with their respective base type. All four molecules are suitable substrates for chain elongation by DNA polymerase, but on incorporation into the growing DNA strand, they terminate synthesis.

Table 6.1 Dye absorption and emission properties (aqueous environment) for several commercial dyes available for DNA sequencing. Absorption and emission maxima are approximate and might be dependent on solvent, solvent properties (e.g. pH) and the biomolecule to which they are attached. NA = information not available.

Dye	*Absorption max. (nm)*	*Emission max. (nm)*	*Dye family*
FAM	490–495	515–520	Fluorescein
R110	500–505	525–530	Rhodamine 110
dR110	NA*	530–535	Rhodamine 110
JOE	520–525	550–555	Dichlorodimethylfluorescein
R6G	525–530	555–560	Rhodamine 6G
dR6G	NA	560–565	Rhodamine 6G
TAMRA	550–555	580–585	Tetramethylrhodamine
dTAMRA	NA	590–595	Tetramethylrhodamine
ROX	580–585	605–610	X-Rhodamine
dROX	NA	615–620	X-Rhodamine
Cy5	650–655	665–670	Pentamethine carbocyanine
Cy5.5	670–675	690–695	Pentamethine carbocyanine
IRDye700	685–690	710–715	Pentamethine carbocyanine
IRDye40	765–770	785–790	Heptamethine carbocyanine
IRDye41	795–800	820–825	Heptamethine carbocyanine
IRDye800	795–800	820–825	Heptamethine carbocyanine

* Information not available





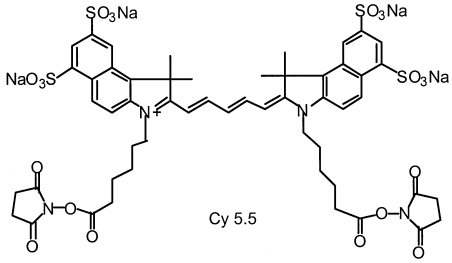
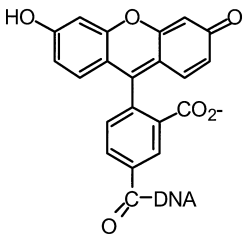
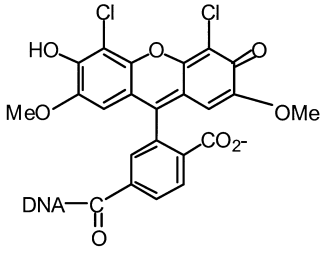
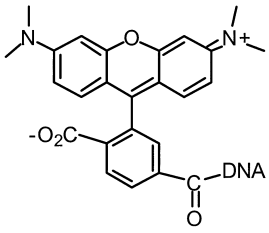


Fig. 6.5 Structures of IRD700 (phosphoramidite functionality), Cy5 (succinimidyl ester functionality), and Cy5.5 (bis-succinimidyl ester functionality). All three dyes are members of the pentamethine carbocyanine dye family which is characterized by five carbons in the conjugating bridge (Fig. 6.2 legend).







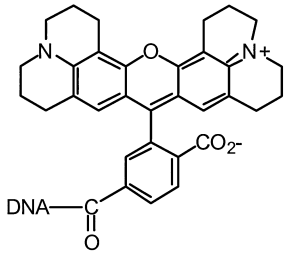


Fig. 6.6 Structures of the dyes FAM, JOE, TAMRA, and ROX. FAM and JOE are members of the fluorescein family whereas TAMRA and ROX are members of the rhodamine family. All four dyes must be purified from isomers that contain alternate sites for the reactive functionality which ultimately couples the dye to DNA. Shown are the 5-isomer for FAM and the 6-isomer for JOE, TAMRA, and ROX.

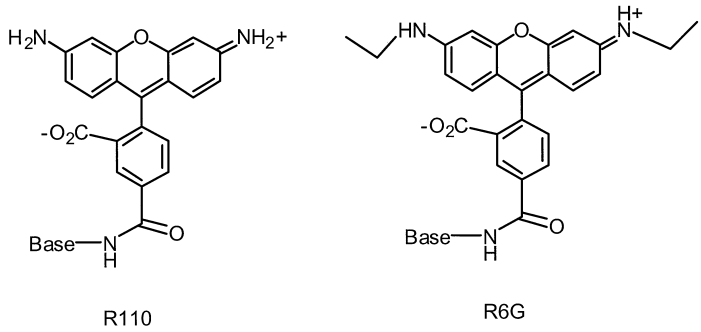


Fig. 6.7 Structures of the R110 and R6G dyes. Both dyes are members of the rhodamine family. Shown are the 5-isomers.

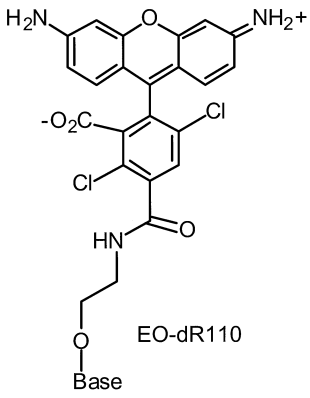


Fig. 6.8 Structure of dichoro-R110 dye linked to a nucleotide base. See text for the effects of adding the two chlorides to R110 (shown in Fig. 6.7). Similar dichloro modifications have been made to dyes TAMRA, ROX, and R6G (shown in Figs. 6.6 and 6.7).

6.3.3 Energy Transfer

When using four-color discrimination the spectral overlap of fluorescence emission of the four fluorophores reduces the signal-to-noise ratio and therefore results in less accurate and shorter read lengths. Also, because the absorption spectra and molar absorptivity of the dyes are not equivalent, the use of a single excitation source for all four dyes compromises sequencing results, because of widely varying fluorescence signal strength.

One method for improving the properties of the dyes makes use of resonance energy transfer, an important process that occurs in the excited state of a fluorophore [36]. Energy transfer can occur between two molecules if the emission spectrum of an absorbing fluorophore (donor) overlaps the absorption spectrum of a nearby acceptor fluorophore. The donor and acceptor molecules are coupled by a dipole–dipole interaction [23]. In addition to spectral overlap, the rate of energy transfer depends on the distance between donor and acceptor and follows an inverse relationship to the 6th power of that distance.

An approach that involves energy transfer in labeled primer chemistry uses the oligonucleotide backbone to separate the donor

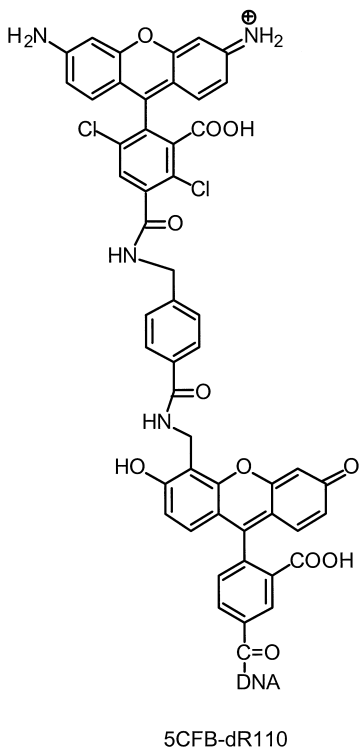


Fig. 6.9 Structure of dichloro-R110 linked to 4′-aminomethyl-fluorescein [34, 35]. This dual dye configuration enables fluorescence resonant energy transfer from fluorescein (donor) to R110 (acceptor) and is a member of a commercially available family of dyes trademarked as BigDyes (PE Biosystems). Other BigDyes are synthesized with dTAMRA, dROX, and dR6G as acceptors, all of which contain the dichloro modification.

and acceptor dyes [37–44]. Another approach uses tethered donor and acceptor dyes for either labeled primers [35] or labeled terminators [34]. These tethered dyes use fluorescein as a donor dye and one of the four dRhodamine dyes (Sect. 6.3.2) as an acceptor dye, and are linked through 4-aminomethyl benzoic acid. The structure of a tethered fluorescein/dR110 is shown in Fig. 6.9.

6.3.4 Fluorescence Lifetime

A fluorophore emits lights as it relaxes from an excited energy state to a ground energy state; such relaxation occurs after the molecule has spent a certain amount of time in the excited state (Sect. 6.3.1). The average time spent in the excited state is known as the fluorescence lifetime of the molecule [23] and it is statistically the same for all molecules having the same structure and exposed to the same environmental conditions. A common characteristic (although not necessarily assumable) is that statistical relaxation of a fluorophore follows an exponential decay profile when examined over several excitation/relaxation cycles. In this circumstance the fluorescence lifetime is specified as the exponential time constant where 63 % of the relaxations occur more quickly than this lifetime average and 37 % occur more slowly.

The lifetime of common visible and near-infrared fluorophores ranges from 0.5–4 ns and depends on their chemical structure. The ability to discriminate among fluorophores is affected by the ratio of their life-

times and the number of photons available to produce the composite lifetime profile histogram [45, 46].

The use of energy transfer to enable common excitation of several dyes has been successfully commercialized (Sect. 6.3.3). As researchers examine alternative approaches for facilitating common excitation and increasing the number of available dye choices, the exploitation of fluorescence lifetime discrimination for DNA sequencing shows promising potential, because the lifetime of a fluorophore is independent of concentration and multiple dyes having overlapping spectral emission can be distinguished [47–54]. Methods for "on-the-fly" lifetime measurements of labeled DNA fragments have been described for capillary electrophoresis [55, 56] and slab gel electrophoresis [57]. Besides enabling common excitation, fluorescence lifetime discrimination also enables the use of common spectral detection optics. Both spectral and lifetime discrimination can be combined in a single design to take advantage of the strengths of each approach [57, 58].

6.4 Biochemistry of DNA Sequencing

The efficient completion of large DNA sequencing projects is now a reality, mainly because of the development of fluorescence-based dideoxynucleotide sequencing chemistries coupled with instrumentation for real time detection of dye-labeled DNA fragments during gel electrophoresis (Sect. 6.5). The commercially available automated sequencers (Sect. 6.5; Tabs. 6.2 and 6.3) can be divided into two groups on the basis of the number of fluorescent dyes used in a sequencing reaction.

The first type uses the one-dye/four-lane approach in which the identity of the chain-terminating nucleotide at each position is determined by running four separate reactions each of which contains the same fluorescent dye but a different dideoxynucleotide (ddATP, ddTTP, ddGTP, ddCTP). The four completed sequencing reactions are loaded in separate lanes of a slab gel (Sect. 6.2, Fig. 6.1), and the automated sequencer must then be able to align the raw data from all four lanes precisely enough to determine the correct base sequence (Sect. 6.6.2).

The second type employs the four-dye/one-lane approach in which a single combined reaction is performed using a fluorescent label specific for each of the four dideoxynucleotides. The combined sequencing reaction can be analyzed in a single gel lane or capillary or microfluidic channel (Sect. 6.5.4), and the automated sequencer must first correct for the different mobility of the four dye-labeled DNA fragment sets before calling bases [59].

6.4.1 Sequencing Applications and Strategies

DNA sequencing is a fundamental technique in genome analysis and it has major applications which fall into two general classes: (1) *de novo* sequencing of unknown DNA, and (2) resequencing segments of DNA for which the sequence is already known. In both classes the DNA to be sequenced is first cloned into a viral or plasmid vector, or is part of an amplified PCR fragment (Sect. 6.4.2).

The approach used to sequence unknown DNA is termed the sequencing strategy and it should provide the correct consensus sequence on both strands of the target DNA using a minimal number of sequencing reactions with minimum overlap (Sect. 6.4.1.1). Large-scale sequencing projects make use of one or more sequencing strategies to completely characterize the entire genome

of an organism [60–63], including archaea [64] and the human genome [65, 66]. On the other hand, many laboratories employ methods of resequencing to characterize the variability of smaller, known DNA segments to find mutations or verify recombinant DNA constructs (Sect. 6.4.1.2).

6.4.1.1

New Sequence Determination

The selection of a sequencing strategy usually depends on the size of the target DNA. For example, random shotgun sequencing is currently the method used in most large-scale DNA sequencing projects [62, 67]. In shotgun sequencing a large segment of target DNA (e.g. a medium-sized BAC clone of 100–120 kilobases) is randomly fragmented by physical shearing or enzymatic digestion into fragment sizes in the range 1 to 5 kilobases. These smaller fragments are then subcloned into bacteriophage M13 or plasmid vectors (Sects. 6.4.2.1 and 6.4.2.2). The cloned inserts are sequenced from "universal" primer binding sites in the flanking vector DNA, and the resulting sequence information compiled by computer into contiguous sequences (i.e. "contigs") to reassemble the original large target DNA.

This method rapidly generates 95 % of the desired sequence, but becomes less efficient as each subsequent random subclone is more likely to yield sequence information already obtained. Typically, each base in the target DNA sequence is read an average of four to six times during this "working draft" phase of the shotgun-sequencing project. Gaps or unresolved regions will still remain, however, and can be filled in by directed approaches during the "finishing" phase of the sequencing project [60, 63, 68, 69].

An adaptation of shotgun sequencing, called whole-genome shotgun assembly (WGSA or WGA), avoids the step of "mapping" the BAC clones by using advanced assembly software to connect the sequences of random clones from the entire genome, rather than from the smaller BAC clones. In WGSA, DNA sequence reads from clone libraries of 2 kilobase pairs (kbp), 10 kbp, and 50 kbp are assembled on to a scaffold-like structure to generate a contiguous sequence [66, 70].

Advantages of shotgun sequencing include no requirement for prior knowledge of the insert sequence and no limitation on the size of the starting target DNA. Additionally, a high degree of parallel processing and automation can be implemented during the initial random phase, with only one or two oligonucleotide sequencing primers required.

Primer walking is a fully directed sequencing strategy. It provides an efficient way to obtain new sequence information and is a good choice for the primary sequencing of small regions (1 to 3 kilobases) of genomic or cDNA clones or as a secondary approach to achieve closure and resolve local ambiguities after an initial shotgun-sequencing phase. Other approaches, for example the enzymatic nested deletion method [71] or transposon insertion [72], have also been used for small-scale *de-novo* sequencing.

The primer-directed method is initiated by sequencing the target DNA from one end using a vector-specific standard primer [73]. A new walking primer is designed using the most distant, reliable sequence data obtained from the first sequencing reaction with the standard primer. This walking primer is then used to sequence the next unknown section of the DNA template. Although, in theory, this primer walking process can be repeated many times to sequence extensive tracts of DNA, its use is generally limited to smaller projects because the successive rounds of sequence

analysis, primer design, and primer synthesis are too expensive and time-consuming [19].

The major benefits of primer walking are that no subcloning is required, the location and direction of each sequencing run is known, and the degree of redundancy needed to obtain the final sequence is minimized. Moreover, read lengths greater than 1000 bases have been reported [74–83] thus reducing the number of walking primers needed to finish a sequencing project.

6.4.1.2 Confirmatory Sequencing

The major purpose of DNA sequencing in many laboratories is to resequence small regions of interest (< 1 kb) using cloned DNA or a PCR product as the template. Resequencing is useful for applications such as confirming plasmid constructs, screening the products of site-directed mutagenesis experiments, or comparing sequences of wild-type and mutant variants associated with genetic disease [84–90]. Because the target region has often been characterized, it is possible to design a primer so that the sequence of interest is within 100 to 150 bases of the sequencing primer. This will provide optimum resolution in the raw sequence data generated by the automated DNA sequencer, and thus the highest base-calling accuracy that can be obtained (Sect. 6.6).

6.4.2 DNA Template Preparation

In the first step of a Sanger dideoxy sequencing reaction, the primer is annealed to a single-stranded DNA template (Sect. 6.2). DNA in this form can be purified directly from viruses such as bacteriophage M13 which have single-stranded genomes. On the other hand, double-stranded DNA such as a plasmid vector containing the target insert must first be converted to the single-stranded form, either by alkali or heat denaturation, before sequencing [19, 91].

The material presented in this section is intended to serve only as a general guide for preparing DNA templates. Specific procedures and applications can be found in several molecular biology manuals [19, 92–96].

6.4.2.1 Single-stranded DNA Template

Several variants of the bacteriophage M13 were constructed for the purpose of generating DNA template for dideoxy sequencing [97]. The DNA to be sequenced is cloned into the double-stranded replicative form of the phage, transformed into *E. coli*, and harvested in large quantity from the culture medium in the form of phage particles containing single-stranded DNA [98]. The purified DNA is ideal for sequencing, because it is single-stranded so that no complementary strand exists to compete with the sequencing primer during the annealing step. Moreover, a universal sequencing primer hybridizes to a complementary portion of the phage DNA immediately adjacent to the multiple cloning site. M13 is still used extensively for high-throughput sequencing applications [67].

6.4.2.2 Double-stranded DNA Template

Many methods have been developed for isolation and purification of plasmid DNA from bacteria [92]. Generally, the process involves five steps: (1) insert foreign (target) DNA into the plasmid vector, (2) transform a suitable bacterial strain with the recombinant plasmid, (3) grow the bacterial culture, (4) harvest and lyse bacteria, and (5) purify the plasmid DNA.

For sequencing applications, double-stranded plasmid DNA containing the tar-

get sequence must be of high purity. Contaminating salt, RNA, protein, DNAses, and polysaccharides from the host bacteria can inhibit dideoxy sequencing reactions and produce a low signal, high background, or spurious bands. Plasmid DNA purified through a cesium chloride gradient is suitable for sequencing if residual salt is removed from the DNA by ethanol precipitation. Commercial plasmid purification kits using anion-exchange resins or silica gel membrane technology are available from Qiagen (Valencia, CA, USA) or Promega Corp. (Madison, WI, USA) These kits are easy to use and provide high-quality DNA.

6.4.2.3

Vectors for Large-insert DNA

Cloning vectors capable of replicating large DNA inserts, such as cosmids (DNA inserts with 35 to 45 kb), P1-derived artificial chromosomes (PAC; DNA inserts from 100 to 150 kb), and bacterial artificial chromosomes (BAC; DNA inserts with up to 300 kb), have been developed for use in genome mapping and large-scale DNA sequencing projects [99–101]. These large-insert clones can be used to construct subclone libraries and are then sequenced by the shotgun approach [96] (Sect. 6.4.1.1).

It is also important to sequence directly on these large DNA clones [102, 103]. Sequence information from the ends of large-insert clones is used in the initial mapping phase of a sequencing project by detecting clones with overlapping sequence. Also, closing gaps and low-quality regions in the "draft" sequence of a large-insert clone can be accomplished more efficiently by sequencing directly off of the cosmid or BAC clone. This process eliminates the need to find the specific subclone sequence or to generate a new subclone library covering the gap.

6.4.2.4

PCR Products

The polymerase chain reaction (PCR) enables a region of DNA located between two distinct priming sites to be amplified [104]. The product of this in vitro nucleic acid amplification is termed the PCR product. If equal amounts of the two primers are used, the PCR product will be a linear double-stranded DNA molecule typically less than 3 kb in size which can serve as template for DNA sequencing [94, 105].

The PCR reaction mix contains significant amounts of reagents such as primers, nucleotides, enzymes, and even unwanted amplified products which must be completely removed from the PCR product before it can be successfully sequenced. Thus, the PCR product should be checked on an agarose gel to verify the presence of a single band of the expected size. The PCR product is the purified using a commercial PCR purification kit (e.g. Promega Corp. Wizard DNA Clean-Up System) or by PEG precipitation [95]. Alternatively, PCR products can be purified by use of agarose gel [93].

6.4.3

Enzymatic Reactions

6.4.3.1

DNA Polymerases

In the original Sanger dideoxy sequencing procedure the Klenow fragment of *E. coli* DNA polymerase I was used for primer extension/termination reactions. The quality of the DNA sequence obtained with the Sanger method was significantly improved by the development of a modified T7 DNA polymerase (Sequenase v2.0, United States Biochemical, Cleveland, OH, USA and Amersham Biosciences, Piscataway, NJ, USA) which has enhanced processivity and a striking uniformity of termination patterns, particularly when manganese ions

are used as a cofactor [106–108]. Both the Klenow fragment and modified T7 DNA polymerase catalyze the synthesis of DNA sequencing fragments in a single pass as the enzyme moves along the template DNA. These enzymes are, however, thermolabile, and thus cannot be used in cycle-sequencing procedures which produce an amplification of signal by repeatedly re-using small amounts of the template DNA [109, 110]. Modified T7 DNA polymerase is effective for sequencing difficult regions with repeats that cause premature "stops" in cycle-sequencing reactions [95].

Cycle-sequencing methods that utilize the thermostable *Thermus aquaticus* (Taq) DNA polymerase have been developed [109–111]. The use of a thermostable DNA polymerase enables repeated rounds of high temperature DNA synthesis involving thermal denaturation of the double-stranded template DNA, primer annealing, and extension/termination of the reaction products. For each cycle, the amount of product DNA will be roughly equivalent to the amount of primed template. A significant benefit of cycle sequencing is, therefore, that only small amounts of DNA template are required, because the number of sequencing reaction products (i.e. "the signal") is linearly amplified during the 20–40 cycles of synthesis. For example, 20–30 ng of a small PCR product or 2–3 µg of a large BAC clone provide sufficient template DNA to complete a cycle-sequencing reaction. Performing the cycle-sequencing reactions at elevated temperatures also minimizes sequencing artifacts due to secondary structure in the template DNA.

Originally, the main disadvantage of cycle sequencing was the poor performance of the native Taq DNA polymerase, which tends to incorporate dideoxynucleotides unevenly compared with deoxynucleotides. As a result, sequencing patterns generated with these enzymes were not uniform (i.e. variable peak heights or band intensities) [112] which reduced the base calling accuracy in automated DNA sequencers (Sect. 6.6). However, genetically modified thermostable polymerases with a high affinity for dideoxynucleotides were introduced to alleviate this problem [113, 114]. These enzymes, Thermo Sequenase from Amersham Biosciences and AmpliTaq FS from Applera/Applied Biosystems (Foster City CA, USA), incorporate dideoxynucleotides at rates similar to deoxynucleotides resulting in uniform peak heights and, therefore, longer, more accurate, sequence-read lengths. The reduced discrimination against dideoxynucleotides that has been engineered into ThermoSequenase and AmpliTaq FS has also resulted in the greater acceptance of fluorescent dye-labeled terminators (Sect. 6.3.2) as substrates in the enzymatic sequencing reaction [113].

6.4.3.2

Labeling Strategy

Automated DNA sequencing uses fluorescent dyes (Sect. 6.3) for detection of electrophoretically resolved DNA fragments. Three methods are used for labeling DNA sequencing reaction products:

1. dye-labeled primer sequencing [9, 10] in which the fluorescent dye is attached to the 5′ end of the oligonucleotide primer;
2. dye-labeled terminator sequencing [12, 115] in which the fluorophores are attached to the dideoxynucleotides or a non-nucleotide terminator [116]; and
3. internal labeling [73, 117, 118] in which a dye-labeled deoxynucleotide is incorporated during the synthesis of a new DNA strand.

Each labeling method has advantages and disadvantages.

Dye-labeled primer sequencing has benefited from the engineered DNA polymerases

which do not discriminate between deoxynucleotides and dideoxynucleotides (Sect. 6.4.3.1). The sequencing electropherograms generated using these enzymes with dye-primers have very even peak heights which makes the base-calling easy and reliable. Signal uniformity also enables heterozygote detection to be based on peak heights and the presence of two bases at the same position [86]. One disadvantage of the dye-primer method is a greater likelihood of increased background level (e.g. spurious bands), because nucleotide chains which terminate prematurely will add to the level of false terminations. Also, the four dye/one lane approach for automated sequencing (Sect. 6.4) requires four separate extension reactions and four dye-labeled primers per template.

The main advantages of dye-terminator sequencing are convenience, because only a single extension reaction is required per template, and the synthesis of a dye-labeled primer is not necessary. In fact, custom unlabeled primers with preferred hybridization sites can be used with dye-terminators and false terminations (i.e. DNA fragments terminated with a deoxynucleotide rather than a dideoxynucleotide) are not observed, because these products are unlabeled. Finally, sequencing with dye-terminators provides a way to read through most compressions. Presumably the large fluorophore at the 3′ end of the DNA fragment modifies or eliminates the in-gel secondary structure that causes compressions [95]. The major disadvantage of dye-terminators is that the pattern of termination varies among DNA polymerases and is less uniform than for dye-labeled primers.

6.4.3.3 The Template–Primer–Polymerase Complex

An important factor in the relative success of a sequencing reaction is the number of template–primer–polymerase complexes formed during the course of a sequencing reaction. The formation of this complex is necessary to produce dye-labeled extension products. A significant number of problems associated with DNA-sequencing reactions can be traced to one or more of these key elements.

For example, the ability of an oligonucleotide primer to bind to the template and interact with the DNA polymerase is a major factor in the overall signal strength of the reaction. Primers should be designed with no inverted repeats or homopolymeric regions, a base composition of approximately 50 % GC, no primer dimer formation, and one or more G or C residues at the 3′ end of primer. These factors affect the stability of the primer–template interaction and thus determine the number of primer–template complexes available to the DNA polymerase under a given set of conditions. For cycle sequencing with thermostable polymerases it is important to design the primer with an annealing temperature of at least 50 °C. Lower annealing temperatures tend to produce higher background and stops in cycle sequencing.

The amount of DNA template used in the dideoxy sequence reaction needs to be within an appropriate range. If the amount of template is too small, few complexes will form and the overall signal level will be too low for automatic base-calling. Additionally, larger amounts of a lower quality template (e.g. salt contaminant carried over from DNA preparation) might inhibit the DNA polymerase resulting in lower signal levels.

The most common factors which limit sequence read length and base-calling accuracy in automated DNA sequencers are impure DNA template, incorrect primer or template concentrations, suboptimum primer selection and annealing, and poor removal of unincorporated dye-labeled dideoxynucleotides.

6.4.3.4

Simultaneous Bi-directional Sequencing

Simultaneous bi-directional sequencing (SBS), also termed "doublex" sequencing [117, 119], is a sequencing method in which both strands of duplex DNA (plasmid or PCR product) are sequenced simultaneously by combining a forward and reverse primer (each labeled with a different fluorescent dye) in the same sequencing reaction. An automated DNA sequencing system with dual lasers, such as the LI-COR Model 4300 (Lincoln, NE, USA) or the European Molecular Biology Laboratory (EMBL, Heidelberg, Germany) two-dye DNA sequencer, can be used for simultaneous detection and analysis of both the forward and reverse sequences of a bi-directional reaction [19, 80].

The benefits of the SBS method are threefold. First, SBS doubles the amount of sequence information from a single sequencing reaction. Second, because a confirming sequence can be generated in the same reaction, it is easier to resolve ambiguities in one strand using the sequence of the complementary strand. Third, time and reagent consumption are halved by combining the forward and reverse sequencing reactions.

6.5 Fluorescence DNA Sequencing Instrumentation

6.5.1

Introduction

In principle, there are only three components of a fluorescence detection system: (1) the excitation energy source; (2) the fluorescent sample; and (3) the fluorescence emission energy detector. In practice, all of these components are sophisticated subsystems whose designs are coordinated to deliver maximum information throughput with optimized signal versus noise discrimination (to achieve high accuracy and data quality). A brief discussion of these components is provided here to give an overview of the factors involved in proper instrument design for DNA sequencing. For a detailed description of general fluorescence-based instrumentation, a comprehensive textbook, for example that by Lakowicz [23], should be consulted. For a review of near-infrared fluorescence instrumentation refer to Middendorf et al. [120].

6.5.1.1

Excitation Energy Sources

Laser-based excitation has usually been used for fluorescence-based DNA sequencing instrumentation, although Millipore introduced a DNA sequencer based on a white light source in 1991 [121] (no longer commercially available). The most common lasers used in today's commercial DNA sequencing instrumentation are the blue/ green argon ion laser (488 nm and 514 nm excitation wavelengths) as in early designs from the mid-1980s [9–15] and far red or near infrared laser semiconductor diodes (650 nm, 680 nm, and 780 nm excitation wavelengths) [122, 123]. The red helium– neon laser (HeNe, 633 nm), the green frequency-doubled solid-state neodymium:yttrium–aluminum–garnet laser (Nd:YAG, 532 nm), and the green second harmonic generation laser (SHG, 532 nm, 473 nm) are also used as excitation sources for two-dimensional fluorescence scanners [124] and a green laser (532 nm) also is used in a commercial capillary DNA sequencer (Tab. 6.3, MegaBACE). It is necessary to combine proper geometric optics and spectral filtering to generate a highly focused excitation source with the proper wavelength necessary for compatibility with the fluorescent sample [13, 15, 120, 122, 125–129]. The sample configuration (e.g. slab gels or capillaries) dictates additional mechanical/optical design criteria.

6.5.1.2

Fluorescence Samples

Dye properties such as absorption wavelength spectrum, molar absorptivity, fluorescence emission wavelength spectrum, fluorescence quantum yield, solubility, stability (e.g. temperature or light), and environmental effects (e.g. pH, quenching, temperature, solvent type) must all be considered when designing a fluorescence system. A particular dye property critical to DNA sequencing performance is its electrophoretic mobility [33, 130–132]. For visible fluorescence (blue or green excitation), fluorophores in the fluorescein and rhodamine families are most commonly used. For far red or near infrared fluorescence the most common dyes are from the polymethine carbocyanine family [28–30, 120, 122, 133, 134]. Energy transfer between acceptor and donor dyes has been successfully implemented as a strategy to manipulate compatibility between fluorophore properties and excitation sources and to provide more even peak heights with greater color separation and therefore improved base calling [34, 35, 37–42, 44, 135]. (A more detailed discussion of fluorescent dye chemistry is given in Sect. 6.6.3.)

6.5.1.3

Fluorescence Detection

Three types of detector have been used in fluorescence DNA sequencing instrumentation:

1. photomultiplier tubes (PMT) [9, 10, 12, 13, 15, 125, 127–129, 136];
2. charge-coupled detectors (CCD) [137–139]; and
3. photodiode detectors (PD), including silicon-avalanche photodiodes (APD) [120, 122, 123].

As for the excitation subsystem, it is necessary to combine proper geometric optics, spectral filtering, and appropriate mechanical design to provide high-sensitivity detection in the proper wavelength range associated with the fluorescent sample.

6.5.1.4

Overview of Fluorescence Instrumentation Related to DNA Sequencing

DNA samples for Sanger-based sequencing purposes [17] (Sects. 6.2 and 6.4) are prepared in such a way that they have three major attributes:

1. the 5′ end of every DNA fragment within a sample begins with the same priming sequence;
2. each DNA fragment is labeled with a fluorescent dye (or dye-pair if energy-transfer is used) either at or near the 5′ end or attached to the 3′ terminal dideoxynucleotide;
3. DNA fragments of different length, but having their 3′ terminus ending in a particular base type (A, C, T, or G) are packaged into the same signal channel (e.g. they have the same type of fluorescence label with one label type for each base type or they are physically isolated from fragments terminated at the other base types such that a geometric channel can be used to distinguish base types).

The process of DNA sequencing performs three functions:

1. maintaining the DNA samples in single-stranded form via a combination of denaturants in the gel and high temperature (45–70 °C);
2. separation of the DNA fragments on the basis of their size with single base sizing resolution; and
3. identification of those fragments via fluorescence optics at a “finish line” location where adequate separation among fragments has occurred.

To accomplish sizing a sieving gel matrix is prepared and loaded either between two parallel glass plates (slab gels) or into a glass capillary or microfluidic channel. Slab-gel matrices are usually cross-linked polyacrylamide (4–6 %) whereas capillary gel matrices are non-crosslinked (for example linear polyacrylamide) [140–143]. The gel thickness for slab-gel sandwiches is 0.1–0.4 mm. The gel diameter for capillary gels ranges from 50–80 μm.

Both slab and capillary gels accomplish sizing (after the sample has been loaded into the gel) via a potential gradient applied across the ends of the gel. The potential gradient drives the negatively charged DNA molecules through the sieving matrix with the length of the DNA molecules determining their relative mobility. Each end of the gel is inserted into a running buffer that also contains an electrode that enables the potential gradient across the gel. This voltage gradient may range from 30–80 V cm^{-1} for a slab gel and from 50–250 V cm^{-1} for a capillary gel.

For slab gels it is important to provide a method for keeping samples separated in geometric lanes. This is accomplished by one of several methods:

1. a "comb" with multiple teeth is inserted at one end of the slab gel sandwich before polymerization of the gel matrix. After polymerization of the gel matrix the comb is removed and leaves open wells within the gel into which the DNA samples are loaded via a pipette tip;
2. subsequent to gel polymerization, a "sharkstooth" comb is inserted into the end of the slab gel; the sample is loaded while the comb remains in place and the teeth of the comb thus separate one sample from another;
3. the sample is loaded into one of several wells permanently fabricated in the top edge of one of the glass plates which then provides geometric isolation among samples [144]; or
4. the sample is first applied to a thin long membrane which is then inserted into an air gap located between the two glass plates at one end of the sandwich (where the gel matrix has been excluded) [145].

For capillary gels or microfluidic channel gels the individual capillaries/channels enable isolation of the samples. To load samples into gel capillaries, the loading end of the capillary is first submerged in a microwell containing the sample and the sample is loaded electrokinetically into the capillary. After a brief period of loading this end of the capillary is then submerged into running buffer. Loading samples into microfluidic channels involves moving samples from a "loading" microfluidic channel to a "separating" microfluidic channel (details are given in Sect. 6.5.4.3).

Detection at the "finish line" is accomplished by exciting the various electrophoresis channels either *en masse* [11, 14, 146–151] or one-by-one, by sequential scanning or use of discrete sources, with one or more laser sources [9, 12, 13, 15, 122, 123, 127–129, 152]. An optical microscope or individual detectors monitor any emitted fluorescence radiation from the sample . The finish line is usually located toward one end of the glass-enclosed gel. For capillary electrophoresis it is necessary to remove the polyimide coating of the capillary at the detection zone. One embodiment of capillary electrophoresis (Tab. 6.3, Model 3700) uses a sheath-flow detection scheme which monitors the sample after it leaves the capillary [146, 147, 153].

6.5.2 Information Throughput

High information throughput is mandatory for addressing the accelerating demand for DNA sequencing. For the purpose of understanding the impact of all the factors affecting throughput, one can model information throughput using the following formula (reproduced, with kind permission from Kluwer Academic Publishers, from Ref. [120], page 22, formula 1):

$$T_i = n \times d \times i/t \quad (1)$$

where T_i is the information throughput, n the number of sample channels, d the information data per sample, i the information independence, and t the time per sample.

6.5.2.1 Sample Channels (n)

Different strategies for increasing the number of signal channels include geometric, spectral, temporal, and intensity discrimination. In almost all approaches to DNA sequencing each base type is assigned to a particular signal channel, thus requiring four signal channels per sample. Although it is theoretically necessary to have only two signal channels per DNA sample, use of redundant channel information reduces errors [128, 154, 155].

The number of geometric channels is related to the number of lanes on a slab gel or the number of capillaries in a capillary-based DNA sequencer. At the time of this review the maximum numbers of geometric channels commercially available are 96 lanes for the slab gel configuration and 384 lanes for capillary gel configuration (Tabs. 6.2 and 6.3), although efforts continue to extend the number of capillaries. Scherer et al. [156] report on a rotary capillary array system designed to analyze over 1000 sequencing separations in parallel.

Spectral discrimination using fluorescent dyes with different wavelength properties is ubiquitous among commercial DNA sequencers and the number of dyes used usually ranges from two to four, although five dyes are sometimes used [157]. There are two commercialized approaches to spectral discrimination:

1. four dyes per single sample based on the early work of Smith et al. [9] and Prober et al. [12]; and
2. one dye for all four bases of each sample, but using different dyes for different samples [120, 123, 158] that are loaded into the same geometric lanes.

The latter approach is based on the multiplex DNA sequencing technique developed by Church and Kieffer-Higgins [159] which had its roots in the genomic sequencing approach of Church and Gilbert [160]. Subsequent to Church's work there were early reports of developing the technique by use of fluorescence [117, 119, 161, 162]. Energy transfer among acceptor and donor dyes is another approach in spectral discrimination design [35, 37, 38, 42].

Temporal discrimination based on fluorescence lifetime has been investigated in the research community [47–57], although DNA sequencers using this type of discrimination are not yet commercially available.

The use of intensity discrimination for DNA sequencing has also been limited to the research community [128, 163–170].

6.5.2.2 Information per Channel (d)

The emphasis on increasing the information per channel has been manifested in efforts to increase base read length [74–83] and the use of confidence values to assess

the quality of each base call [171–174]. Both of these efforts are discussed in detail in Sects. 6.6.2 and 6.6.3. For a first order approximation, the base read length is related to the square root of the separation distance from the loading well to the detection location.

6.5.2.3 Information Independence (I)

This attribute relates to sequence alignment strategies [175] and approaches to reducing systematic base-calling errors that affect sequence alignment. For example, in shotgun DNA sequencing, the depth of sequence coverage affects the amount of "draft" versus "finished" sequence, because of statistical gaps between contiguous alignments (contigs) [176]. Judicious choices of clones by use of tiling [177] and finishing [69] strategies affects the cost/output ratio.

Another example of optimizing information dependence/independence ratios involves primer walking, in which newly synthesized primers based on information from a previous DNA sequencing run are used to extend the read through a clone [178]. Too much overlap between successive "walks" through the clone increases the overall cost per base sequenced.

Reduction of systematic errors in base calling is achieved by incorporating independent biochemical procedures such as choice of polymerase, choice of DNA strand, choice of dye chemistry (e.g. labeled primers or labeled terminators) and choice of signal channel (e.g. four dyes/sample or four geometric lanes/sample) [63].

6.5.2.4 Time per Sample (t)

Efforts to minimize the amount of time to obtain DNA sequence data involve three components: (1) sample preparation, (2) electrophoresis run times, and (3) post-run sample information processing and analysis. Robotics and cycle-sequencing methods have greatly reduced the time (and cost) of sample preparation. The use of high voltage gradients (Sect. 6.5.1.4) in combination with either ultrathin slab gels [126, 179–182] or capillary electrophoresis has significantly reduced run times, but at the expense of read length [183, 184]. Efforts to increase read length for capillary DNA sequencing have also been successful [78, 79, 82]. Extending the read length is significant in reducing the time and cost of achieving highly accurate and large contiguous regions of a DNA sequence.

6.5.3 Instrument Design Issues

Proper design of fluorescence instrumentation for DNA sequencing involves comprehensive analysis of both signal and noise components. Middendorf et al. [120] describe the design of the LI-COR Model 4200 DNA sequencer, the principles of which can be extrapolated to the design of other sequencers also, because the report illustrates the relationship among the three above-mentioned components of a fluorescence detection system (Sects. 6.5.1.2–6.5.1.4). Middendorf et al. [120] also describe the formulaic relationship among fluorescence excitation, dye properties, and fluorescence emission; the formulaic relationship between fluorescence emission and detector signal; and the formulaic components of system noise (shot noise, thermal noise) in the LI-COR Model 4200 DNA sequencer.

For a 0.4 mm thick gel and a 4.5 mm well width, the sensitivity of the LI-COR 4200 DNA sequencer is about 15 amol [185]. For thinner gels (0.2 cm) and narrower wells (2.25 mm), the sensitivity is about 5–10 amol (unpublished results). This compares with

sensitivity of 50–100 amol, 150–200 amol, 250–400 amol, and 250–800 amol for FAM, HEX, TAMRA, and ROX dyes, respectively in slab gel electrophoresis [137].

6.5.4 Forms of Commercial Electrophoresis used for Fluorescence DNA Sequencing

Three commercially available forms of electrophoresis are currently used for DNA sequencing–slab gels, capillary gels, and microfluidic grooved channel gels (the commercial products are compared in Tabs. 6.2 and 6.3). All three forms use on-line detection in a "finish line" format which provides spatial information relating to the geometric channel dimension and temporal information relating to the bands within each channel. This temporal information is significantly different than that derived from spatially scanning a two-dimensional gel after stopping the electrophoresis run.

6.5.4.1 Slab Gels

Commercial slab gel electrophoresis systems (Tab. 6.2) include the Applera/Applied Biosystems Model 377 (discontinued by manufacturer, but still in use); LI-COR Models 4300S and 4300L (replaced discontinued Models 4200S and 4200L, respectively); Amersham Biosciences Models ALFexpress II and SEQ4X4; the Bayer Diagnostics OpenGene DNA Sequencing System (formerly Visible Genetics Models Micro-Gene

Table 6.2 Comparison of the slab gel automated fluorescence DNA sequencers/analyzers currently available commercially. Read length (in bases) is for accuracies ranging from 98–99 %, depending on manufacturer. All product names are trademarked under their respective manufacturers. Data were extracted from the Internet home pages of each manufacturer (where more descriptive detail can be found) and from Boguslavsky [186].

Model	*Company*	*Source*	*Detection*	*# dyes*	*# lanes*	*Gel length (cm)*	*Read length*	*Run time*
377	PE Biosystems	Ar Laser	CCD	4–5	18, 36, 64, 96	36, 48 (WTR)	550, 650, 750	3, 9, 11 h
4300S	LI-COR	Laser diodes	APD	2	32, 48, 64, 96	15, 31 (WTR)	400, 700	3, 6 h
4300L	LI-COR	Laser diodes	APD	2	32, 48, 64, 96	15, 31, 56 (WTR)	400, 700, 1000	3, 6, 10 h
ALFexpress	Amersham PB	HeNe Laser	PD	1	40	NA	NA	NA
SEQ4X4	Amersham PB	Laser diodes	PD	1	16	14	300	40 min
Clipper	Visible Genetics	Laser diodes	PD	2	16	14	400	40 min
Tower	Visible Genetics	Laser diodes	PD	2	16	28	800	4 h
DSQ2000L	Shimadzu	Ar Laser	PMT	1	10 samples	610 mm	1200	NA
DSQ600L	Shimadzu	Ar Laser	PMT	1	10 samples	260 mm	350	3 h
DSQ1000	Shimadzu	Ar Laser	NA	1	10 samples	NA	1000	17–20 h
DSQ500	Shimadzu	Ar Laser	NA	1	10 samples	NA	350	2–3 h
BaseStation	MJ Research	Ar Laser	PMT	4	96	NA	500	2 h
NucleoScan	Nucleotech	Solid State	NA	1	48	32	300	1 h

WTR = well-to-read distance. NA = information not available

Table 6.3 Comparison of the capillary gel automated fluorescence DNA sequencers/analyzers currently available commercially. Read length (in bases) is for accuracies ≥98 % unless marked by an asterisk (*), in which case the accuracy is ≥99 % (quality/PHRED values greater than 20 – Sect. 6.6.4). All product names are trademarked under their respective manufacturers. Data were extracted from the Internet home pages of each manufacturer (where more descriptive detail can be found) and from Boguslavsky [186].

Model	*Company*	*Source*	*Detection*	*# dyes*	*# lanes*	*Gel length (cm)*	*Read length (bases)*	*Run time*	*Maximum bases/day*
CEQ8000	Beckman Coulter	Laser diodes	NA	4	8	NA	700	1.7 h	80,000
CEQ8800	Beckman Coulter	Laser diodes	NA	4	8	NA	700	1.7 h	80,000
310	Applied Biosystems	Ar Laser	CCD	4	1	47, 61	425, 600	0.63, 2.7 h	14,985
3100-Avant	Applied Biosystems	Ar Laser	CCD	4	4	36, 80	500, 950	0.67, 4 h	72,000
3100	Applied Biosystems	Ar Laser	CCD	4	16	36, 80	500, 950	0.67, 4 h	288,000
3700	Applied Biosystems	Ar Laser	CCD	4	96	NA	350, 500	2.3, 4 h	350,000
3730	Applied Biosystems	Ar Laser	CCD	4	48	36, 50	500*, 650*, 800*	0.6, 1, 2 h	960,000*
3730xl	Applied Biosystems	Ar Laser	CCD	4	96	36, 50	500*, 650*, 800*	0.6, 1, 2 h	1,920,000*
MegaBACE 500	Amersham	Ar Laser, Green Laser	PMT	4	48	40 (WTR)	600	2 h	355,000*
MegaBACE 1000	Amersham	Ar Laser, Green Laser	PMT	4	96	40 (WTR)	600	2 h	649,000*
MegaBACE 4000	Amersham	Ar Laser	PMT	4	384	40 (WTR)	600	2 h	2,600,000*
SCE9610	SpectuMedix	Ar Laser	CCD	4–30	24, 48, 96, 192	NA	500	3.4 h	672,000
RISA-384	Shimadzu	Ar Laser	NA	4	384	NA	600, 1000	NA	1,000,000
BioMEMS-768	Network Biosystems	NA	NA	4	768	NA	800	NA	5,000,000

WTR = well-to-read distance. NA = information not available

Clipper and Long-Read Tower); Shimadzu Models DSQ-2000L and DSQ-600L; the MJ Research Model BaseStation; and the NucleoTech Model NucleoScan2000. There is significant variation among these systems with regard to the number of spectral channels (dyes), geometric channels (lanes), information per sample (read length, which depends on gel length), and time per read (run time).

6.5.4.2

Capillary Gels

Commercial capillary gel electrophoresis systems (Tab. 6.3) include Applera/Applied Biosystems Models 310, 3100-Avant, 3100, 3700 (discontinued but still in use), 3730, and 3730xl; Amersham Biosciences Models MegaBACE 500, MegaBACE 1000 and MegaBACE 4000; Beckman Coulter Models CEQ8000 and CEQ8800 (replaced discontinued Model CEQ2000); the SpectruMedix Aurora DNA Sequencers (24, 48, 96, or 192 capillaries; also known as Model SCE9610); and the Shimadzu Biotech RISA-384 [187]. Excellent descriptions of capillary array electrophoresis technology are given by Bashkin et al. [188, 189], Dovichi [153], Marsh et al. [190], Pang et al. [139], Behr et al. [191], and Dolnik [192], and references cited therein. Sample purification is important to achieving high performance in capillary DNA sequencing [193, 194].

6.5.4.3

Micro-Grooved Channel Gel Electrophoresis

Instead of using capillaries for DNA separation (with associated electrokinetic loading from microwell plates), grooved channels can be etched in substrates by use of photolithography technology similar to that employed by the semiconductor industry [143, 195–203]. In the January 2000 issue of Electrophoresis is a paper symposium on miniaturization and includes several reviews of microdevice electrophoresis, including Becker and Gärtner [204], McDonald et al. [205], Dolnik et al. [206], and Carrilho [207].

Loading of the sample into the grooved separation channel is significantly different from that of capillary electrophoresis. A "cross-T" interface (or variations using offsets in the junction) between a sample-loading channel and the separation channel enables a sample plug to be injected into the separation channel without creating the bias toward loading only shorter fragments commonly associated with the electrokinetic loading of capillary electrophoresis. The sample can be loaded into the loading channel using electroosmotic pumping or by electrophoresis using electrodes that connect to the two ends of the loading channel. Hybrid devices combining microfabricated "T" injectors with capillary separations have also been investigated in an attempt to extend read length [208, 209].

There is also potential for extending the lifetime of grooved channels compared with that of capillaries, because of the possibility of using high temperatures in connection with various solvents to refurbish the channels. (Using high temperatures with capillaries would damage the polyimide coating used to strengthen the capillary and reduce breakage on bending). The micro-grooved plate is more conducive for interfacing with low volume, upstream reagent processes that require significantly smaller quantities of reagent, and thus reduce cost.

At the time of this review one commercially available DNA sequencer uses micro-grooved channel plates, the Network Biosystems BioMEMS-768 (available from Shimadzu; Tab. 6.3). Collaboration between Agilent Technologies (Palo Alto, CA, USA) and Caliper Technologies (Mountain View, CA, USA) has resulted in a commercial product (Agilent 2100 Bioanalyzer) that uses this micro-grooved technology for sep-

aration of larger DNA fragments but not for DNA sequencing. Other academic and industrial investigation of this technology for DNA sequencing is in progress. Industrial efforts include reports by Agilent and Caliper [210], CuraGen [211], and PE Biosystems (Foster City, CA, USA) [212, 213].

6.5.5 Non-electrophoresis Methods for Fluorescence DNA Sequencing

Several techniques for sequencing short fragments of DNA without use of electrophoresis have been reported in the literature. The emphasis of several of these lies in the importance of detecting single nucleotide polymorphisms (SNP) for diagnostic applications. The efforts usually involve monitoring the extension or removal of the 3′ base, one base at a time.

Technology based on the removal of bases from the 3′ end has been developed by Brenner and colleagues at Lynx Therapeutics [214–219]. This method involves repeated cycles of ligation and cleavage of labeled probes at the 3′ terminus of target DNA. A similar method has been report by Jones [220].

Several groups have investigated the method of single-base extension for DNA sequencing in which the extended base type is determined, one at a time, by fluorescence. Macevicz [221] uses repeated cycles of ligation whereas others have used reversible terminators of polymerase extension in which the terminators are labeled with a distinct, yet removable, tag for each of the four base types [222, 223]. The use of photocleavable fluorescent nucleotides as reversible terminators whereby the photocleavable linker is attached to nucleotide base has also been incorporated into a parallel DNA-sequencing chip system [224].

Another ligase method involves the hybridization and subsequent ligation of short labeled extension oligonucleotides to a DNA template at a position adjacent to the 3′ (or 5′) end of previously hybridized oligonucleotides [225]. In this sequencing method, the labeled ligation product is formed wherein the position and type of label incorporated into the labeled ligation product provides information concerning the nucleotide residue in the DNA template with which it is base-paired.

A clever version of DNA sequencing by single-base extension (called pyrosequencing), that has been commercialized, involves the use of pyrophosphate detection [226–234]. Pyrosequencing involves measurement of the absolute amount of natural nucleotide incorporation by detecting the amount of pyrophosphate released on incorporation. The process utilizes a four-enzyme mixture, including DNA polymerase. The released pyrophosphate is converted to adenosine triphosphate (ATP) by ATP sulfurylase, which is then sensed by luciferase to generate light. Apyrase is used to remove unreacted nucleotides.

A non-enzyme-based technique that has shown utility for resequencing applications is sequencing by hybridization [235–238]. This technique involves hybridizing a library of short oligonucleotides to a DNA template and mathematically transforming the hybridization pattern into a sequence based on the individual sequences of the oligonucleotides from the library that actually hybridize.

6.5.6 Non-fluorescence Methods for DNA Sequencing

Several non-fluorescence techniques for sequencing DNA have been investigated, including matrix-assisted-laser-desorption/ionization–time-of-flight (MALDI–TOF)

mass spectrometry [239–243]. The use of stable non-radioactive isotopes for labeling and detecting bases has been investigated for DNA sequencing by mass spectrometry [244, 245].

6.6 DNA Sequence Analysis

6.6.1 Introduction

The fundamental objective of data analysis is to determine in an automated fashion the DNA sequence from the fluorescence signals in gel electrophoresis generated by the DNA sequence fragments. The performance metrics of the data analysis software are read length, accuracy, and confidence values of the resulting sequence. This is challenging, because of the variation of quality among multiple electrophoresis samples and runs [246].

One approach to automated sequence analysis that depends on minimum variation from one electrophoresis run to another requires adherence to rigid biological and electrophoresis procedures. This approach enables the implementation of an inflexible model that is relatively intolerant to data variations. If, however, the data characteristics lie outside the pre-defined specifications, the automated analysis performance might be significantly compromised.

Another approach that requires considerably less adherence to such rigid procedures is adaptive automation. The algorithms dynamically adjust to optimally fit the data in order to be highly tolerant to the wide variability in data quality [247, 248]. In addition to evaluating local sequence properties such as amplitude, peak time, peak width, and peak fluorescence spectra, it is important to understand the interdependence of these properties among neighboring peaks [114, 249–251]. The best results from automated sequence analysis (long reads, high accuracy, and robust quality) are obtained using a combination of prudent laboratory quality-control measures and adaptive automation analysis.

During analysis the data generated by the fluorescence signal of the automated DNA sequencer are subjected to multiple processes. With each intelligent data reduction performed by the analysis software the data are further mathematically transformed such that the sequence information can be more readily and more accurately obtained. Major software deliverables of the analysis software include: lane detection and tracking, trace generation, base calling, and quality value generation.

6.6.2 Lane Detection and Tracking

Lane detection and tracking is the process of identifying the lane boundaries of DNA sequencing fragments throughout a complete gel image. This process is done automatically, but most analysis software packages make provision for optional visual verification and editing (retracking). Lane tracking is required for slab gel-based sequencers but is not required for capillary-based sequencers (Sect. 6.5.4.2).

Lane tracking is a critical step in the sequence-determination process, because its performance can directly affect the accuracy of the base calls for an entire sample or even a multi-sample gel run. Lane tracking is challenging when there is a wide range of sequence-image configurations of different quality [252]. Characteristics of these different sequence image configurations include comb sizes and types, sample loading formats, gel sizes, sequence chemistries, and

gel matrices. The lane tracking algorithm should also be able to deal effectively with image distortions such as non-uniform lane widths, lane drift, overlapping signals between lanes, variable background noise and signal intensities, a large signal dynamic range, and gel or image streaks/blobs.

For the lane tracker to effectively handle a wide range of DNA sequence-loading options, a minimum set of *a priori* loading information is specified by the user for a given gel. This information includes comb type (rectangular or shark) and tooth size (well-to-well distance), number of samples, and sample-loading format.

After initial computation of the lane boundaries, the image analysis software can, at the user's discretion, display these computed boundaries graphically for manual verification or retracking. The algorithm might also perform a quality assessment of its initial lane-finding performance. If the measure of the quality of its results is low (because of invalid sample-loading information, sample-loading errors, or a poor quality image) it alerts the user while graphically displaying its suggested lane boundaries.

Under normal slab gel electrophoresis conditions the true lane tracks might contain positional variation throughout an electrophoresis run. Such variation is different from one run to another. The adaptive algorithm responds to these (challenging) variations in a manner analogous to that of a human. Adaptive processing of stochastic images includes dynamic noise filtering, dy-

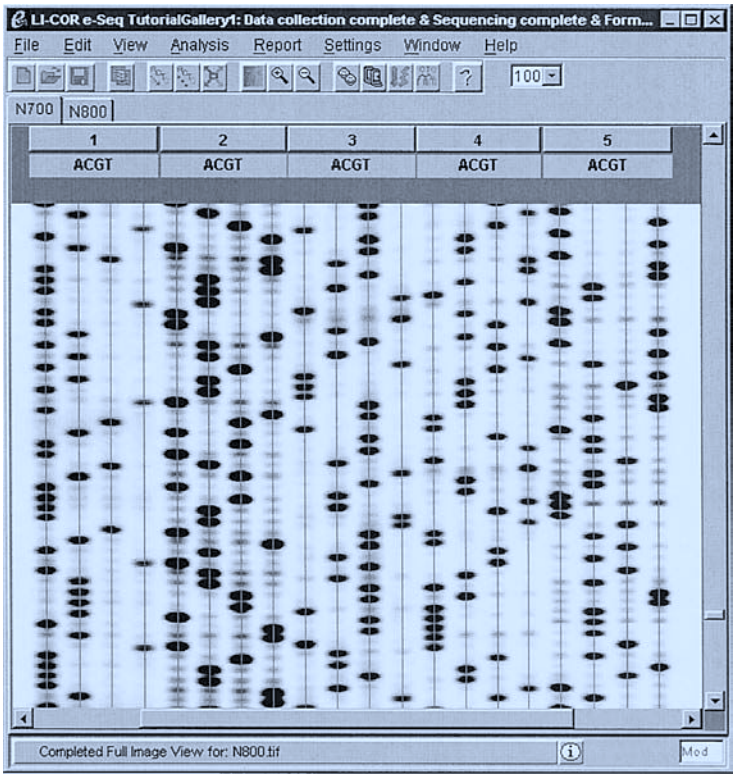


Fig. 6.10 Lane-finder results from the LI-COR Model 4200 DNA Sequencer.

namic background subtraction, and pattern recognition. Artificial-intelligence techniques for band-feature detection and lane-center determination are also used, including neural network techniques [253].

The lane-detection process is initiated by reception of the image file and the *a priori* load configuration. The image data are filtered to remove high-frequency random noise and the background image "surface" is characterized according to its topology. Next, band features are detected by pattern recognition algorithms and the locations of the band centers are determined. By use of iterative optimization the band centers are then partitioned and linked together in associated lane groups. The lane tracker is designed to make adjustments dynamically for lane drift but is restrained in making excessive adjustments. The resulting output of the lane tracker is a multiple set of lane track locations from the beginning to the end of the gel image (Fig. 6.10).

6.6.3 Trace Generation and Base Calling

After creating the lane-track information (for slab gels only; not necessary for capillary gels), trace data are generated for each lane and the sequential base locations, or base calls, are determined. The objective in base calling is to extend the reading of the banding pattern as far as possible with the highest accuracy, in an environment of variable gel or image quality [254].

The accuracy of the base-caller software is mostly affected by the quality of the sequencing reactions [114, 249] and the gel electrophoresis conditions [180, 255–257]. Some of the challenges to accurate base calling include non-uniformity in band-to-band spacing, variable band spreading, non-uniform band mobility, and overlapping (poorly resolved) peaks. Additional potential sources of error include weak or variable signal strengths, ghost bands, variable band morphologies (often because of loading-well distortions and/or salt gradients between the sample and the running buffer), and undesired excess signal artifacts.

For slab gels, the base caller initially transforms data within each of the two-dimensional (2D) image lane tracks into one-dimensional (1D) lane trace profiles. This step is not performed for capillary gels, because the initial data are already formatted as a 1D trace. The 2D image data format associated with slab gels contains additional information, however, that can be utilized to improve the base calling accuracy. For example, on reducing the dimensionality of the signal, the signal-to-noise ratio is enhanced in that summing pixel data across the lane width increases the signal linearly but the noise increases only according to the square root of the number of pixels. In addition, pattern recognition of the full two-dimensional nature of DNA bands assists in analyzing overlapping bands. Care must, however, be taken in the 2D to 1D transformation process such that no signal information is lost or that distortions are not introduced when creating the 1D trace data.

A primary intra-lane image distortion that requires correction before dimension reduction is band tilt (as a result of thermal effects, non-uniform salt concentrations, or well-loading errors) [258]. The computer algorithm dynamically calculates how much band tilt is present in each lane and produces undistorted lane trace profiles. The resulting lane traces are also dynamically corrected relative to background signal levels.

A composite sequence trace is then created by overlaying each of the four associated base profiles (A, T, G, and C) [259]. For the purpose of displaying the composite trace, each base profile is uniquely colored (e.g. red = T, green = A, blue = C, black = G) for

visual identification. An idealized composite trace would consist of evenly spaced, non-overlapping peaks, each corresponding to the labeled fragments that terminate at a particular base in the sequence strand [171]. The non-ideal trace requires further processing, however, for example mobility correction and deconvolution.

Mobility correction is performed on the traces to compensate for mobility inequalities among lanes, because of thermal gradients in the gel [252, 258]. It can also result from errors in gel preparation and loading. To eliminate the mobility shift effect it is necessary for the software algorithm to dynamically examine the band signals throughout the gel. Also, for loading all four base types from one sample into a single lane it is necessary to correct for mobility shifts because of the use of different dye labels for each base type [33, 130–132].

The next processing step is mathematical enhancement of the resolution of the band signals by deconvolution [174, 260, 261]. Resolution has been shown to be the factor limiting increasing read lengths and it degrades as a function of electrophoresis run time [77, 183, 184, 262–275]. Mathematically, resolution is represented by the ratio of band spacing to band spreading. The result of deconvolution is to enhance the data in such a way as to increase this resolution ratio. Deconvolution techniques should not alter band positional information but should reduce band broadening, improve resolution, and reduce signal overlap, as shown in Fig. 6.11 [81].

The final stage in the analysis is identification of the individual bases by intelligent and adaptive processing of the transformed traces. These adaptive techniques characterize band signal intensities, spacing, and spreading. The performance of the LI-COR base caller on genomic DNA data is shown in Fig. 6.12.

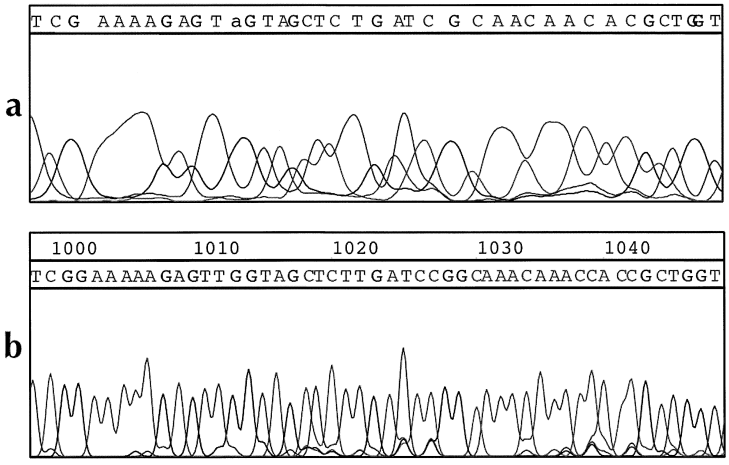


Fig. 6.11 Resolution enhancement by deconvolution: (a) non-deconvoluted trace (b) deconvoluted trace. Both traces represent the section of the pGEM sequence from base 1000 to 1040 as collected on a LI-COR Model 4200 DNA Sequencer using 66 cm gel procedures (56 cm well-to-read distance, 0.2 mm gel thickness, 100 bases h^{-1} run speed). For the region displayed 10 errors were made by the LI-COR base caller (all deletions) when analyzing the upper, non-deconvoluted trace whereas no errors resulted when analyzing the lower, deconvoluted trace.

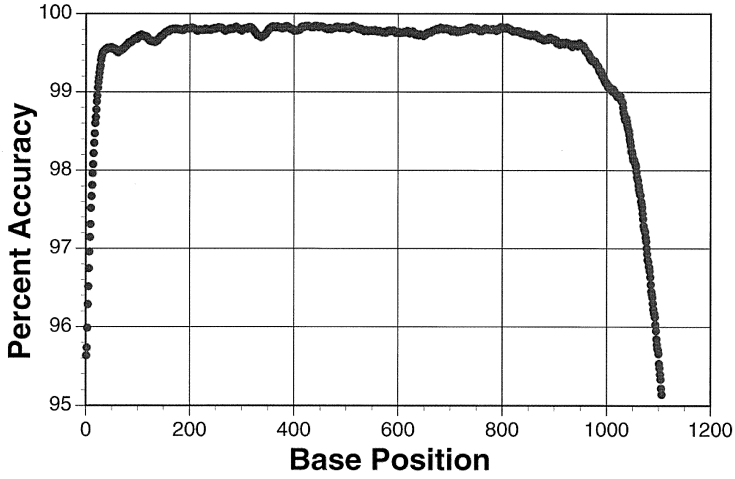


Fig. 6.12 Dependence on base position of the accuracy of the LI-COR Model 4200 DNA Sequencer base caller. The accuracy represents data from 2200 samples of a >2.2 million base-pair project where the average read length was greater than 1000 base pairs and the average accuracy over all reads was 99.65 %. (Data courtesy of Drs P. Brottier, H. Crespeau, and P. Wincker, Genoscope National Sequencing Center, EVRY Cedex, France.)

6.6.4 Quality/Confidence Values

High-throughput sequencing established the need for reliable quality-control measures for the sequence base calls [87, 89, 171, 172, 254, 276] where an estimate of the probability of error is given for each base call. This estimate is a function of selected measures of data quality and enables automation in performing sequence assemblies, quality control, and benchmarking. Assembly software programs such as PHRAP [172], CONSED [277] or CAP3 [278] use these base-specific quality values to improve the accuracy of assembly by weighting the base calls according to their quality values when generating a consensus sequence [173].

To begin a calibration procedure, a pool of measurable sequence data characteristics or properties that correlation strongly with the performance of the base caller is identified. The properties enabling discrimination between correct and incorrect base calls are then selected from the pool. Effective data quality properties include signal-to-noise ratio, resolution, and band-to-band spacing.

When the most effective base-quality properties have been determined, a correlation function (quality predictor) is built. The quality predictor receives as input a set of these quality properties for each base call. The predictor's output is generated by correlating the quality properties with a predicted accuracy (or probability of error) based on past performance under similar conditions.

Proper calibration of the quality predictor requires the formation of a large database containing several million redundant base calls and a known consensus sequence associated with those base calls. (It is impor-

tant that the redundancy is based on diverse sequence conditions.) A set of quality characteristics is then determined for each base call as is identification of its correctness. Using the established database of base calls (and knowledge of both correct and incorrect calls), a correlation between the accuracy and the quality properties is statistically defined. This correlation (or quality predictor) is then stored in the form of a lookup table for further access during subsequent base calling [172]. As an example of the need for the large database, it can be shown that to make a prediction for an accuracy of 1 error out of 10,000 base calls, it is necessary to typify several tens of thousands of base calls with a similar set of quality characteristics.

Results from the LI-COR Model 4200 base caller quality predictor on genomic data are shown in Figs 6.13 and 6.14, in which the quality values are defined as ten times the logarithm (base 10) of the estimated error probability for that base call. Quality values, often referred to as PHRED values [171, 172] of 20 and 30 correspond to predicted error rates of one in one hundred and one in one thousand, respectively.

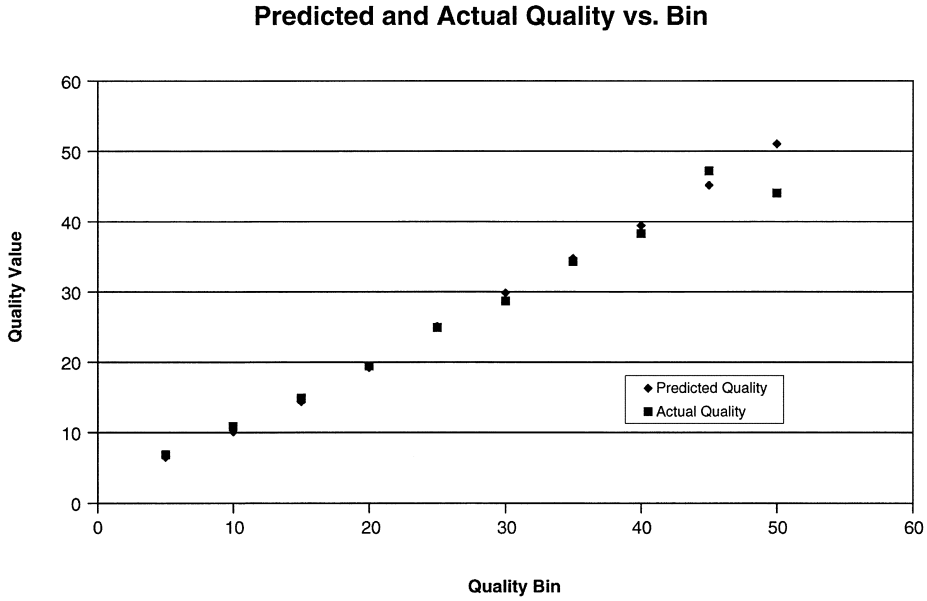


Fig. 6.13 Comparison of predicted and actual quality using the LI-COR Model 4200 DNA Sequencer base caller. Predicted quality values based on 1,518,306 actual base calls were binned into five-score bins. The actual quality values were determined by how many actual errors (based on the consensus sequence) there were for each predictive quality bin, divided by the total number of base calls in that predictive quality bin. (Data courtesy of Drs P. Brottier, H. Crespeau, and P. Wincker, Genoscope National Sequencing Center, EVRY Cedex, France.)

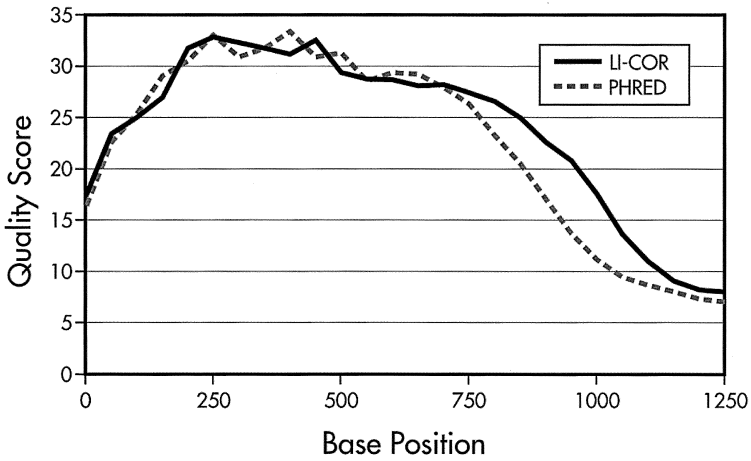


Fig. 6.14 Dependence of quality values on base position for LI-COR and PHRED base callers. These quality values reflect the actual error rates in the aligned parts of sequences in data set TitanGV2 (data courtesy of Drs P. Brottier, H. Crespeau, and P. Wincker, Genoscope National Sequencing Center, EVRY Cedex, France). The graph illustrates that PHRED and LI-COR base callers perform substantially the same for the first 700 bases. After 700 bases, the LI-COR base caller makes fewer errors than PHRED, resulting in a higher quality score.

6.7 DNA Sequencing Approaches to Achieving the $1000 Genome

6.7.1 Introduction

A major change in the technology platform used for DNA sequencing must be implemented to significantly enhance throughput while reducing costs. Examples of current levels of throughput for a single capillary instrument based on Sanger technology include the Applera/Applied Biosystems Model 3730xl at approximately 1.9 megabases per day (equivalent to 22 bases s^{-1}; from http:// docs.appliedbiosystems.com/ pebiodocs/ 00113233.pdf) and the Amersham Biosciences Model MegaBACE 4000 at approximately 2.6 raw megabases per day (equivalent to 30 bases s^{-1}; from http:// www1.amershambiosciences.com/aptrix/ upp00919.nsf/(FileDownload)?OpenAgent&docid = 9B930E7451F83B40C1256C2B00085189&file = 63004308RevAB.pdf). Microfluidic DNA sequencing instrumentation based on Sanger technology, for example the BioMEMS-768 to be commercialized by Network Biosystems, promises throughputs of approximately five raw megabases per day (equivalent to 58 bases s^{-1}; from www.genomems.com/sequencer.asp). A summary of daily production from currently available high-throughput sequencing instrumentation is given in Tab. 6.3. Current costs per mammalian genome, including sample preparation, DNA sequence acquisition, and sequence assembly range from $10M to $50M (http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-002.html), or approximately 0.5–2 cents per high-accuracy "finished" base.

The ability to sequence a human genome accurately in one day would require at least a throughput of 175,000 bases s^{-1} (assuming a minimum of fivefold redundancy to achieve 99.99 % accuracy). This is more than a 3000-fold improvement in throughput compared with current approaches. A $1000 5× genome would necessitate cost reduction of at least four orders of magnitude. Certainly physical limitations such as those considered by the semiconductor industry as it increases the number of transistors per chip [279, 280] need to be taken into consideration as resources are appropriated for the extremely challenging endeavor of increasing DNA sequencing throughput by several orders of magnitude. These assessments include the trade-off between accuracy and information rates as imposed by thermodynamics and modeled using information (communication) theory [281, 282] and the relationship among Heisenberg's Uncertainty Principle, Abbe's Diffraction Limit, and thermal diffusion [283]. Economic issues associated with all components of the DNA sequencing business infrastructure, including shifts in reagent and instrumentation commercialization and contract sequencing factories, also need to be addressed, given the challenge to reduce costs to less than 0.00001cents per raw base in a manner profitable to all enterprises in the sequencing value chain. The practical Biblical advice given by Jesus in Luke 14:28ff with regard to first assessing the costs – "*For which one of you, when he wants to build a tower, does not first sit down and calculate the cost to see if he has enough to complete it? Otherwise, when he has laid a foundation and is not able to finish, all who observe it begin to ridicule him, saying, This man began to build and was not able to finish.*" – becomes as relevant as Pierce's perspective that "*from the point of view of information theory, the most interesting relation between physics and information theory lies in the evaluation of the unavoidable limitations imposed by the laws of physics on our ability to communicate*" (Ref. [281], page 198) and that "*communication theory can be valuable in telling us what can't be accomplished as well as in suggesting what can be*" (ibid, page 179).

Single-molecule detection (SMD) technologies [284–286] might provide a new level of performance if the potential to significantly simplify sample preparation, reduce reagent consumption, and miniaturize the sequencing engine is successfully fulfilled. The sources of error at the single-molecule level are considerably different from those of the traditional Sanger approach, because the detection of every molecule is important. For example, photobleaching, fluorophore blinking [287], and "dead-on-arrival" fluorophores become relevant considerations when using fluorescence SMD. Diffusion can also be a major source of error, regardless of whether or not the detection mechanism involves fluorescence. If the errors associated with SMD sequencing are stochastic, the amount of over-sampling required to compensate for such errors might hopefully be reduced to a level that nullifies any requirements of potential ergodic equivalence between the individual sampling approach of SMD sequencing and the ensemble sampling approach of Sanger sequencing. In other words, over-sampling in Sanger sequencing results from both the use of multiple sequence runs and the multiple molecules within an ensemble (band) that comprises a single Sanger-based data point. SMD sequencing, by its very nature, can only use multiple sequence runs to provide over-sampling for error reduction. If the current level of over-sampling associated with the multiple molecules within an ensemble comprising a single Sanger sequencing data point is necessary to generate the desired level of accuracy, then a

smaller amount of molecules in that ensemble would not suffice to represent the true statistical behavior and the accuracy would be degraded. If this were so then, by definition, ergodic equivalence would require the same amount of sampling between the two approaches (Sanger or SMD) to achieve the same level of accuracy, thus negating the main advantage promised by SMD sequencing because of the excessive need for multiple sequence runs. The hope is that the over-sampling currently associated with Sanger sequencing is "overkill" and, therefore, that SMD will impart more information than that provided by the ensemble average [288].

Four biochemistry strategies have been reported in the scientific literature for DNA sequencing using SMD: (1) DNA degradation, (2) DNA synthesis, (3) DNA hybridization, and (4) nanopore filtering. The first three strategies usually involve fluorescence detection, whereas the nanopore-filtering strategy is predicated on the detection of natural, unlabeled bases. The first approaches to DNA sequencing using SMD were based on the DNA degradation strategy whereby individual bases were removed from a DNA fragment and their base type was identified one-by-one. Efforts using this approach continue to be reported (Sect. 6.7.2). SMD methods using the DNA synthesis strategy usually involve using a polymerase to add one nucleotide at a time to a DNA fragment hybridized to a template strand, and then identifying the incorporated base one-by-one. Several variants of this approach have been reported (Sect. 6.7.3). Hybridization methods utilize a library of oligonucleotide probes spatially positioned at unique locations on a target DNA fragment so that their positions are assayed at the single-molecule level (Sect. 6.7.4). Nanopore filtering involves identifying one-by-one each base of a DNA fragment as that fragment passes through a nanopore (Sect. 6.7.5). Additional information (not included in the references to this chapter) about these methods can be found in the patent literature and the trade magazine literature by using internet search engines. Corporate websites describing technology not yet appearing in the refereed literature are identified in the text where appropriate.

6.7.2 DNA Degradation Strategy

Early attempts at rapid sequencing of 40 kilobase or larger fragments of DNA at a rate of 100 to 1000 bases per second have been explored by Los Alamos National Laboratory (LLNL) [289–291]. Their approach involved the concept of fluorescently labeling every base of a newly synthesized DNA strand using four types of dye for each of the four base types and then attaching the DNA strand to a solid support which is moved into a flowing sample stream, cleaving the labeled 3′ bases one by one with an exonuclease, and finally detecting the cleaved base by using fluorescence detection. More recent progress in demonstrating the feasibility of the LLNL concept was reported in 2003; in this work one of the four base types was labeled and detected after exonuclease cleavage [292]. The seminal effort of the LLNL group set the foundation for investigating single-molecule detection (SMD) as a technology platform for appreciably increasing DNA-sequencing throughput while substantially reducing the cost.

Related research efforts coordinated by the German Ministry of Education and Research (BMBF) and the Swedish National Board for Industrial and Technical Development (NUTEK) were initially reported in 1997 [293] with a comprehensive follow-up collectively reported in 2001 in a special issue of the Journal of Biotechnology (Vol-

ume 86, Issue 3) and included an editorial and nine articles describing aspects of the sample preparation, biochemistry, microfabrication, and detection of this multi-laboratory endeavor [294–303]. Brakmann and Löbermann describe the step of labeling of the target DNA [304] and the step of exonucleolytic degradation of the target DNA [305] in their investigations of SMD sequencing via degradation. A review of the two enzymatic tasks (polymerase and exonuclease activity) has also been written by Brakmann [306]. Further investigations into the synthesis of fluorescently labeled nucleotides and the polymerase-mediated fluorescent replacement of bases in a DNA molecule have been conducted by Gnothis SA, a Swiss commercial enterprise [307, 308].

Ulmer [309] developed a method for DNA sequencing involving the use of an exonuclease to cleave from a single DNA strand the next available single nucleotide on the strand and then detecting the cleaved nucleotide by transporting it into a fluorescence-enhancing matrix such that the natural fluorescence spectrum of the four nucleotides is exploited to determine the base sequence. To detect the natural fluorescence it was necessary to cool the nucleotides to cryogenic temperatures.

The use of fluorescence polarization to discriminate among the four base types cleaved by an exonuclease instead of the spectral discrimination associated with four different dyes has been investigated by Yan and Myrick [310] as another approach to SMD sequencing via DNA degradation.

6.7.3 DNA Synthesis Strategy

Several groups have investigated DNA sequencing via DNA synthesis with a variety of biochemical and optical methods to enhance signal-to-background ratios, including reagent cycling, polymerase residence time discrimination, electronic charge discrimination, fluorescence energy transfer, quenching, photobleaching, evanescent wave excitation, and zero-mode cavity waveguides and a variety of fluorescence labeling motifs, including labeling of the nucleotide base, sugar, or phosphate and labeling of the polymerase.

One of the more common methods of the DNA synthesis strategy is to use reversible terminators of DNA synthesis in combination with a step-wise sequencing process involving repeated cycles of interrogation (see, for example, www.solexa.com and www.genovoxx.de). This method involves incorporation into a growing DNA strand of a single fluorescently-labeled nucleotide which then blocks the polymerase from adding another base, either because of a terminating moiety at the 3′ position of the newly incorporated nucleotide (e.g. Solexa) or because of steric hindrance between the polymerase and the fluorescent label (e.g. Genovoxx). The incorporated nucleotide is identified on the basis of its fluorescence spectrum and then the fluorescent label or 3′ block is removed. If the fluorescent dye remains on the growing strand it must be photo-deactivated before the next cycle.

A related technology amplifies an individual DNA molecule into one of several sequence-specific colonies, with each colony localized in a unique spatial position within a thin acrylamide gel attached to a microscope slide [311]. Each stepwise cycle of incorporation of a base-labeled nucleotide includes fluorescence removal via either a reducing agent acting on a disulfide linkage between the fluorophore and the nucleotide base or the use of near-UV light to break a photocleaveable linker (see also Refs. [224] and [312]). Even though the fluorescent label is conjugated to the base, it still blocks

the polymerase from adding more than one nucleotide for a sufficient amount of cycle time to assay which base type was incorporated. In a similar fashion researchers at 454 Corporation (www.454.com) start with an individual DNA molecule which is then PCR-amplified within one of 300,000 microwells containing as little as 39.5 picoliter (pL) and arranged in a honeycomb array [313]. Sequencing is performed within each well using the pyrosequencing technology (Sect. 6.5.5) with read lengths up to 100 bases [314].

In 2003, the Quake lab at Caltech published results reporting the first actual determination of sequence fingerprints up to five base pairs in length of individual DNA strands with single-base resolution [315]. DNA templates were anchored on to a quartz slide then combined with a primer and polymerase, with sequencing performed by means of cyclic, stepwise introduction of base-labeled nucleotides combining evanescent excitation, fluorescence energy transfer, and photobleaching. Helicos BioSciences (www.helicosbio.com) has started to commercialize the Quake technology.

Approaches not based on the stepwise cycling of reversible terminators require detection approaches that discriminate between excess unincorporated labeled nucleotides and the individual nucleotide that has just been incorporated into the growing DNA strand by the polymerase. An effort at Cornell that used zero-mode cavity waveguides to reduce the detection volume and fluorescence correlation spectroscopy (FCS) for temporal discrimination demonstrated the observation of single molecule DNA polymerase activity at high (micromolar) concentrations of labeled nucleotides with successful background discrimination [316]. This work is being commercialized by Nanofluidics (www.nanofluidics.com). Visigen Biotechnologies (www.visigenbio.com) is investigating the use of fluorescence energy transfer between a labeled engineered polymerase and the incoming labeled nucleotide for discriminating against bulk labeled nucleotides. Mobious Genomics (www.mobious.com) is investigating the use of surface plasmon resonance to detect polymerase conformation changes as a function of the particular type of nucleotide incorporated. This approach does not involve fluorescence labeling.

6.7.4 DNA Hybridization Strategy

Approaches that use the DNA hybridization strategy either move single DNA fragments that contain multiple labeled probes past a detector (US Genomics) or stretch out DNA on a substrate and use two-dimensional detection of either labeled probes hybridized to the DNA or immobilized fragments of the DNA after they have been subjected to restriction endonucleases (OpGen). The US Genomics (www.usgenomics.com) technology involves the nanofabrication of either narrow slits [317, 318] or a series of posts [319] to geometrically untangle a linear string of mostly single-stranded DNA to enable detection of labeled probes hybridized to the DNA. The temporal distribution of the fluorescence signal generated as the DNA fragment moves past the detector is translated to spatial information. The technology is currently used for detecting single-base mutations but might have potential for DNA sequencing. Related technology for periodically transporting DNA molecules through a glass nanopipette (fabricated to have an inner radius of ~50 nm by use of a laser-based pipette puller) used single molecule fluorescence detection to determine the number of molecules delivered per voltage pulse [320].

The approach of OpGen (www.opgen.com) in which DNA is stretched into an untangled form for spatially-resolved detection of restriction fragments lends itself well to DNA mapping [321–325] with the hope that the analysis of "DNA molecules bound to chemically modified surfaces" ... "can be used as a viable platform to ascertain the sequences of short nucleotide strings and ultimately, complete genomes" (taken from abstract of a University Wisconsin–Madison analytical seminar, September 24, 2003).

6.7.5 Nanopore Filtering Strategy

In 1996 seminal work on the use of lipid bilayer membrane channel as nanopores through which single-stranded DNA molecules could be translocated was reported [326]. This has led to a large number of efforts to assess the feasibility of using nanopores for single-molecule DNA sequencing, many of which have received federal funding (see, for example, http://crisp.cit.nih.gov) but so far with no published results. Critical to the success of this strategy is the ability to resolve the differences between individual bases within the channel [327, 328]. Other challenges include clogging at the opening of the pore as a result of DNA secondary structure such as hairpin loops, bidirectional movement of the DNA within the pore owing to thermal diffusion, and the formation of robust nanopores, including the possibility of solid-state nanopores, rather than the α-hemolysin channels most commonly studied [329–336]. The challenges of using nanopores for DNA sequencing have been reviewed by Slater et al. (Ref. [337], p. 3883). Detection approaches within the nanopore have generally been limited to ionic current measurements.

References

1 Brownlee, G.G., Sanger, F. and Barrell, B.G. (1967), Nucleotide sequence of 5S-ribosomal RNA from Escherichia coli, Nature 215, pp. 735–736.

2 Fellner, P. and Sanger, F. (1968), Sequence analysis of specific areas of 16S and 23 ribosomal RNAs, Nature 219, pp. 236–238.

3 Adams, J.M., Jeppesen, P.G., Sanger, F. and Barrell, B.G. (1969), Nucleotide sequence from the coat protein cistron of R17 bacteriophage RNA, Nature 223, pp. 1009–1014.

4 Sanger, F., Donelson, J.E., Coulson, A.R., Kossel, H. and Fischer, D. (1973), Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage f1 DNA, Proc. Natl. Acad. Sci. USA 70, pp. 1209–1213.

5 Sanger, F., Donelson, J.E., Coulson, A.R., Kossel, H. and Fischer, D. (1974), Determination of a nucleotide sequence in bacteriophage f1 DNA by primed synthesis with DNA polymerase, J. Mol. Biol. 90, pp 315–333.

6 Gilbert, W. and Maxam A. (1973), The nucleotide sequence of the lac operator, Proc. Natl. Acad. Sci. USA 70, pp. 3581–3584.

7 Sanger, F. and Coulson, A.R. (1975), A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, J. Mol. Biol. 94, pp. 441–448.

8 Maxam, A. M. and Gilbert, W. (1977), A new method for sequencing DNA, Proc. Natl. Acad. Sci. U.S.A. 74, pp. 560–564.

9 Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, B.H. and Hood, L.E. (1986), Fluorescence detection in automated DNA sequence analysis, Nature 321, pp. 674–679.

10 Ansorge, W., Sproat, B.S., Stegemann, J. and Schwager, C. (1986), A non-radioactive automated method for DNA sequence determination, J. Biochem. Biophs. Methods 13, pp. 315–323.

11 Ansorge, W., Sproat, B., Stegemann, J., Schwager, C. and Zenke, M. (1987), Automated DNA sequencing: Ultrasensitive detection of fluorescent bands during electrophoresis, Nucl. Acids Res. 15, 4593–4602.

12 Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A. and Baumeister, K. (1987), A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides, Science 238, pp. 336–341.

13 Brumbaugh, J.A., Middendorf, L.R., Grone, D.L. and Ruth, J.L. (1988), Continuous, on-line, DNA sequencing using multifluorescently tagged primers, Proc. Natl. Acad. Sci. USA 85, pp. 5610–5614.

14 Kambara, H., Nishikawa, T., Katayama, Y. and Yamaguchi, T. (1988), Optimization of parameters in a DNA sequenator using fluorescence detection, Bio/Technology 6, pp. 816–821.

15 Middendorf, L.R., Brumbaugh, J.A., Grone, D.L., Morgan, C.A. and Ruth, J.L. (1988), Large scale DNA sequencing, American Biotechnology Laboratory 6, pp. 14–22.

16 DeLisi, C. (1988), The human genome project, Amer. Sci. 76, pp. 488–493.

17 Sanger, F., Nicklen, S. and Coulson, A.R. (1977), DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. USA 74, pp 5463–5467.

18 Alphey, L. (1997), DNA Sequencing. New York: Springer.

19 Ansorge, W., Voss, H. and Zimmermann, J. (1997), DNA Sequencing Strategies. New York: John Wiley and Sons.

20 Smith, L.M. (1989), DNA sequence analysis: past, present, and future, Intl. Biotech. Lab. 7, pp. 8–19.

21 Middendorf, L.R., Bruce, J.C., Bruce, R.C., Eckles, R.D., Roemer, S.C. and Sloniker, G.D. (1993), A versatile infrared laser scanner/ electrophoresis apparatus, Proc. SPIE 1885, pp. 423–434.

22 Kessler, C. (Ed) (1992), Nonradioactive Labeling and Detection of Biomolecules, Berlin: Springer.

23 Lakowicz, J. (1999), Instrumentation for Fluorescence Spectroscopy, in: Principles of Fluorescence Spectroscopy (Lakowicz, J., Ed.), 2nd Ed., pp. 25–61. New York: Kluwer Academic/Plenum Publishers.

24 Stokes, G.G. (1852), On the change of refrangibility of light, Phil. Trans. R. Soc. London 142, pp. 463–562.

25 Strekowski, L., Lipowska, M. and Patonay, G. (1992), Facile derivatizations of heptamethine cyanine dyes, Synth. Comm. 22, 2593–2598.

26 Narayanan, N., Little, G., Raghavachari, R. and Patonay, G. (1995), New near infrared dyes for applications in bioanalytical methods, Proc. SPIE 2388, pp. 6–15.

27 Shealy, D. B., Lipowska, M., Lipowski, J., Narayanan, N, Sutter, S., Strekowski, L. and Patonay, G. (1995), Synthesis, chromatographic separation, and characterization of near-infrared labeled DNA oligomers for use in DNA sequencing, Anal. Chem. 67, pp. 247–251.

28 Narayanan, N., Little, G., Raghavachari, R., Gibson, J., Lugade, A., Prescott, C., Reiman, K., Roemer, S., Steffens, D., Sutter, S. and Draney, D. (1998), New NIR dyes: Synthesis, spectral properties and applications in DNA analyses, in: Near-Infrared Dyes for High Technology Applications (S. Daehne et al., Eds.), pp. 141–158. The Netherlands: Kluwer Academic Publishers.

29 Matsuoka, M., Editor (1990), Infrared Absorbing Dyes, pp. 19–33, New York: Plenum Press.

30 Mujumdar, R.B., Ernst, L.A., Mujumdar, S.R. and Waggoner, A.L. (1989), Cyanine dye labeling reagents containing isothiocyanate groups, Cytometry 10, pp. 11–19.

31 Mujumdar, R.B., Ernst, L.A., Mujumdar, S.R., Lewis, C.J. and Waggoner, A.L. (1993), Cyanine dye labeling reagents: sulfoindocyanine succinimidyl esters, Bioconjugate Chem. 4, pp. 105–111.

32 Zhu, Z., Chao, J., Yu, H. and Waggoner, A.S. (1994), Directly labeled DNA probes using fluorescent nucleotides with different length linkers, Nucl. Acids Res. 22, pp. 3418–3422.

33 Tu, O., Knott, T., Marsh, M., Bechtol, K., Harris, D., Barker, D. and Bashkin, J. (1998), The influence of fluorescent dye structure on the electrophoretic mobility of end-labeled DNA, Nucl. Acids Res. 26, pp. 2797–2802.

34 Rosenblum, B.B., Lee, L.G., Spurgeon, S.L., Khan, S.H., Menchen, S.M., Heiner, C.R. and Chen, S.M. (1997), New dye-labeled terminators for improved DNA sequencing patterns, Nucl. Acids Res. 25, pp. 4500–4504.

35 Lee, L.G., Spurgeon, S.L., Heiner, C.R., Bensen, S.C., Rosenblum, B.B., Menchen, S.M., Graham, R.J., Constantinescu, A., Upadhya, K.G. and Cassel, J.M. (1997), New energy transfer dyes for DNA sequencing, Nucl. Acids Res. 25, pp. 2816–2822.

36 Förster, Th. (1948), Intermolecular energy migration and fluorescence, Ann. Phys. (Leipzig) 2, pp. 55–75.

37 Ju, J., Ruan, C., Fuller, C. W., Glazer, A. N., and Mathies, R. A., (1995), Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis, , Proc. Natl. Acad. Sci. U.S.A. 92, pp 4347–4351.

38 Ju, J., Kheterpal, I., Scherer, J. R., Ruan, C., Fuller, C. W., Glazer, A. N. and Mathies, R. A., (1995), Design and synthesis of fluorescence energy transfer dye-labeled primers and their applications for DNA sequencing and analysis, Anal. Biochem. 231, pp 131–140.

39 Ju, J., Glazer, A. N., and Mathies, R. A., (1996), Energy transfer primers: A new fluorescence labeling paradigm for DNA sequencing and analysis, Nat. Med. 2, pp 246–249.

40 Hung, S.C., Ju, J., Mathies, R.A. and Glazer, A.N. (1996), Cyanine dyes with high absorption cross section as donor chromophores in energy transfer primers, Anal. Biochem. 243, pp. 15–27.

41 Hung, S.C., Ju, J., Mathies, R.A. and Glazer, A.N. (1996), Energy transfer primers with 5- or 6-carboxyrhodamine-6G as acceptor chromophores, Anal. Biochem. 238, pp. 165–170.

42 Metzker, M.L., Lu, J. and Gibbs, R.A. (1996), Electrophoretically uniform fluorescent dyes for automated DNA sequencing, Science 271, 1420–1422.

43 Hung, S.C., Mathies, R.A. and Glazer, A.N. (1997), Optimization of spectroscopic and electrophoretic properties of energy transfer primers, Anal. Biochem. 252, pp. 78–88.

44 Hung, S.C., Mathies, R.A. and Glazer, A.N. (1998), Comparison of fluorescence energy transfer primers with different donor–acceptor dye combinations, Anal. Biochem. 255, pp. 32–38.

45 Köllner, M. and Wolfrum, J. (1992), How many photons are necessary for fluorescence-lifetime measurements?, Chem. Phys. Letters 200, pp. 199–204.

46 Köllner, M. (1993), How to find the sensitivity limit for DNA sequencing based on laser-induced fluorescence, Appl. Optics 32, pp. 806–820.

47 Chang, K., and Force, R. K. (1993), Time-resolved laser induced fluorescence study on dyes used in DNA sequencing, Appl. Spectrosc. 47, pp. 24–29.

48 Han, K.-T., Sauer, M., Schulz, A., Seeger, S. and Wolfrum, J. (1993), Time-resolved fluorescence studies of labelled nucleosides, Ber. Bunsenges. Phys. Chem. 97, pp. 1728–1730.

49 Sauer, M., Han, K.-T., Ebert, V., Muller, R., Schulz, A., Seeger, S. and Wolfrum, J. (1994), Design of multiplex dyes for the detection of different biomolecules, Proc. SPIE 2137, pp. 762–774.

50 Legendre, B. L., Williams, D. C., Soper, S. A., Erdmann, R., Ortmann, U., and Enderlein, J. (1996), An all solid-state near-infrared time-correlated single photon counting instrument for dynamic life-time measurements in DNA sequencing applications, Rev. Sci. Inst. 67, pp. 3984–3989.

51 Soper, S.A., Davidson, Y.Y., Flanagan, J.H., Legendre, Jr., B.L., Owens, C., Williams, D.C. and Hammers, R.P. (1996), Micro-DNA sequence analysis using capillary electrophoresis and near-IR fluorescence detection, Proc. SPIE 2680, pp. 235–246.

52 Müller, R., Herten, D., Lieberwirth, U., Neumann, M., Sauer, M., Schulz, A., Siebert, S., Drexhage, K.H. and Wolfrum, J. (1997), Time-resolved DNA identification in capillary gel electrophoresis with semiconductor lasers, Proc. SPIE 2980, pp. 116–126.

53 Nunnally, B.K., He, H., Li, L.-C., Tucker, S.A. and McGown, L.B. (1997), Characterization of visible dyes for four-decay fluorescence detection in DNA sequencing, Anal. Chem. 69, pp. 2392–2397.

54 Flanagan, J.H. Jr., Owens, C.V., Romero, S.E., Waddell, E., Kahn, S.H., Hammer, R.P. and Soper, S.A. (1998), Near-infrared heavy-atom-modified fluorescent dyes for base-calling in DNA-sequencing applications using temporal discrimination, Anal. Chem. 70, pp. 2676–2684.

55 Li, L.-C. and McGown, L. B. (1996), On-the-fly frequency domain fluorescence lifetime detection in capillary electrophoresis, Anal. Chem. 68, pp. 2737–2743.

56 Li, L.-C., He, H., Nunnally, B. K., and McGown, L. B. (1997), On-the-fly fluorescence lifetime detection of labeled DNA primers, J. Chromatogr. 695, pp. 85–92.

57 Lassiter, S.J., Stryjewski, W., Legendre, B.L. Jr., Erdmann, R., Wahl, M., Wurm, J., Peterson, R., Middendorf, L. and Soper, S.A. (2000), Time-resolved fluorescence imaging of slab gels for lifetime base-calling in DNA sequencing applications, Anal. Chem 72, pp. 5373–5382.

58 Prummer, M., Hübner, C.G., Sick, B., Hecht, B., Renn, A. and Wild, U.P. (2000), Single-molecule identification by spectrally and time-resolved fluorescence detection, Anal. Chem. 72, pp. 443–447.

59 Hawkins, T.L., Du, Z., Halloran, N.D. and Wilson, R.K. (1992), Fluorescence chemistries for automated primer-directed DNA sequencing, Electrophoresis 13, pp. 552–559.

60 Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. and Waterston, R. (1992), The C. elegans genome sequencing project: a beginning, Nature 356, pp. 37–41.

61 Fleischmann, R.D., Adams, M.D., White O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J.D., Scott, J., Shirley, R., Liu, L., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O. and Venter, J.C. (1995), Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, Science 269, pp. 496–512.

62 Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O. and Hunkapiller, M. (1998), Shotgun sequencing of the human genome, Science 280, pp. 1540–1542.

63 Kukanskis, K.A., Siddiquee, Z., Shohet, R.V. and Garner, H.R. (2000), Mix of sequencing technologies for sequence closure: an example, BioTechniques 28, pp. 630–632, 634.

64 She, Q., et al. (2001), The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2, Proc. Natl. Acad. Sci. U.S.A. 98, pp. 7835–7840.

65 International Human Genome Sequencing Consortium (2001), Initial sequencing and analysis of the human genome, Nature 409, pp. 860–921.

66 Venter, J.C., et al. (2001), The sequence of the human genome, Science 291, pp. 1304–1351.

67 Martin-Gallardo, A., Lamerdin, J., and Carrano, A. (1994), Shotgun sequencing, in: Automated DNA Sequencing and Analysis (M.D. Adams et al., Eds.), pp. 37–41. London: Academic Press.

68 Hunkapiller, T., Kaiser, R.J., Koop, B.F. and Hood, L. (1991), Large-scale and automated DNA sequence determination, Science 254, pp. 59–67.

69 Roach, J.C., Siegel, A.F., van den Engh, G., Trask, B. and Hood, L. (1999), Gaps in the Human Genome Project, Nature 401, pp.843–845.

70 Istrail, S., et al. (2004), Whole-genome shotgun assembly and comparison of human genome assemblies, Proc. Nat. Acad. Sci. U.S.A. 101, pp. 1916–1921.

71 Liu, L.I. and Fleischmann, R.D. (1994), Construction of exonuclease III generated nested deletion sets for rapid DNA sequencing, in: Automated DNA Sequencing and Analysis (M.D. Adams et al., Eds.), pp.65–70. London: Academic Press.

72 Martin, C.H., Mayeda, C.A., Davis, C.A., Strathmann, M.P. and Palazzolo, M.J. (1994), Transposon-facilitated sequencing: an effective set of procedures to sequence DNA fragments smaller than 4 kb, in: Automated DNA Sequencing and Analysis (M.D. Adams et al., Eds.), pp.60–64. London: Academic Press.

73 Voss, H., Wirkner, U., Schwager, C., Zimmermann, J., Stegemann, J., Hewitt, N.A. and Ansorge, W. (1993), Automated DNA sequencing system resolving 1000 bases with fluorescein-15-dATP as internal label, Methods Mol. Cell. Biol. 3, pp. 153–155.

74 Nishikawa, T. and Kambara, H. (1992), High resolution-separation of DNA bands by electrophoresis with a long gel in a fluorescence-detection DNA sequencer, Electrophoresis 13, pp. 495–499.

75 Grothues, D., Voss, H., Stegemann, J., Wiemann, S., Sensen, C., Zimmermann, J., Schwager, C., Erfle, H., Rupp, T. and Ansorge, W. (1993), Separation of up to 1000 bases on a modified A.L.F. DNA sequencer, Nucl. Acids Res. 21, pp. 6042–6044.

76 Zimmermann, J., Wiemann, S., Voss, H., Schwager, C. and Ansorge, W. (1994), Improved fluorescent cycle sequencing protocol allows reading nearly 1000 bases, BioTechniques 17, pp. 302–308.

77 Middendorf, L., Gartside, B., Humphrey, P., Roemer, S., Sorensen, D., Steffens, D. and Sutter, S. (1995), Enhanced throughput for infrared automated DNA sequencing, Proc. SPIE 2386, pp. 66–78.

78 Carrilho, E., Ruiz-Martinez, M.C., Berka, J., Smirnov, I., Goetzinger, W., Miller, A.W., Brady, D. and Karger, B.L. (1996), Rapid DNA sequencing of more than 1000 bases per run by capillary electrophoresis using replaceable linear polyacrylamide solutions, Anal. Chem. 19, pp. 3305–3313.

79 Klepárnik, K., Foret, F., Berka, J., Goetzinger, W., Miller, A.W. and Karger, B.L. (1996), The use of elevated column temperature to extend DNA sequencing read lengths in capillary electrophoresis with replaceable polymer matrices, Electrophoresis 17, 1860–1866.

80 Roemer, S.C, Brumbaugh, K. A., Boveia, V. and Gardner, J. (1997), Simultaneous bi-directional cycle sequencing, Microbial and Comparative Genomics 2, pp. 206 (Abstract A-33 only).

81 Roemer, S., Boveia, V., Humphrey, P., Amen, J. and Osterman, H. (1998), Improvements in Long Read Automated Sequencing, Microbial and Comparative Genomics 3, pp. C-67 (Abstract A-59 only).

82 Salas-Solano, O., Carrilho, E., Kotlar, L., Miller, A.W., Goetzinger, W., Sosic, Z. and Karger, B.L. (1998), Routine DNA sequencing of 1000 bases in less than one hour by capillary electrophoresis with replaceable linear polyacrylamide solutions, Anal. Chem. 70, pp. 3996–4003.

83 Zhou, H., Miller A.W., Sosic, Z., Buchholz, B., Barron A.E., Kotler, L. and Karger, B.L. (2000), DNA sequencing up to 1300 bases in

two hours by capillary electrophoresis with mixed replaceable linear polyacrylamide solutions, Anal. Chem. 72, pp. 1045–1052.
84 Larder, B.A., Kohli, A., Kellam, P., Kemp, S.D., Kronick, M. and Henfrey, R.D. (1993), Quantitative detection of HIV-1 drug resistance mutations by automated DNA sequencing, Nature 365, pp. 671–673.
85 Kwok, P.Y., Carlson, C., Yager, T.D., Ankener, W. and Nickerson, D.A. (1994), Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products, Genomics 23, pp. 138–144.
86 Perkin–Elmer/ABI (1995), Comparative PCR sequencing manual, Foster City, CA.
87 Nickerson, D.A., Tobe, V.O. and Taylor, S.L. (1997), PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing, Nucl. Acids Res. 25, pp. 2745–2751.
88 Plaschke, J., Voss, H., Hahn, M., Ansorge, W. and Schackert, H.K. (1998), Doublex sequencing in molecular diagnosis of hereditary diseases, BioTechniques 24, pp. 838–841.
89 Rieder, M.J., Taylor, S.L., Tobe, V.O. and Nickerson, D.A. (1998), Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome, Nucl. Acids Res. 26, pp. 967–973.
90 Crawford D.C., Carlson C.S., Rieder M.J., Carrington D.P., Yi Q., Smith J.D., Eberle M.A., Kruglyak L. and Nickerson D.A. (2004), Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations, Am J Hum Genet. 74, pp. 610–622.
91 Chen, E.Y. and Seeburg, P.H. (1985), Supercoil sequencing: a fast and simple method for sequencing plasmid DNA, DNA 4, pp. 165–170.
92 Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989), Molecular Cloning, Second Edition. Cold Spring Harbor Laboratory Press.
93 Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (1992), Short Protocols in Molecular Biology. New York: John Wiley and Sons.
94 Fanning, S. and Gibbs, R.A. (1997), PCR in genome analysis, in: Genome Analysis, A Laboratory Manual (B. Birren et al., Eds.), pp. 249–299. Cold Spring Harbor Laboratory Press.
95 Wilson, R.K. and Mardis, E.R. (1997), Fluorescence-based DNA sequencing, in: Genome Analysis, A Laboratory Manual (B. Birren et al., Eds.), pp. 301–395. Cold Spring Harbor Laboratory Press.
96 Wilson, R.K. and Mardis, E.R. (1997), Shotgun sequencing, in: Genome Analysis, A Laboratory Manual (B. Birren et al., Eds.), pp. 397–454. Cold Spring Harbor Laboratory Press.
97 Messing, J. (1983), New M13 vectors for cloning, Methods Enzymol. 101, pp. 20–78.
98 Messing, J. and Bankier, A.T. (1989), The use of single-stranded DNA phage in DNA Sequencing, in: Nucleic Acids Sequencing (C.J. Howe and E.S. Ward, Eds.), pp. 1–36. Oxford: IRL Press.
99 Craxton, M. (1993), Cosmid sequencing, in: Methods in Molecular Biology, Vol. 23 DNA Sequencing Protocols (H. Griffin, Ed.), pp. 149–167. Totowa: Humana Press.
100 Ioannou, P.A., Amemiya, C.T., Garnes, J., Kroisel, P.M., Shizuya, H., Chen, C., Batzer, M.A. and de Jong, P.J. (1994), A new bacteriophage P1-derived vector for the propagation of large human DNA fragments, Nature Genetics 6, pp. 84–89.
101 Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, T. and Simon, M. (1992), Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector, Proc. Natl. Acad. Sci. USA 89, pp. 8794–8797.
102 Boysen, C., Simon, M.I. and Hood, L. (1997), Fluorescence-based sequencing directly from bacterial and P1-derived artificial chromosomes, BioTechniques 23, pp. 978–982.
103 Fajas, L., Staels, B. and Auwerx, J. (1997), Cycle sequencing on large DNA templates, BioTechniques 23, pp. 1034–1036.
104 Mullis, K.B. and Faloona, F.A. (1987), Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction, in: Methods in Enzymology (R. Wu, Ed.), pp. 335–350. London: Academic Press.

105 Innis, M., Gelfand, D., Sninsky, J. and White, T. (1990), PCR Protocols: A Guide to Methods and Applications. San Diego: Academic Press.

106 Tabor, S. and Richardson, C.C. (1987), DNA sequence analysis with a modified bacteriophage T7 DNA polymerase, Proc. Natl. Acad. Sci. USA 84, pp. 4767–4772.

107 Tabor, S. and Richardson, C.C. (1989), Effect of manganese ions on the incorporation of dideoxynucleotides by bacteriophage T7 DNA polymerase and *E. coli* DNA polymerase I, Proc. Natl. Acad. Sci. USA 86, pp. 4076–4080.

108 Voss, H., Schwager, C., Kristensen, T., Duthie, S., Olsson, A., Erfle, H., Stegemann, J., Zimmermann, J. and Ansorge, W. (1989), One-step reaction protocol for automated DNA sequencing with T7 DNA polymerase results in uniform labeling, Methods Mol. Cell. Biol. 1, pp. 155–159.

109 Murray, V. (1989), Improved double-stranded DNA sequencing using the linear polymerase chain reaction, Nucl. Acids Res. 17, pp. 8889.

110 Craxton, M. (1991), Linear amplification sequencing, a powerful method for sequencing DNA, Methods: a Companion to Methods in Enzymology 3, pp. 20–26.

111 Innis, M.A., Myambo, K.B., Gelfand, D.H. and Brow, M.D. (1988), DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA, Proc. Natl. Acad. Sci. USA 85, pp. 9436–9440.

112 Parker L.T., Deng Q., Zakeri H., Carlson C., Nickerson D.A. and Kwok P.Y. (1995), Peak height variations in automated sequencing of PCR products using Taq dye-terminator chemistry. BioTechniques 19, pp. 116–121.

113 Reeve, M.A. and Fuller, C.W. (1995), A novel thermostable polymerase for DNA sequencing, Nature 376, pp. 796–797.

114 Parker, L.T., Zakeri, H., Deng, Q., Spurgeon, S., Kwok, P.Y. and Nickerson, D.A. (1996), AmpliTaq DNA polymerase, FS dye-terminator sequencing: analysis of peak height patterns, BioTechniques 21, pp. 694–699.

115 Lee, L.G., Connell, C.R., Woo, S.L., Cheng, R.D., McArdle, B.F., Fuller, C.W., Halloran, N.D. and Wilson, R.K. (1992), DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments, Nucleic Acids Res. 20, pp. 2471–2483.

116 Roemer, S., Boveia, V., Buzby, P., DiMeo, J., Draney, D., Johnson, C., Narayanan, N., Olive, M., Vanek, M. and Osterman, H. (2000), New dye-labeled acycloterminators for automated infrared DNA sequencing, Advances in Genome Biology and Technology I (Abstract only).

117 Voss, H., Nentwich, U., Duthie, S., Wiemann, S., Benes, V., Zimmermann, J. and Ansorge, W. (1997), Automated cycle sequencing with Taquenase: Protocols for internal labeling, dye primer and "doublex" simultaneous sequencing, BioTechniques 23, pp. 312–318.

118 Steffens, D.L., Jang, G.Y., Sutter, S.L., Brumbaugh, J.A., Middendorf, L.R., Muhlegger, K., Mardis, E.R., Weinstock, L.A. and Wilson, R.K. (1995), An infrared fluorescent dATP for labeling DNA, Genome Res. 5, pp. 393–399.

119 Wiemann, S., Stegemann, J., Grothues, D., Bosch, A., Estivill, X., Schwager, C., Zimmermann, J., Voss, H. and Ansorge W. (1995), Simultaneous on-line DNA sequencing on both strands with two fluorescent dyes, Anal. Biochem. 224, pp. 117–121.

120 Middendorf, L., Amen, J., Bruce, R., Draney, D., DeGraff, D., Gewecke, J., Grone, D., Humphrey, P., Little, G., Lugade, A., Narayanan, N., Oommen, A., Osterman, H., Peterson, R., Rada, J., Raghavachari, R. and Roemer, S. (1998), Near-infrared fluorescence instrumentation for DNA analysis, in: Near-Infrared Dyes for High Technology Applications (S. Daehne et al., Eds.), pp. 21–54. © 1998 Kluwer Academic Publishers. Printed in the Netherlands.

121 Millipore Corporation (1991), BaseStation Automated DNA Sequencer, Lit. No. PM016, Bedford, MA.

122 Middendorf, L.R., Bruce, J.C., Bruce, R.C., Eckles, R.D., Grone, D.L., Roemer, S.C., Sloniker, G.D., Steffens, D.L., Sutter, S.L., Brumbaugh, J.A. and Patonay, G. (1992), Continuous, on-line DNA sequencing using a versatile infrared laser scanner/electrophoresis apparatus, Electrophoresis 13, pp. 487–494.

123 Yager, T.D., Baron, L., Batra, R., Bouevitch, A., Chan, D., Chan, K., Darasch, S., Gilchrist, R., Izmailov, A., Lacroix, J.-M., Marchelleta, K., Renfrew, J., Rushlow, D., Steinbach, E., Ton, C., Waterhouse, P., Zaleski, H., Dunn, J.M. and Stevens, J. (1999), High

performance DNA sequencing, and detection of mutations and polymorphisms, on the Clipper sequencer, Electrophoresis 20, pp. 1280–1300.
124 Patton, W.F. (2000), Making blind robots see: the synergy between fluorescent dyes and imaging devices in automated proteomics, BioTechniques 28, pp. 944–957.
125 Swerdlow, H., Zhang, J.Z., Chen, D.Y., Harke, H.R., Grey, R., Wu, S. and Dovichi, N.J. (1991), Three DNA sequencing methods using capillary gel electrophoresis and laser-induced fluorescence, Anal. Chem. 63, pp. 2835–2841.
126 Brumley, Jr., R.L. and Smith, L.M. (1991), Rapid DNA sequencing by horizontal ultrathin gel electrophoresis, Nucl. Acids Res. 19, pp. 4121–4126.
127 Huang, X.C., Quesada, M.A. and Mathies, R.A. (1992), Capillary array electrophoresis using laser-excited confocal fluorescence detection, Anal. Chem. 64, pp. 967–972.
128 Huang, X.C., Quesada, M.A. and Mathies, R.A. (1992), DNA sequencing using capillary array electrophoresis, Anal. Chem. 64, pp. 2149–2154.
129 Kheterpal, I., Scherer, J.R., Clark, S.M., Radharkrishnan, A., Ju, J., Ginther, C.L., Sensabaugh, G.F. and Mathies, R.A. (1996), DNA sequencing using a four-color confocal fluorescence capillary array scanner, Electrophoresis 17, pp. 1852–1859.
130 Tan, H. and Yeung, E.S. (1997), Characterization of dye-induced mobility shifts affecting DNA sequencing in poly(ethylene oxide) sieving matrix, Electrophoresis 18, pp. 2893–2900.
131 O'Brien, K.M., Ironside, M.A., Athanasiou, M.C., Basit, M.A., Evans, G.A. and Garner, H.R. (1998), Correcting data shifts in gel files created by Model 377 DNA sequencers, BioTechniques 24, pp. 1002–1003.
132 O'Brien, K.M., Schageman, J.J., Major, T.H., Evans, G.A. and Garner, H.R. (1998), Improving read lengths by recomputing the matrices of Model 377 DNA sequencers, BioTechniques 24, pp. 1014–1016.
133 Ernst, L.A., Gupta, R.K., Mujumdar, R.B. and Waggoner, A.L. (1989), Cyanine dye labeling reagents for sulfhydryl groups, Cytometry 10, pp. 3–10.
134 Patonay, G. and Antoine, M.D. (1991), Near-infrared fluorogenic labels: New approach to an old problem, Anal. Chem. 63, pp. 321A–327A.
135 Glazer, A.N. and Mathies, R.A. (1997), Energy-transfer fluorescent reagents for DNA analyses, Curr. Opin. Biotechnol. 8, pp. 94–102.
136 Brumley, Jr., R.L. and Luckey, J.A. (1996) An improved high-throughput DNA fragment analyzer employing horizontal ultrathin gel electrophoresis, Proc. SPIE 2680, pp. 349–361.
137 Nordman, E. and Connell, C. (1996), New Optical Design for automated DNA sequencer, Proc. SPIE 2680, pp. 290–293.
138 Ueno, K. and Yeung, E.S. (1994), Simultaneous monitoring of DNA fragments separated by electrophoresis in a multiplexed array of 100 capillaries, Anal. Chem. 66, pp. 1424–1431.
139 Pang, H., Pavski, V. and Yeung, E.S. (1999), DNA sequencing using 96-capillary array electrophoresis, J. Biochem. Biophys. Methods 42, pp. 121–132.
140 Goetzinger, W., Kotler, L., Carrilho, E., Ruiz-Martinez, M.C., Salas-Solano, O. and Karger, B.L. (1998), Characterization of high molecular mass linear polyacrylamide powder prepared by emulsion polymerization as a replaceable polymer matrix for DNA sequencing by capillary electrophoresis, Electrophoresis 19, pp. 242–248.
141 Heller, C. (1998), Finding a universal low viscosity polymer for DNA separation (II), Electrophoresis 19, pp. 3114–3127.
142 Barron, A.E. and Zuckermann, R.N. (1999), Review: Bioinspired polymeric materials: in-between proteins and plastics, Curr. Opin. Chem. Biol. 3, pp. 681–687.
143 Schmalzing, D., Koutny, L., Salas-Solano, O., Adourian, A., Matsudaira, P. and Ehrlich, D. (1999), Review: Recent developments in DNA sequencing by capillary and microdevice electrophoresis, Electrophoresis 20, pp. 3066–3077, and references 11–15 therein.
144 Kolner, D.E., Mead, D.A. and Smith L.M. (1992), Ultrathin DNA sequencing gels using microtrough vertical electrophoresis plates, BioTechniques 13, pp. 338–339.
145 Erfle, H., Ventzki, R., Voss, H., Rechmann, S., Benes, V., Stegemann, J. and Ansorge, W. (1997), Simultaneous loading of 200 sample lanes for DNA sequencing on vertical and horizontal, standard and ultrathin gels, Nucl. Acids. Res. 25, pp. 2229–2230.
146 Kambara, H. and Takahashi, S. (1993), Multiple-sheathflow capillary array DNA analyser, Nature 361, pp. 565–566.

147 Takahashi, S., Murakami, K., Anazawa, T. and Kambara, H. (1994), Multiple sheath-flow gel capillary-array electrophoresis for multicolor fluorescent DNA detection, Anal. Chem. 66, pp. 1021–1026.

148 Chen, D., Peterson, M.D., Brumley, Jr., R.L., Giddings, M.C., Buxton, E.C., Westphall, M., Smith, L. and Smith L.M. (1995), Side excitation of fluorescence in ultrathin slab gel electrophoresis, Anal. Chem. 67, pp. 3405–3411.

149 Anazawa, T., Takahashi, S. and Kambara, H. (1996), A capillary array gel electrophoresis system using multiple laser focusing for DNA sequencing, Anal. Chem. 68, pp. 2699–2704.

150 Quesada, M., Dhadwal, H., Fisk, D. and Studier, F.W. (1998), Multi-capillary optical waveguides for DNA sequencing, Electrophoresis 19, pp. 1415–1427.

151 Anazawa, T., Takahashi, S. and Kambara, H. (1999), A capillary-array electrophoresis system using side-entry on-column laser irradiation combined with glass rod lenses, Electrophoresis 20, pp. 539–546.

152 Quesada, M.A., Rye, H.S., Gingrich, J.C., Glazer, A.N. and Mathies, R.A. (1991), High-sensitivity DNA detection with a laser-excited confocal fluorescence gel scanner, BioTechniques 10, pp. 616–625.

153 Dovichi, N.J. (1997), Review: DNA sequencing by capillary electrophoresis, Electrophoresis 18, pp. 2393–2399.

154 Nelson, M., VanEtten, J.L. and Grabherr, R. (1992), DNA sequencing of four bases using three lanes, Nucl. Acids Res. 20, pp. 1345–1348.

155 Nelson, M., Zhang, Y., Steffens, D.L., Grabherr, R. and VanEtten, J.L. (1993), Sequencing two DNA templates in five channels by digital compression, Proc. Natl. Acad. Sci. USA, pp. 1647–1651.

156 Scherer, J.R., Kheterpal, I., Radhakrishnan, A., Ja, W.W. and Mathies, R.A. (1999), Ultra-high throughput rotary capillary array electrophoresis scanner for fluorescent DNA sequencing and analysis, Electrophoresis 20, pp. 1508–1517.

157 Roque-Biewer, M., Sharaf, M., Taylor, W., Labrenz, J., Menchen, S. and Tynan, K. (1998), Expanding the capability of sequencing and fragment analysis by the introduction of a 5th dye on a multiple capillary electrophoresis instrument, Microbial and Comparative Genomics 3, pp. C-92 (Abstract C-15 only).

158 Ventzki, R., Stegemann, J., Benes, V., Rechmann, S. and Ansorge, W. (1998), Simultaneous loading of 200 sample lanes for DNA sequencing on vertical and horizontal, standard and ultrathin gels, Microbial and Comparative Genomics 3, pp. C-57 (Abstract A-27 only).

159 Church, G.M. and Kieffer-Higgins, S. (1988), Multiplex DNA sequencing, Science 240, pp. 185–188.

160 Church, G.M. and Gilbert, W. (1984), Genomic sequencing, Proc. Natl. Acad. Sci. USA 81, pp. 1991–1995.

161 Yang, M.M. and Youvan, D.C. (1989), A prospectus for multispectral-multiplex DNA sequencing, Bio/Technology 7, pp. 576–580.

162 Kambara, H., Nagai, K. and Hayasaka, S. (1991), Real time automated simultaneous double-stranded DNA sequencing using two-color fluorophore labeling, Bio/Technology 9, pp. 648–651.

163 Ansorge, W., Voss, H., Wirkner, U., Schwager, C., Stegemann, J., Pepperkok, R., Zimmermann, J. and Erfle. J. (1989), Automated Sanger DNA sequencing with one label in less than four lanes on gel, J. Biochem. Biophys. Methods 20, pp. 47–52.

164 Ansorge, W., Zimmermann, J., Schwager, C., Stegemann, J., Erfle, H. and Voss, H. (1990), One label, one tube, Sanger DNA sequencing in one and two lanes on a gel, Nucl. Acids Res. 18, pp. 3419–3420.

165 Pentoney, Jr., S.L., Konrad, K.D. and Kaye, W. (1992), A single-fluor approach to DNA sequence determination using high performance capillary electrophoresis, Electrophoresis 13, pp. 467–474.

166 Chen D., Harke, H. and Dovichi, N.J. (1992), Two-label peak-height encoded DNA sequencing by capillary gel electrophoresis: three examples, Nucl. Acids Res. 20, pp. 4873–4880.

167 Starke H.R., Yan, J.Y., Zhange, J.Z., Mühlegger, K., Effgen, K. and Dovichi, N.J. (1994) Internal fluorescence labeling with fluorescent deoxynucleotides in two-label peak-height encoded DNA sequencing by capillary electrophoresis, Nucl. Acids Res. 22, pp. 3997–4001.

168 Li, Q. and Yeung, E.S. (1995), Simple two-color base-calling schemes for DNA sequencing based on standard four-label Sanger chemistry, Appl. Spect. 49, pp. 1528–1533.

169 Williams, D.C. and Soper, S.A. (1995), Ultrasensitive near-IR fluorescence detection for capillary gel electrophoresis and DNA sequencing applications, Anal. Chem. 67, pp. 3427–3432.

170 Negri, R., Costanzo, G., Saladino, R. and DiMauro, E. (1996), One-step, one-lane chemical DNA sequencing by *N*-Methylformamide in the presence of metal ions, BioTechniques 21, pp. 910–917.

171 Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998), Base-calling of automated sequencer traces using *Phred*. I. Accuracy Assessment, Genome Res. 8, pp. 175–185.

172 Ewing, B. and Green, P. (1998), Base-calling of automated sequencer traces using *Phred*. II. Error probabilities, Genome Res. 8, pp. 186–194.

173 Richterich, P. (1998), Estimation of errors in "raw" DNA sequences: A validation study, Genome Res. 8, pp. 251–259.

174 Richterich, P., Humphrey, P. and Amen, J. (1998), Optimization of LI-COR trace files for processing by PHRED and PHRAP, Microbial and Comparative Genomics 3, pp. C-91 (Abstract C-10 only).

175 Gaasterland T. and Sensen, C.W. (1996), Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture, Biochimie 78, pp. 302–310.

176 Lander, E.S. and Waterman, M.S. (1988), Genomic mapping by fingerprinting random clones: a mathematical analysis, Genomics 2, pp.231–239.

177 Batzoglou, S., Berger, B., Mesirov, J. and Lander, E.S. (1999), Sequencing a genome by walking with clone-end sequences: A mathematical analysis, Genome Res. 9:1163–1174.

178 Voss, H., Wiemann, S., Grothues, D., Sensen, C., Zimmermann, J., Schwager, C., Stegemann, J., Erfle, H., Rupp, T. and Ansorge, W. (1993), Automated low-redundancy large-scale DNA sequencing by primer walking, BioTechniques 15, pp. 714–721.

179 Kostichka, A.J., Marchbanks, M.L., Brumley, R.L. Jr., Drossman, H. and Smith, L.M. (1992), High speed automated DNA sequencing in ultrathin slab gels, BioTechnology 10, pp. 78–81.

180 Carninci, P., Volpatti, F. and Schneider, C. (1995), A discontinuous buffer system increasing resolution and reproducibility in DNA sequencing on high voltage horizontal ultrathin-layer electrophoresis, Electrophoresis 16, pp. 1836–1845.

181 Smith, L.M., Brumley, R.L. Jr., Buxton, E.C., Giddings, M., Marchbanks, M. and Tong, X. (1996), High-speed automated DNA sequencing in ultrathin slab gels, Methods Enzymol. 271, pp. 219–237.

182 Yager, T.D., Dunn, J.M. and Stevens, S.K. (1997), High-speed DNA sequencing in ultrathin slab gels, Curr. Opin. Biotechnol. 8, pp. 107–113.

183 Luckey, J.A. and Smith, L.M. (1993), Optimization of electric field strength for DNA sequencing in capillary gel electrophoresis, Anal. Chem. 65, pp. 2841–2850.

184 Yan, J.Y., Best, N., Zhang, J.Z., Ren, H.J., Jiang, R., Hou, J. and Dovichi, N. (1996), The limiting mobility of DNA sequencing fragments for both cross-linked and non-cross-linked polymers in capillary electrophoresis: DNA sequencing at 1200 V cm^{-1}, Electrophoresis 17, pp. 1037–1045.

185 McIndoe, R.A., Hood, L. and Baumgartner, R.E. (1996), An analysis of the dynamic range and linearity of an infrared DNA sequencer, Electrophoresis 17, pp. 652–658.

186 Boguslavsky, J. (2000), DNA sequencers reach production scale, Drug Discovery and Development Magazine January/February, 36–38.

187 Shibata, K., Itoh, M., Aizawa, K., Nagaoka, S., Sasaki, N., Carninci, P., Konno, H., Akiyama, J., Nishi, K., Kitsunai, T., Tashiro, H., Itoh, M., Sumi, N., Ishii, Y., Nakamura, S., Hazama, M., Nishine, T., Harada, A., Yamamoto, R., Matsumoto, H., Sakaguchi, S., Ikegami, T., Kashiwagi, K., Fujiwake, S., Inoue, K. and Togawa, Y. (2000), RIKEN integrated sequence analysis (RISA) system – 384-format sequencing pipeline with 384 multicapillary sequencer, Genome Res. 10, pp. 1757–1771.

188 Bashkin, J.S., Bartosiewicz, M., Roach, D., Leong, J. Barker, D. and Johnston, R. (1996), Implementation of a capillary array electrophoresis instrument, J. Cap. Elec. 3, pp. 61–68.

189 Bashkin, J., Marsh, M., Barker, D. and Johnston, R. (1996), DNA sequencing by capillary electrophoresis with a hydroxyethylcellulose sieving buffer, Appl. Theor. Electrophor. 6, pp. 23–28.

190 Marsh, M., Tu, O., Dolnik, V., Roach, D., Solomon, N., Bechtol, K., Smietana, P., Wang, L., Li, X., Cartwright, P., Marks, A.,

Barker, D., Harris, D. and Bashkin, J. (1997), High-throughput DNA sequencing on a capillary array electrophoresis system, J. Capillary Electrophor. 4, pp. 83–89.

191 Behr, S., Mätzig, M., Levin, A., Eickhoff, H. and Heller, C. (1999), A fully automated multicapillary electrophoresis device for DNA analysis, Electrophoresis 20, pp. 1492–1507.

192 Dolnik, V. (1999), Review: DNA sequencing by capillary electrophoresis, J. Biochem. Biophys. Methods 41, pp. 103–119.

193 Ruiz-Martinez, M.C., Salas-Solano, O., Carrilho, E., Kotler, L. and Karger, B.L. (1998), A sample purification method for rugged and high-performance DNA sequencing by capillary electrophoresis using replaceable polymer solutions. A. Development of the cleanup protocol, Anal. Chem. 70, pp. 1516–1527.

194 Salas-Solano, O., Ruiz-Martinez, M.C., Carrilho, E., Kotler, L. and Karger, B.L. (1998), A sample purification method for rugged and high-performance DNA sequencing by capillary electrophoresis using replaceable polymer solutions. B. Quantitative determination of the role of sample matrix components on sequencing analysis, Anal. Chem. 70, pp. 1528–1535.

195 Woolley, A.T., Sensabaugh, G.F. and Mathies, R.A. (1997), High-speed DNA genotyping using microfabricated capillary array electrophoresis chips, Anal. Chem. 69, pp. 2181–2186.

196 Schmalzing, D., Adourian, A., Koutny, L., Ziaugra, L., Matsudaira, P. and Ehrlich, D. (1998), DNA sequencing on microfabricated electrophoretic devices, Anal. Chem. 70, pp. 2303–2310.

197 Simpson, P.C., Roach, D., Woolley, A.T., Thorsen, T., Johnston, R., Sensabaugh, G.F. and Mathies, R.A. (1998), High-throughput genetic analysis using microfabricated 96-sample capillary array electrophoresis microplates, Proc. Natl. Acad. Sci. USA 95, pp. 2256–2261.

198 Waters, L.C., Jacobson, S.C., Kroutchinina, N., Khandurina, J., Foote, R.S. and Ramsey, J.M. (1998), Multiple sample PCR amplification and electrophoretic analysis on a microchip, Anal. Chem. 70, pp. 5172–5176.

199 Liu, S., Shi, Y., Ja, W.W. and Mathies, R.A. (1999), Optimization of high-speed DNA sequencing on microfabricated capillary electrophoresis channels, Anal. Chem. 71, pp. 566–573.

200 Ramsey, J.M. (1999), The burgeoning power of the shrinking laboratory, Nat. Biotechnol. 17, pp. 1061–1062.

201 Shi, Y., Simpson, P.C., Scherer, J.R., Wexler, D., Skibola, C., Smith M.T. and Mathies, R.A. (1999), Radial capillary array electrophoresis microplate and scanner for high-performance nucleic acid analysis, Anal. Chem. 71, pp. 5354–5361.

202 Schmalzing, D., Tsao, N., Koutny, L., Chisholm, D., Srivastava, A., Adourian, A., Linton, L., McEwan, P., Matsudaira, P. and Ehrlich, D. (1999), Toward real-world sequencing by microdevice electrophoresis, Genome Res. 9, pp. 853–858.

203 Schmalzing, D., Belenky, A., Novotny, M.A., Koutny, L., Salas-Solano, O., El-Difrawy, S., Adourian, A., Matsudaira, P. and Ehrlich, D. (2000), Microchip electrophoresis: a method for high-speed SNP detection, Nucl. Acids Res. 28, pp. e43 (electronic article).

204 Becker, H. and Gärtner (2000), Polymer microfabrication methods for microfluidic analytical applications, Electrophoresis 21, pp. 12–26.

205 McDonald, J.C., Duffy, D.C., Anderson, J.R., Chiu, D.T., Wu, H., Schueller, O.J.A. and Whitesides, G.M. (2000), Fabrication of microfluidic systems in poly(dimethylsiloxane), Electrophoresis 21, pp. 27–40.

206 Dolnik, V., Liu, S. and Jovanovich, S. (2000), Capillary electrophoresis on microchip, Electrophoresis 21, pp. 41–54.

207 Carrilho, E. (2000), DNA sequencing by capillary array electrophoresis and microfabricated array systems, Electrophoresis 21, pp. 55–65.

208 Liu, S. (2003), A microfabricated hybrid device for DNA sequencing, Electrophoresis 24, pp. 3755–3761.

209 Liu, S., Elkin, C. and Kapur, H. (2003), Sequencing of real-world samples using a microfabricated hybrid device having unconstrained straight separation channels, Electrophoresis 24, pp. 3762–3768.

210 Mueller, O., Hahnenberger, K., Dittmann, M., Yee, H., Dubrow, R., Nagle, R. and Ilsley, D. (2000), A microfluidic system for high-speed reproducible DNA sizing and quantitation, Electrophoresis 21, pp. 128–134.

211 Simpson, J.W., Ruiz-Martinex, M.C., Mulhern, G.T., Berka, J., Latimer, D.R., Ball, J.A., Rothberg, J.M. and Went, G.T. (2000), A transmission imaging spectrograph and

microfabricated channel system for DNA analysis, Electrophoresis 21, pp. 135–149.
212 Ren, H., Karger, A.E., Oaks, F., Menchen, S., Slater, G. and Drouin, G. (1999), Separating DNA sequencing fragments without a sieving matrix, Electrophoresis 20, pp. 2501–2509.
213 Backhouse, C., Caamano, M., Oaks, F., Nordman, E., Carrillo, A., Johnson, B. and Bay, S. (2000), DNA sequencing in a monolithic microchannel device, Electrophoresis 21, pp. 150–156.
214 Brenner, S. (1996), DNA sequencing by stepwise ligation and cleavage, U.S. Patent 5 552 278.
215 Brenner, S. (1996), DNA sequencing by stepwise ligation and cleavage, U.S. Patent 5 599 675.
216 Brenner, S. (1997), Massively parallel sequencing of sorted polynucleotides, U.S. Patent 5 695 934.
217 Brenner, S. and DuBridge, R.B. (1998), DNA sequencing by stepwise ligation and cleavage, U.S. Patent 5 714 330.
218 Brenner, S., Williams, S.R., Vermaas, E.H., Storck, T., Moon, K., McCollum, C., Mao, J-I, Luo, S., Kirchner, J.J., Eletr, S., DuBridge, R.B., Burcham, T. and Albrecht, G. (2000), In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs, Proc. Natl. Acad. Sci. USA 97, pp. 1665–1670.
219 Albrecht, G., Brenner, S., DuBridge, R.B., Lloyd, D.H. and Pallas, M.C. (2000), Massively parallel signature sequencing by ligation of encoded adaptors, U.S. Patent 6 013 445.
220 Jones, D.H. (1997), An iterative and regenerative method for DNA sequencing, BioTechniques 22, pp. 938–946.
221 Macevicz, S.C. (1998), DNA sequencing by parallel oligonucleotide extension, U.S. Patent 5 750 341.
222 Canard, B. and Sarfati, R.S. (1994), DNA polymerase fluorescent substrates with reversible 3′-tags, Gene 148, pp. 1–6.
223 Metzker, M.L., Raghavachari, R., Richards, S., Jacutin, S.E., Civitello, A., Burgess, K. and Gibbs, R.A. (1994), Termination of DNA synthesis by novel 3′-modified-deoxyribonucleoside 5′-triphosphates, Nucl. Acids Res. 22, pp. 4259–4267.
224 Li, Z., Bai, X., Ruparel, H., Kim, S., Turro, N.J. and Ju, J. (2003), A photocleavable fluorescent nucleotide for DNA sequencing and analysis, Proc. Natl. Acad. Sci. U.S.A., pp. 414–419.
225 Brennan, T.M. and Heyneker, H.L. (1995), Methods and composition for determining the sequence of nucleic acids, U.S. Patent 5 403 708.
226 Hyman, E.D. (1988), A new method of sequencing DNA, Anal. Biochem. 174, pp. 423–436.
227 Nyrén, P. (1987), Enzymatic method for continuous monitoring of DNA polymerase activity, Anal. Biochem. 167, pp. 235–238.
228 Nyrén, P., Pettersson, B. and Uhlén, M. (1993), Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay, Anal. Biochem. 208, pp. 171–175.
229 Nyrén, P.J. (1996), Apyrase immobilized on paramagnetic beads used to improve detection limits in bioluminometric ATP monitoring, J. Biolumin. Chemilumin. 9, pp. 29–34.
230 Ronaghi, M., Karamuhamed, S., Pettersson, B., Uhlén, M. and Nyrén, P. (1996), Real-time DNA sequencing using detection of pyrophosphate release, Anal. Biochem. 242, pp. 84–89.
231 Ronaghi, M., Uhlén, M., and Nyrén, P. (1998), A sequencing method based on real-time pyrophosphate, Science 281, pp. 363–365.
232 Ronaghi, M., Pettersson, B., Uhlén, M., and Nyrén, P. (1998), PCR-introduced loop structure as primer in DNA sequencing, BioTechniques 25, pp. 876–884.
233 Ronaghi, M., Nygren, M., Lundeberg, J. and Nyrén, P. (1999), Analyses of secondary structures in DNA by pyrosequencing, Anal. Biochem. 267, pp. 65–71.
234 Ahmadian, A., Lundeberg, J., Nyrén, P., Uhlén, M and Ronaghi, M. (2000), Analysis of the *p53* tumor suppressor gene by pyrosequencing, BioTechniques 28, pp. 140–147.
235 Drmanac, R., Labat, I., Brukner, I. and Crkvenjakov, R. (1989), Sequencing of megabase plus DNA by hybridization: theory of the method, Genomics 4, pp. 114–128.
236 Drmanac, R., Drmanac, S., Labat, I., Crkvenjakov, R., Vicentic, A. and Gemmell, A. (1992), Sequencing by hybridization: towards an automated sequencing of one million M13 clones arrayed on membranes, Electrophoresis 13, pp. 566–573.

237 Drmanac, R., Drmanac, S., Strezoska, Z., Paunesku, T., Labat, I., Zeremski, M., Snoddy, J., Funkhouser, W.K., Koop, B., Hood, L. et al. (1993), DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing, Science 260, pp. 1649–1652.

238 Drmanac R. and Drmanac S. (1999), cDNA screening by array hybridization, Methods Enzymol. 303, pp. 165–178.

239 Köster, H., Tang, K., Fu, D.-J., Braun, A., van den Boom, D., Smith, C.L., Cotter, R.J. and Cantor, C.R. (1996), A strategy for rapid and efficient DNA sequencing by mass spectrometry, Nat. Biotechnol. 14, pp. 1123–1128.

240 Fu, D.J., Tang, K., Braun, A., Reuter, D., Darnhofer-Demar, B., Little, D.P., O'Donnell, M.J., Cantor, C.R. and Köster, H. (1998), Sequencing exons 5 to 8 of the *p53* gene by MALDI-TOF mass spectrometry, Nat. Biotechnol. 16, pp. 381–384.

241 Kirpekar, F., Nordhoff, E., Larsen, L.K., Kristiansen, K., Roepstorff, P. and Hillenkamp, F. (1998), DNA sequence analysis by MALDI mass spectrometry, Nucl. Acids Res. 26, pp. 2554–2559.

242 Taranenko, N.I., Allman, S.L., Golovlev, V.V., Taranenko, N.V., Isola, N.R. and Chen, C.H. (1998), Sequencing DNA using mass spectrometry for ladder detection, Nucl. Acids Res. 26, pp. 2488–2490.

243 Griffin, T.J., Hall, J.G., Prudent, J.R. and Smith, L.M. (1999), Direct genetic analysis by matrix-assisted laser desorption/ionization mass spectrometry, Proc. Natl. Acad. Sci. USA 96, pp. 6301–6306.

244 Arlinghaus, H.F., Kwoka, M.N., Guo, X.-Q. and Jacobson, K.B. (1997), Multiplexed DNA sequencing and diagnostics by hybridization with enriched stable isotope labels, Anal. Chem. 69, pp. 1510–1517.

245 Chen, X., Fei, Z., Smith, L.M., Bradbury, E.M. and Majidi, V. (1999), Stable-isotope-assisted MALDI-TOF mass spectrometry for accurate determination of nucleotide compositions of PCR products, Anal. Chem. 71, pp. 3118–3125.

246 Serpico, S.B. and Vernazza, G. (1987), Problems and prospects in image processing of two-dimensional gel electrophoresis, Opt. Engr. 26, pp. 661–668.

247 Giddings, M.C., Brumley, R.L., Haker, M. and Smith, L.M. (1993), An adaptive object oriented strategy for base calling in DNA sequence analysis, Nucl. Acids Res. 21, pp. 4530–4540.

248 Giddings, M.C., Severin, J., Westphall, M., Wu, J. and Smith, L.M. (1998), A software system for data analysis in automated DNA sequencing, Genome Res. 8, pp. 644–665.

249 Zakeri, H., Amparo, G., Chen, S.M., Spurgeon, S. and Kwok, P.Y. (1998), Peak height pattern in dichloro-rhodamine and energy transfer dye terminator sequencing, BioTechniques 25, pp. 406–410, 412–414.

250 Davies, S., Eizenman, M., Pasupathy, S., Muller, W. and Slater, G. (1999), Models of local behavior of DNA electrophoresis peak parameters, Electrophoresis 20, pp. 1443–1454.

251 Song, J.M. and Yeung, E.S. (2000), Alternative base-calling algorithm for DNA sequencing based on four-label multicolor detection, Electrophoresis 21, pp. 807–815.

252 Starita-Geribaldi, M., Houri, A. and Sudaka, P. (1993), Lane distortions in gel electrophoresis patterns, Electrophoresis 14, pp. 773–781.

253 Golden, J.B. III, Torgersen, D. and Tibbetts, C. (1993), Pattern recognition for automated DNA sequencing: I. On-line signal conditioning and feature extraction for base-calling, Intelligent Systems for Molecular Biology 1, pp. 136–144.

254 Chen, W.Q. and Hunkapiller, T. (1992), Sequence accuracy of large DNA sequencing projects, DNA Sequence–J.DNA Sequencing and Mapping 2, pp. 335–342.

255 Swerdlow, H., Dew-Jager, K.E., Brady, K., Grey, R., Dovichi, N.J. and Gesteland, R. (1992), Stability of capillary gels for automated sequencing of DNA, Electrophoresis 13, pp. 475–483.

256 Swerdlow, H., Dew-Jager, K. and Gesteland, R.F. (1994), Reloading and stability of polyacrylamide slab gels for automated DNA sequencing, BioTechniques 16, pp. 684–685.

257 Desruisseaux, C., Slater, G.W. and Drouin, G. (1998), The gel edge electric field gradients in denaturing polyacrylamide gel electrophoresis, Electrophoresis 19, pp. 627–634.

258 Koutny, L.B. and Yeung, E.S. (1992), Automated image analysis for distortion compensation in sequencing gel electrophoresis, Appl. Spect 46, pp. 136–141.

259 Dear, S. and Staden, R. (1992), A standard file format for data from DNA sequencing

instruments, DNA Sequence-J.DNA Sequencing and Mapping 3, pp. 107–110.
260 Wu, A. and Mislan, D. (1992), Automated DNA sequencing: an image processing approach, Appl. Theor. Electrophor 3, pp. 223–228.
261 Ives, J.T., Gesteland, R.F. and Stockham, T.G. (1994), An automated film reader for DNA sequencing based on homomorphic deconvolution, IEEE Trans. Biomed. Engr. 41, pp, 509–518.
262 Nishikawa, T. and Kambara, H. (1991), Analysis of limiting factors of DNA band separation by a DNA sequencer using fluorescence detection, Electrophoresis 12, pp. 623–631.
263 Sanders, J.Z., Petterson, A.A., Hughes, P.J., Connell, C.R., Raff, M., Mechen, S., Hood, L.E. and Teplow, D.B. (1991), Imaging as a tool for improving length and accuracy of sequence analysis in automated fluorescence-based DNA sequencing, Electrophoresis 12, pp. 3–11.
264 Aldroubi, A. and Garner, M.M. (1992), Minimal electrophoresis time for DNA sequencing, BioTechniques 13, pp. 620–624.
265 Grossman, P.D., Menchen, S. and Hershey, D. (1992), Quantitative analysis of DNA sequencing electrophoresis, Genet. Anal. Tech. Appl. 9, pp. 9–16.
266 Slater, G.W. and Drouin, G. (1992), Why can we not sequence thousands of DNA bases on a polyacrylamide gel?, Electrophoresis 13, pp. 574–582.
267 Luckey, J.A., Norris, T.B. and Smith, L.M. (1993), Analysis of resolution in DNA sequencing by capillary gel electrophoresis, J. Phys. Chem. 97, pp. 3067–3075.
268 Luckey, J.A. and Smith, L.M. (1993), A model for the mobility of single-stranded DNA in capillary gel electrophoresis, Electrophoresis 14, pp. 492–501.
269 Ribeiro, E.A. and Sutherland, J.C. (1993), Resolving power: A quantitative measure of electrophoretic resolution, Anal. Biochem. 210, pp. 378–388.
270 Slater, G.W. (1993), Theory of band broadening for DNA gel electrophoresis and sequencing, Electrophoresis 14, pp. 1–7.
271 Slater, G.W., Mayer, P. and Grossman, P.D. (1995), Diffusion, joule heating, and band broadening in capillary gel electrophoresis of DNA, Electrophoresis 16, pp. 75–83.
272 Fang, Y., Zhang, J.Z., Hou, J.Y., Lu, H. and Dovichi, N.J. (1996), Activation energy of the separation of DNA sequencing fragments in denaturing noncross-linked polyacrylamide by capillary electrophoresis, Electrophoresis 17, pp. 1436–1442.
273 Nishikawa, T. and Kambara, H. (1996), Characteristics of single-stranded DNA separation by capillary gel electrophoresis, Electrophoresis 17, pp. 1476–1484.
274 Weiss, G.H. and Kiefer, J.E. (1997), Some properties of a measure of resolution in gel electrophoresis and capillary zone electrophoresis, Electrophoresis 18, pp. 2008–2011.
275 Guttman, A., Benedek, K. and Kalász (1998), On the separation parameters in DNA sequencing by capillary gel electrophoresis, American Laboratory (April), pp. 63–65.
276 Buetow, K.H., Edmonson, M.N. and Cassidy, A.B. (1999), Reliable identification of large numbers of candidate SNPs from public EST data, Nat. Genet. 21, pp. 323–325.
277 Gordon, D., Abajian, C. and Green, P. (1998), Consed: a graphical tool for sequence finishing, Genome Res. 8, pp. 195–202.
278 Huang, X. and Madan A. (1999), CAP3: A DNA sequence assembly program, Genome Res. 9, pp. 868–877.
279 Moore, G.E. (1965), Cramming more components on to integrated circuits, Electronics 38, pp. 114–117.
280 Meindl, J.D., Chen, Q. and Davis, J.A. (2001), Limits on silicon nanoelectronics for terascale integration, Science 293, pp. 2044–2049.
281 Pierce, J.R. (1980), An introduction to information theory: symbols, signals and noise, New York: Dover Publications.
282 Lloyd, S. (2000), Ultimate physical limits to computation, Nature 406, pp. 1047–1054.
283 Stelzer, E.H.K. and Grill, S. (2000), The uncertainty principle applied to estimate focal spot dimensions, Opt. Commun. 173, pp. 51–56.
284 Dovichi, N.J., Martin, J.C., Jett, J.H., Trkula, M. and Keller, R.A. (1984), Laser-induced fluorescence of flowing samples as an approach to single-molecule detection in liquids, Anal. Chem. 56, pp. 348–354.
285 Peck, K., Stryer, L., Glazer, A.N. and Mathies, R.A. (1989), Single-molecule fluorescence detection: autocorrelation criterion and experimental realization with phycoerythrin, Proc. Natl. Acad. Sci. U.S.A., pp. 4087–4091.

286 Mathies, R.A., Peck, K. and Stryer, L. (1990), Optimization of high-sensitivity fluorescence detection, Anal. Chem. 62, pp. 1786–1791.

287 Kuno, M., Fromm, D.P., Hamann, H.F., Gallagher, A. and Nesbitt, D.J. (2000), Nonexponential "blinking" kinetics of single CdSe quantum dots: A universal power law behavior, J. Chem. Phys. 112, pp. 3117–3120.

288 Moerner, W.E. and Fromm, D.P. (2003), Methods of single-molecule fluorescence spectroscopy and microscopy, Rev. Sci. Instrum. 74, pp. 3597–3619.

289 Jett, J.H., Keller, R.A., Martin, J.C., Marrone, B.L., Moyzis, R.K., Ratliff, R.L., Seitzinger, N.K., Shera, E.B. and Stewart, C.C. (1989), High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules, J. Biomol. Struct. Dyn. 7, pp. 301–309.

290 Davis, L.M., Fairfield, F.R., Harger, C.A., Jett, J.H., Keller, R.A., Hahn, J.H., Krakowski, L.A., Marrone, B.L., Martin, J.C., Nutter H.L., et al. (1991), Rapid DNA sequencing based upon single molecule detection, Genet. Anal. Tech. Appl. 8, pp. 1–7.

291 Harding, J.D. and Keller, R.A. (1992), Review: Single-molecule detection as an approach to rapid DNA sequencing, Trends Biotechnol. 10, pp. 55–57.

292 Werner, J.H., Cai, H., Jett, J.H., Reha-Krantz, L., Keller, R.A. and Goodwin, P.M. (2003), Progress towards single-molecule DNA sequencing: a one color demonstration, J. Biotech. 102, pp. 1–14.

293 Dörre, K., Brakmann, S., Brinkmeier, M., Han, K.-T., Riebeseel, K., Schwille, P., Stephan, J., Wetzel, T., Lapczyna, M., Stuke, M., Bader, R., Hinz, M., Seliger, H., Holm, J., Eigen, M. And Rigler, R. (1997), Techniques for single molecule sequencing, Bioimaging 5, pp. 139–152.

294 Rigler, R. and Seela, F. (2001), DNA-Sequencing at the single molecule level, J. Biotechnol. 86, pp. 161.

295 Eggeling, C., Berger, S., Brand, L., Fries, J.R., Schaffer, J., Volkmer, A. And Seidel, C.A.M. (2001), Data registration and selective single-molecule analysis using multi-parameter fluorescence detection, J. Biotechnol. 86, pp 163–180.

296 Sauer, M., Angerer, B., Ankenbauer, W., Földes-Papp, Z., Göbel, F., Han, K.-T., Rigler, R., Schulz, A., Wolfrum, J. and Zander, C. (2001), Single molecule DNA sequencing in submicrometer channels: state of the art and future prospects, J. Biotechnol. 86, pp 181–201.

297 Földes-Papp, Z., Angerer, B., Thyberg, P., Hinz, M., Wennmalm, S., Ankenbauer, W., Seliger, H., Holmgren, A. And Rigler, R. (2001), Fluorescently labeled model DNA sequences for exonucleolytic sequencing, J. Biotechnol. 86, pp 203–224.

298 Földes-Papp, Z., Angerer, B., Ankenbauer, W. And Rigler, R. (2001), Fluorescent high-density labeling of DNA: error-free substitution for a normal nucleotide, J. Biotechnol. 86, pp 237–253.

299 Dörre, K., Stephan, J., Lapczyna, M., Stuke, M., Dunkel, H. And Eigen, M. (2001) Highly efficient single molecule detection in micro-structures, J. Biotechnol. 86, pp 225–236.

300 Stephan, J., Dörre, K., Brakmann, S., Winkler, T., Wetzel, T., Lapczyna, M., Stuke, M., Angerer, B., Ankenbauer, W., Földes-Papp, Z., Rigler, R. and Eigen, M. (2001), Towards a general procedure for sequencing single DNA molecules, J. Biotechnol. 86, pp 255–267.

301 Seela, F., Feiling, E., Gross, J., Hillenkamp, F., Ramzaeva, N., Rosemeyer, H. and Zulauf, M. (2001), Fluorescent DNA: the development of 7-deazapurine nucleoside triphosphates applicable for sequencing at the single molecule level, J. Biotechnol. 86, pp 269–279.

302 Hinz, M., Gura, S., Nitzan, B., Margel, S. and Seliger, H. (2001), Polymer support for exonucleolytic sequencing, J. Biotechnol. 86, pp 281–288.

303 Augustin, M.A., Ankenbauer, W. and Angerer, B. (2001), Progress towards single-molecule sequencing: enzymatic synthesis of nucleotide-specifically labeled DNA, J. Biotechnol. 86, pp 289–301.

304 Brakmann, S. and Löbermann, S. (2001), High-density labeling of DNA: Preparation and characterization of the target material for single-molecule sequencing, Angew. Chem. 40, pp. 1427–1429.

305 Brakmann, S. and Löbermann, S. (2002), A further step towards single-molecule sequencing: *Escherichia coli* exonuclease III degrades DNA that is fluorescently labeled at each base pair, Angew. Chem. 41, pp. 3215–3217.

306 Brakmann, S. (2004), Optimal enzymes for single-molecule sequencing, Curr. Pharm. Biotech. 5, pp. 119–126.

307 Giller, G., Tasara, T., Angerer, B., Mühlegger, K., Amacker, M. And Winter, H. (2003), Incorporation of reporter molecule-labeled nucleotides by DNA polymerases. I. Chemical synthesis of various reporter group-labeled 2′-deoxyribonucleoside-5′-triphosphates, Nucl. Acids. Res. 31, pp. 2630–2635.

308 Tasara, T., Angerer, B., Damond, M., Winter, H., Dörhöfer, S., Hübscher, U. and Amacker, M. (2003), Incorporation of reporter molecule-labeled nucleotides by DNA polymerases. II. High-density labeling of natural DNA, Nucl. Acids Res. 31, pp. 2636–2646.

309 Ulmer, K.M. (1997), Methods and apparatus for DNA sequencing, U.S. Patent 5 674 743.

310 Yan, Y. and Myrick, M.L. (2001), Identification of nucleotides with identical fluorescent labels based on fluorescence polarization in surfactant solutions, Anal. Chem. 73, pp. 4508–4513.

311 Mitra, R.D., Shendure, J., Olejnik, J., Krzymansak-Olejnik, E. and Church, G.M. (2003), Fluorescent in situ sequencing on polymerase colonies, Anal. Biochem. 320, pp. 55–65.

312 Seo, T.S., Bai, X., Ruparel, H., Li, Z., Turro, N.J. and Ju, J. (2004), Photocleavable fluorescent nucleotides for DNA sequencing on a chip constructed by site-specific coupling chemistry, Proc. Natl. Acad. Sci. U.S.A. 101, pp. 5488–5493.

313 Leamon, J.H., Lee, W.L., Tartaro, K.R., Lanza, J.R., Sarkis, G.J., deWinter, A.D., Berka, J. and Lohman, K.L. (2003), A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions, Electrophoresis 24, pp. 3769–3777.

314 Kling, J. (2003), Ultrafast DNA sequencing, Nat. Biotechnol. 21, pp. 1425–1427.

315 Braslavsky, I., Hebert, B., Kartalov, E. And Quake, S.R. (2003), Sequence information can be obtained from single DNA molecules, Proc. Natl. Acad. Sci. U.S.A. 100, pp. 3960–3964.

316 Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G. and Webb, W.W. (2003), Zero-mode waveguides for single-molecule analysis at high concentrations, Science 299, pp. 682–686.

317 Tegenfeldt, J.O., Bakajin, O., Chou, C.-F., Chan, S.S., Austin, R., Fann, W., Liou, L., Chan, E., Duke, T. and Cox, E.C. (2001), Near-field scanner for moving molecules, Phys. Rev. Lett. 86, pp. 1378–1381.

318 Tegenfeldt, J.O., Prinz, C., Cao, H., Huang, R.L., Austin, R.H., Chou, S.Y., Cox, E.C. and Sturm, J.C. (2004), Micro- and nanofluidics for DNA analysis, Anal. Bioanal. Chem. 378, pp. 1678–1692.

319 Jonietz, E. (2002), The personal genome sequencer, Technology Review, November, pp. 76–79.

320 Ying, L., Bruckbauer, A., Rothery, A.M., Korchev, Y.E. and Klenerman, D. (2002), Programmable delivery of DNA through a nanopipet, Anal. Chem. 74, pp. 1380–1385.

321 Cai, W., Aburatani, H., Stanton, V.P. Jr., Housman, D.E., Wang, Y.K. and Schwartz, D.C. (1995), Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces, Proc. Natl. Acad. Sci. U.S.A. 92, pp. 5164–5168.

322 Aston, C., Mishra, B. and Schwartz, D.C. (1999), Optical mapping and its potential for large-scale sequencing projects, Trends Biotechnol. 17, pp. 297–302.

323 Lim, A., Dimalanta, E.T., Potamousis, K.D., Yen, G., Apodoca, J., Tao, C., Lin, J., Qi, R., Skiadas, J., Ramanathan, A., Perna, N.T., Plunkett, G. 3rd, Burland, V., Mau, B., Hackett, J., Blattner, F.R., Anantharaman, T.S., Mishra, B. and Schwartz, D.C. (2001), Shotgun optical maps of the whole Escherichia coli O157:H7 genome, Genome Res. 11, pp. 1584–1593.

324 Zhou, S., Deng, W., Anantharaman, T.S., Lim, A., Dimalanta, E.T., Wang, J., Wu, T., Chunhong, T., Creighton, R., Kile, A., Kvikstad, E., Bechner, M., Yen, G., Garic-Stankovic, A., Severin, J., Forrest, D., Runnheim, R., Churas, C., Lamers, C., Perna, N.T., Burland, V., Blattner, F.R., Mishra, B., Schwartz, D.C. (2002), A whole-genome shotgun optical map of *Yersinia pestis* strain KIM, Appl. Environ, Microbiol. 68, pp. 6321–6331.

325 Zhou, S., Kvikstad, E., Kile, A., Severin, J., Forrest, D., Runnheim, R., Churas, C., Hickman, J.W., Mackenzie, C., Choudhary, M., Donohue, T., Kaplan, S. and Schwartz, D.C. (2003), Whole-genome shotgun optical mapping of *Rhodobacter sphaeroides* strain 2.4.1 and its use for whole-genome shotgun sequence assembly, Genome Res. 13, pp. 2142–2151.

326 Kasianowicz, J.J., Brandin, E., Branton, D. and Deamer, D.W. (1996), Characterization of individual polynucleotide molecules using a membrane channel, Proc. Natl. Acad. Sci. U.S.A. 93, pp. 13770–13773.

327 Wang, H. and Branton, D. (2001), Nanopores with a spark for single-molecule detection, Nat. Biotechnol. 19, pp. 622–623.

328 Vercoutere, W., Winters-Hilt, S., Olsen, H., Deamer, D., Haussler, D. and Akeson, M. (2001), Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel, Nat. Biotechnol. 19, pp. 248–252.

329 Akeson, M., Branton, D., Kasianowicz, J.J., Brandin, E. and Deamer, D.W. (1999), Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules, Biophys. J. 77, pp. 3227–3233.

330 Meller, A., Nivon, L., Brandin, E., Golovchenko, J. and Branton, D. (2000), Rapid nanopore discrimination between single polynucleotide molecules, Proc. Natl. Acad. Sci. U.S.A. 97, pp. 1079–1084.

331 Meller, A., Nivon, L. and Branton, D. (2001), Voltage-driven DNA translocations through a nanopore, Phys. Rev. Lett. 86, pp. 3435–3438.

332 Meller, A. and Branton, D. (2002), Single molecule measurements of DNA transport through a nanopore, Electrophoresis 23, pp. 2583–2591.

333 Howorka, S., Cheley, S. and Bayley, H. (2001), Sequence-specific detection of individual DNA strands using engineered nanopores, Nat. Biotechnol. 19, pp. 636–639.

334 Deamer, D.W. and Branton, D. (2002), Characterization of nuclei acids by nanopore analysis, Acc. Chem. Res. 35, pp. 817–825.

335 Sauer-Budge, A.F., Nyamwanda, J.A., Lubensky, D.K. and Branton, D. (2003), Unzipping kinetics of double-stranded DNA in a nanopore, Phys. Rev. Lett. 90, pp. e238101 (Epub).

336 Vercoutere, W.A., Winters-Hilt, S., DeGuzman, V.S., Deamer, D., Ridino, S.E., Rodgers, J.T., Olsen, H.E., Marziali, A. and Akeson, M. (2003), Discrimination among individual Watson-Crick base pairs at the termini of single DNA hairpin molecules, Nucl. Acids Res. 31, pp. 1311–1318.

337 Slater, G.W., Desruisseaux, C., Hubert, S.J., Mercier, J.-F., Labrie, J., Boileau, J., Tessier, F. and Pépin, M.P. (2000), Theory of DNA electrophoresis: A look at some current challenges, Electrophoresis 21, pp. 3873–3887.

7
Proteomics and Mass Spectrometry for the Biological Researcher

Sheena Lambert and David C. Schriemer

7.1
Introduction

The concept of the proteome has been with us for approximately 10 years and has made a dramatic impact on how biological research is conducted. Capabilities and influence continue to grow, having had an impact on areas of biological inquiry from agriculture to zoology [1]. As with any rapidly advancing discipline, it becomes difficult to apply a rigorous and widely accepted definition. A "historical" definition of proteome parallels the definition of the genome. That is, proteome refers to the full set of proteins expressed in a given organism. Early stages of the discipline thus focused on sequencing the proteome as a logical progression from genomics. It quickly became apparent that such a definition of the word proteome was dangerously misleading because it suggested that a sequencing or mapping effort is, indeed, an essential activity within proteomics. In recent years there have been fewer attempts to deliver a wholesale catalogue of all proteins in a given organism. It is illustrative to follow the yearly, cumulative citation of the words "proteome" and "proteomics" in the scientific literature (Fig. 7.1). Although it is a stretch to suggest that the word "proteome" is falling out of favor, it is clear that the term does not convey the same research-level utility as the word "proteomics". This latter term is more functional, because it refers to the processes, methods and instrumentation used to study aspects of the proteome. At the risk of over-interpreting this trend in citation, it suggests we are still very much in the definition and "tool-building" phase of proteomics (interestingly, one sees the inverse with the terms genome and genomics). Quite simply, the proteome is a more complex and extensive congregation of biomolecules than the genome, and its complete mapping by singular initiatives is well beyond current capabilities.

Developing a functional definition of *proteomics*, reflecting current capabilities, is thus a more useful an undertaking. It can be defined as the spatio-temporal harvesting, identification, and quantitation of all proteins in a given sample, to study a specific biochemical process. Proteomics considers the flux in all protein chemical properties postulated to be critical to this process (for example, post-translational modifications). The last five years of effort has led to the refinement of our expectations – we see greater utility in targeting proteomics to

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

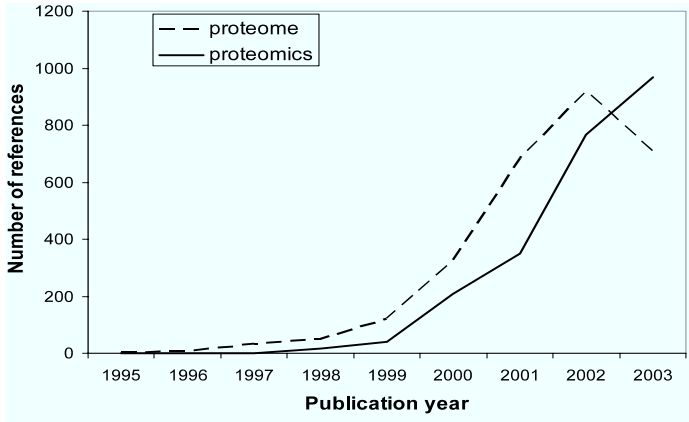


Fig. 7.1 A literature survey of the use of the terms "proteome" and "proteomics".

deliver unbiased data for defined biochemical problems at the expense of wholesale cataloging efforts. In this context we can better understand its relationship to genomics. Genomics enables proteomic initiatives, by provision of the raw sequence. Proteomics returns insight on how the organism utilizes this sequence at the protein level. This utilization can be as simple as confirming that gene products detected at the level of mRNA are present as translated proteins, and as complex as mapping the multiplicity of phosphorylation events induced by extracellular stimuli. The literature of the last few years has reflected this refinement. We continue to see discovery-style applications [2], but increasingly proteomics methods are used to reveal novel proteins in a functional context. For example, proteins associated with known tumor suppressors help expand our understanding of transcriptional regulators [3–5]. Functionally-directed proteomics projects have ranged from "small scale" studies from the perspective of the number of proteins discovered, to large-scale methods such as the yeast interactome [6, 7]. Differential expression studies have been used to monitor and compare up/down regulation of proteins in diseased and "normal" states [8], in response to stimuli [9] and therapeutic treatment [10], for example. In addition to phosphorylation, focused studies involve the elucidation of specific protein structural features such as glycosylation [11, 12]. A wide range of bioconjugation strategies and chemistries have also emerged; these have been designed to improve our ability to identify, quantitate, and characterize [13]. Steady improvement of the instrumentation for purification and analysis continue to expand the range of applications for high-throughput discovery initiatives. Of particular note is the search for disease-related proteins in biological fluids [14, 15]; this has led to a shift in the high-throughput application of proteomics away from drug-target discovery and toward biomarker discovery.

Three fundamental issues must be considered before conducting a proteomics analysis of biological samples – sample complexity, the capacity of the analytical system, and the quality of available databases. If a fundamental goal in proteomics is to identify all proteins in a given sample, a defining technological challenge in proteomics is the absence of amplification. Many proteins are present at a level of no more than a few copies per cell and unless cloned and over-expressed they cannot be signifi-

cantly amplified in their native environment. Because many other proteins are present at extremely high levels the dynamic range of protein expression in a given cellular system is wide – approximately nine orders of magnitude [16]. Tab. 7.1 provides a numerical description of this issue [17]. Whereas the low end of the scale stresses the need for sensitive detection systems, the high end creates problems relating to the capacity of common enrichment systems. Both conventional detection and enrichment systems have finite dynamic ranges of 3–4 orders of magnitude [18], well below that required (Fig. 7.2). In other words, the highly expressed protein prohibits the application of sensitive enrichment methods like microscale chromatography and capillary electrophoresis [19]. This means substantial *preprocessing* of the sample is required to shrink the dynamic range presented by the sample to a degree manageable with the combined enrichment and detection system [20, 21]. Thus the essential

Table 7.1 Avogadro's Challenge, adapted from Ref. [17]. The minimum number of cells required to provide sufficient protein for visualization by 2D gel electrophoresis and analysis by mass spectrometry. Assuming that no losses occur during protein harvesting and separation, for 10^9 cells and 1000 protein copies, 1.6 pmol protein is present. For 25, 50, and 100 kDa proteins this translates into 4, 8, and 16 ng, respectively. A challenge for proteomics will be to minimize protein losses during purification and separation and to reduce the dynamic range of protein expression presented by a given cell or organism. With permission from Wiley-VCH Verlag.

Number of cells	*Protein copies/ cell*	*Total number of proteins*	*Moles of protein*	*Avogadro's challenge*			*Visualization detection limits*
				ng for 25 kDa protein	*ng for 50 kDa protein*	*ng for 100 kDa protein*	
1.0×10^9	10^6	1×10^{15}	1600 pmol	41,528.00	83,056.00	116,112.00	Coomassie blue
1.0×10^9	10^5	1×10^{14}	160 pmol	4152.00	8304.00	16,608.00	Coomassie blue
1.0×10^9	10^4	1×10^{13}	16 pmol	415.00	830.00	1660.00	Silver staining
1.0×10^9	10^3	1×10^{12}	1.6 pmol	41.00	82.00	164.00	Silver staining
1.0×10^9	10^2	1×10^{11}	160 fmol	4.00	8.00	16.00	Silver staining
1.0×10^9	10	1×10^{10}	16 fmol	0.40	0.80	1.60	Radio
1.00×10^8	10^6	1×10^{14}	160 pmol	4152.00	8304.00	16,608.00	Coomassie blue
1.00×10^8	10^5	1×10^{13}	16 pmol	415.00	830.00	1660.00	Silver staining
1.00×10^8	10^4	1×10^{12}	1.6 pmol	41.00	82.00	164.00	Silver staining
1.00×10^8	10^3	1×10^{11}	160 fmol	4.00	8.00	16.00	Silver staining
1.00×10^8	10^2	1×10^{10}	16 fmol	0.40	0.80	1.60	Radio
1.00×10^8	10	1×10^9	1.6 fmol	0.04	0.08	0.16	Radio
1.00×10^7	10^6	1×10^{13}	16 pmol	415.00	830.00	1660.00	Silver staining
1.00×10^7	10^5	1×10^{12}	1.6 pmol	41.00	82.00	164.00	Silver staining
1.00×10^7	10^4	1×10^{11}	160 fmol	4.00	8.00	16.00	Silver staining
1.00×10^7	10^3	1×10^{10}	16 fmol	0.40	0.80	1.60	Radio
1.00×10^7	10^2	1×10^9	1.6 fmol	0.04	0.08	0.16	Radio
1.00×10^7	10	1×10^8	0.2 fmol	0.004	0.008	0.016	Radio
1.00×10^6	10^6	1×10^{12}	1.6 pmol	41.00	82.00	164.00	Silver staining
1.00×10^6	10^5	1×10^{11}	160 fmol	4.00	8.00	16.00	Silver staining
1.00×10^6	10^4	1×10^{10}	16 fmol	0.40	0.80	1.60	Radio
1.00×10^6	10^3	1×10^9	1.6 fmol	0.04	0.08	0.16	Radio
1.00×10^6	10^2	1×10^8	0.2 fmol	0.004	0.008	0.016	Radio

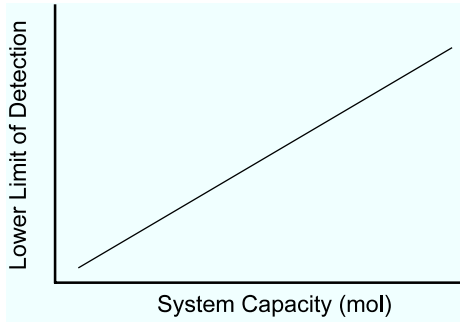


Fig. 7.2 Relationship between the lower limit of detection for a given protein or peptide and the overall capacity of the analytical system. In conventional proteomics limits of detection for any given analyte increase as the overall mass of the sample increases.

effort in proteomics involves reduction of sample complexity before analysis. The third major issue is database quality. Identification of proteins relies very heavily on comparing analytical datasets to lists of protein or DNA sequences. These databases range from high quality annotated genomes (e.g. *E. coli* and yeast) to those which are complex and of comparatively low quality (e.g. human) [22, 23]. Many interesting organisms have no sequence databases. Analytical approaches must be selected with this in mind, as we shall see.

A single chapter is insufficient to portray the full range of research in proteomics and its applications. We will focus on protein identification strategies, presented in a manner that guides the researcher through the many decision-making steps required to conduct fruitful analyses (i.e. maximizing the number of proteins identified in a given sample). The wide variety of analytical options for conducting proteomics will be presented in the context of sample complexity. The concepts presented in this chapter form the basis for creation of a dynamic web tool for managing the multitude of decisions required when conducting such analyses [24].

7.2 Defining the Sample for Proteomics

All biological samples presented for analysis are both highly heterogeneous and compositionally unstable – proteins are both chemically processed and degraded. Proteomic analysis should endeavor to reflect this heterogeneity and instability. The samples presented for proteomic analysis are extremely diverse, which has led to the use of numerous approaches for sample preparation.

Ideally, proteomics should be capable of sampling the heterogeneity present in a biological sample, because heterogeneity is functionally important in biology. For example, understanding the sub-cellular localization of a protein, and how it changes in response to a stimulus, is a key exercise in the determination of protein function [25]. As with any analytical discipline, the sampling strategy is ultimately dictated by the limits of detection and speed of the analytical methodology used. In this regard, modern proteomics has not yet reached a stage where heterogeneity profiling can be performed routinely. Current mass spectrometry systems used for proteome profiling require approximately 10^6–10^7 cells per analysis to provide sufficient protein for identification purposes (Tab. 7.1). Any heterogeneity within this sample will be averaged away. Nevertheless, certain guiding principles can be followed at the sample selection and definition stage to help create informative data sets.

7.2.1 Minimize Cellular Heterogeneity, Avoid Mixed Cell Populations

An important goal in proteomics is the ability to type disease tissues at the molecular level. This is particularly crucial in cancers, where the exact type and stage of a tumor

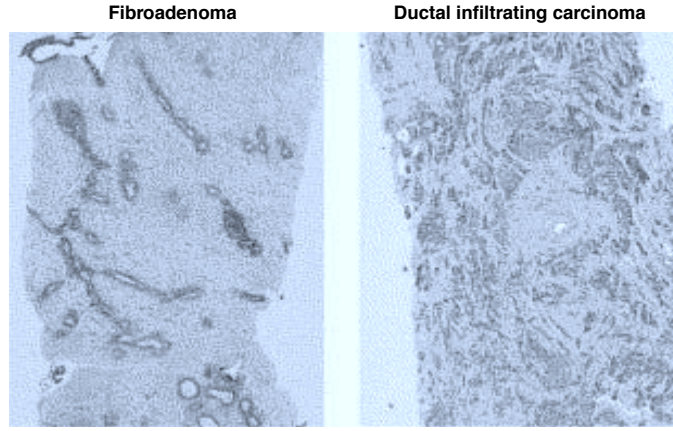


Fig. 7.3 Histological images of representative fibroadenoma and ductal infiltrating carcinoma samples. Each image shows a section of core biopsy. No normal epithelial breast tissue was present in specimens subjected to proteomic analysis [27]. With permission from Elsevier.

strongly affects the treatment selected. A recent study on tissue proteomics from 85 patients with brain cancer involved the careful collection of tumor tissue only [26]. The sampling strategy and subsequent proteomic analysis in this case was sufficient to delineate a set of tumors with a biologically distinct subset of astrocytomas of a particularly aggressive nature. The proteomics results matched nicely with standard histology. In this instance tumor heterogeneity was sufficiently marked to be revealed by the proteomic analysis. The study also shows the value in conducting large numbers of sample analyses to improve the quality of the correlations between cancer type and molecular fingerprint.

In many instances biopsies are cleaner than surgical resections, thus although sample volumes are lower the samples are conceptually ideal for proteomics. This is illustrated in a recent proteomic analysis conducted on breast biopsies [27]. The authors collected multiple core biopsies of breast lesions, added protease inhibitors to prevent degradation of the sample by endogenous enzymes, and froze them immediately. The biopsies were typed by a pathologist by visual assessment of heterogeneity and subsampled for proteomics analysis. This was performed to ensure the absence of normal epithelial breast tissue in the samples. Although the limited sample size forced the authors to focus on higher-expression proteins, they confirmed the presence of key proteins previously linked to breast carcinoma and discriminated between carcinomas and fibroadenomas. Although the differences could be determined by conventional histology (Fig. 7.3), this work involved a rigorous and logical sampling strategy that can be applied in other situations when histological differences are not visible. A particularly promising approach to selective sampling of specific cell types from heterogeneous tissue samples is laser-capture microdissection (LCM), which is already being applied to proteomics [28].

7.2.2 Use Isolated Cell Types and/or Cell Cultures

Tissue samples are not truly homogenous even when presented as careful selections guided by histology and LCM. Such samples would typically contain asynchronous cells with additional variation depending upon their unique microenvironment. This level of heterogeneity will, at the very least,

affect the expression level of many proteins. Cultured cell lines are useful alternatives, because the variability can be controlled to a greater extent. The benefits of cell lines can be illustrated by a proteomic study conducted on human T cells [29]. In this work, the authors carefully cultured and stimulated CD4+ lymphocytes, and then selectively polarized fractions of the culture toward T1 helper and T2 helper cells using the required interleukins. A functional assessment of polarization was made before conducting a proteomic comparison of expression levels between the differentially stimulated fractions. This level of control over growth conditions and stimulation promoted informative comparisons and helped identify differentially expressed proteins with a minimum of analytical variance. In general, cell and tissue cultures are excellent entry-points into proteomics experiments, because sufficiently large samples can be prepared under controlled conditions. One caution is that evolution and adaptation of the line through numerous passages can alter the significance of the findings. Whole cell proteomics have been particularly successful in the analysis of lower complexity organisms such as bacteria and yeast [30–32]. Cultures of these organisms can be prepared in high abundance, with greater ease and compositional homogeneity. The small size of their proteomes make whole-cell proteomic initiatives feasible – for example recent studies suggest that approximately 5000 proteins are expressed at any given time in *S. cerevisiae* [33, 34].

7.2.3 Minimize Intracellular Heterogeneity

Virtually all tissue and whole-cell proteomics experiments of the types described above reflect higher expression-level proteins in the sample. None of the analytical techniques used in proteomics have the dynamic range required for detection of the lower range of protein concentration nor do they offer empirical evidence of protein localization. Substantial recent activity in proteomics profiling has focused on the analysis of *subcellular* fractions, in an effort to minimize the dynamic range problem and recover some measure of location specificity. Not surprisingly, integral membrane and membrane-associated proteins have been studied extensively, because of their significance in cell signaling and their high value as drug targets [35, 36]. Organellar proteomics of mitochondria [37, 38], the nucleolus [39, 40], and the phagosome [41, 42] are examples of some recent initiatives that have led to new knowledge. A key component of all these projects was the implementation of fractionation procedures specific for the targeted organelle [17]. Some of the work on the analysis of the plasma membrane has revealed proteins normally associated with the mitochondria, thus raising the question of the validity of the association. Prefractionation procedures are not always perfect but elegant examples of subtractive techniques such those as described in the work on the proteomics of the nuclear membrane [35] suggest ways around current difficulties. The proteomic analysis of lipid rafts, or membrane microdomains, is more controversial because of the lack of consensus on an enrichment procedure [43]. In general, a proper proteomics strategy involves a validation step, in which the proteins identified in the discovery phase are localized more precisely, for example by immunostaining and microscopy.

7.2.4 Minimize Dynamic Range

Sub-cellular fractionation helps to minimize the complexity of the sample intended

for proteomic identification and has the additional benefit of reducing the dynamic range of protein concentration. Conventional chromatographic or electrophoretic methods of protein prefractionation have been used to reduce the dynamic range further and, whether this is applied to the proteins themselves or their enzymatic digests, results clearly show that prefractionation minimizes the dynamic range in each fraction presented for analysis [44]. Because these separation methods are based on bulk properties such as hydrophobicity, charge and size they tend not to discriminate between high-concentration and low-concentration proteins. Thus, the dynamic range within fractions is usually, but not always, compressed. For example, low-abundance proteins with retention properties similar to those of serum albumin will occur in a fraction with a wide dynamic range. For this reason multidimensional separation techniques such as 2D gel electrophoresis and hyphenation of different types of chromatography pervade modern proteomics. An extra dimension of separation can further reduce the dynamic range issue.

The problem is significant for unstructured biological fluids such as serum. Recent efforts have focused on extraction of the main, high abundance proteins by affinity chromatography (for albumin, antitrypsin, hemopexin, etc.) thus enriching the remaining solution for lower-abundance proteins [45–47]. This minimizes the downstream sample handling required to identify the proteins. Affinity-based procedures are also used for selective targeting of proteins within complex samples. Protein chips or arrays based on selective recognition of predetermined proteins are useful fr quantitation from such samples [48]. These can be regarded as high-throughput variants of the ubiquitous immunoassay. Immunoprecipitation of a given protein and its associated proteins is another means of circumventing the dynamic range issue and conveying functional relevance at the same time. Large-scale interaction maps have been assembled by using this sample handling procedure [6, 49, 50]. More targeted efforts show that the methods are reasonably reproducible but interlaboratory variance is still an issue, as are nonspecific associations [7].

7.2.5
Maximize Concentration/Minimize Handling

Detection systems used in proteomics are "concentration-dependent" rather than mass-dependent. Sample fractionation techniques should therefore preserve – or better – increase the concentration of the recovered proteins. A final step before identification is usually a preconcentration step, often based on chromatography or electrophoresis. In short, all the considerations discussed here indicate that substantial sample processing is required to deliver high identification power. Proteins are notoriously unstable in solution outside their biological environment and thus multistep processes tend to compound problems with sample loss as a result of denaturing, precipitation and adsorption.

7.3
New Developments – Clinical Proteomics

There are new developments in proteomics that directly affect the sampling strategy problem and serve to highlight the technical challenges in the field. Researchers are investigating the feasibility of direct sampling of tissue using mass spectrometry, so-called "imaging mass spectrometry" [51]. The current state of the art supports the

mapping of hundreds of peptides and proteins from spot sizes of approximately 50 μm, with the potential of low-micron spatial resolution [52]. Not all of the proteins and peptides present in the spot are represented in the resulting data sets – there is a bias toward soluble, low-molecular-weight protein – but the data suggest that the spectral patterns generated correlate well with histological categorization. Key challenges that lie ahead include improving the resolution and minimizing this bias in detection but for now the mass spectral patterns generated seem to provide imaging capability.

The concept of proteomic patterns has also appeared in large-scale biomarker discovery projects. Biomarkers are molecular "signposts" that are tracked in a clinical environment to predict disease onset, progression, and response to treatment [53]. In the discovery of such biomarkers, large numbers of clinical samples (e.g. tissue, serum, urine) require profiling for their protein content using the tools of proteomics. Numerous groups have applied the concept of *spectral patterns* of clinical protein samples as surrogates for rigorous compositional analysis [15, 54]. The approach has generated intense interest in the clinical sciences, although the concept of patterns rather than protein(s) has some inherent weaknesses [53].

Both of these new applications emphasize that sample processing in modern proteomics is very labor intensive and that simplifications designed to increase throughput come at the expense of detectable protein dynamic range and identity. In an attempt to overcome the averaging effect in the analysis of large samples, others have turned to single-cell proteomics. Although the standard tools of proteomics (e.g. mass spectrometers) currently lack the detection limits needed to identify the proteins in samples this small, multidimensional capillary electrophoresis with laser-induced fluorescence detection shows promise for fingerprinting the protein complement of single cells [55, 56].

7.4 Mass Spectrometry – The Essential Proteomic Technology

When the sample has been defined along the lines described above, it requires matching with the appropriate analytical methods for protein identification. Before discussing the range of options available we must introduce the engine of modern discovery proteomics – mass spectrometry. Mass spectrometry is a technique used to determine the mass-to-charge ratio (m/z) of gas-phase ions on the basis of their behavior in a mass analyzer.

The revolution in mass spectrometry as it affects biology arose from the development of two combined sample-introduction/ionization methods. One of the key hurdles to overcome was the volatilization of large, labile biomolecules. Until the late 1980s there were no reliable sample-introduction methods for proteins and peptides. At this time two new methods for simultaneous volatilization and ionization appeared – matrix-assisted laser desorption/ionization (MALDI) [57] and electrospray ionization (ESI) [58]. MALDI is a pulsed-ionization technique in which proteins or peptides are dissolved and mixed with an energy-absorbing organic compound. The resulting solution is spotted on to a conductive target and dried, leaving a deposit of sample cocrystallized with the energy-absorbing matrix (Fig. 7.4). Conventional MALDI experiments use pulsed UV lasers, most commonly the simple nitrogen laser, which emits at 337 nm for approximately 3 ns per pulse. In a process still not completely

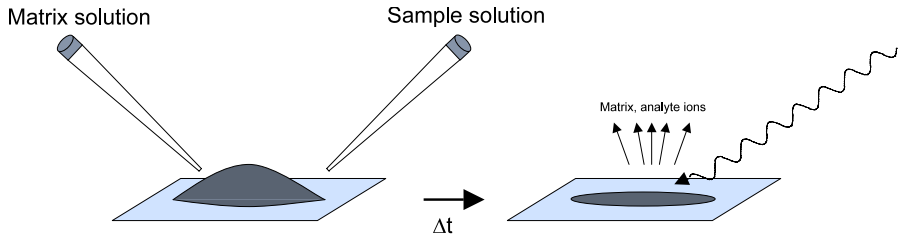


Fig. 7.4 Schematic diagram of the MALDI process. A solid sample of organic matrix material and soluble biological sample is cast from solution on to a plate and irradiated with a pulsed laser beam to generate an ablation plume of ions that is sampled by the mass spectrometer.

understood, irradiation of the cocrystallized deposit leads to an ablative event in which the entrained sample molecules are carried into the gas phase and ionized, the latter probably by gas-phase proton transfer from the excited state of the matrix to the protein or peptide [59]. An energy transfer "bottleneck" exists such that the labile sample molecules are generally not dissociated in a conventional MALDI experiment. Because MALDI is a pulsed ion generation technique, it is most often configured with a mass spectrometer designed to detect ion pulses, namely a time-of-flight (TOF) mass spectrometer. MALDI–TOF has emerged as the workhorse instrument for protein and peptide detection, with a theoretically unlimited MW detection range with current limits of detection down to sub-amol (10^{-18} mol) [60]. The method has only modest tolerance of common buffers and solution components found in biological samples and thus requires some form of "clean-up" before analysis. Peptide and protein ions that are generated by this technique are usually singly charged by adduction with a proton, although very large proteins can acquire a small charge-state distribution.

Electrospray ionization is a continuous ion-generation technique, as opposed to the pulsed nature of MALDI. In this process, gaseous ionized molecules are emitted from a conductive solution sprayed through a strong electric field (Fig. 7.5). Again, as with MALDI, the mechanism for ion generation is still incompletely understood, although key processes include the production of charged droplets, coulombic explosion of desolvating droplets, and redox chemistry [61]. Solvated ions are generated by the technique and are sampled by the mass spectrometer. Given that a continuous stream of solvated ions is produced, scanning-type mass spectrometers are a natural fit. Quadrupole instruments were the earliest successful designs for peptide and protein detection. These instruments have a limited m/z range (<3000) but the large charge-state distributions generated by the electrospray process enables analysis of high-molecular-weight proteins even with such a platform. It is not uncommon for a protein with a molecular weight of 50,000 to carry 10–50 protons, thus bringing its m/z at least partially within the detectable range of the quadrupole instrument. The electrospray process is even more sensitive to contaminants than the MALDI technique, and thus a common instrumental configuration involves liquid chromatography (LC) interfaced with an electrospray mass spectrometer. As will be shown in later sections, these classical

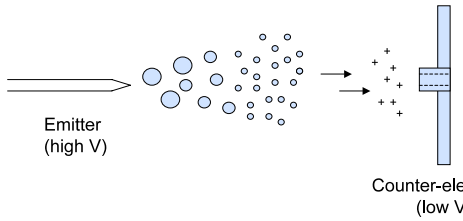


Fig. 7.5 Schematic diagram of the electrospray process. Solvated ions are emitted from ever-shrinking droplets, and sampled by the mass spectrometer.

"blends" of ionization technique and mass spectrometer have been substantially reworked, leading to a diverse collection of hyphenated tools for protein analysis best described in relation to the intended applications.

Mass spectrometry would be of relatively minor importance in protein analysis were it only applied to the detection of intact proteins in a mixture. Gel electrophoresis can be used to measure the molecular weight of proteins, although not with the same accuracy. It is more tolerant of impurities, usually has lower limits of detection, and can be interfaced with immunoaffinity procedures for characterization and validation. Mass spectrometry is, however, best viewed as a detection system for *direct identification* of proteins, and is most efficiently utilized when merged with protein separation technologies such as gel electrophoresis (one can consider MS to be conceptually like a protein gel stain, albeit with far greater analytical power). The identification arises as a result of compositional analysis of the protein (Fig. 7.6). The breakthrough that led to the application of mass spectrometry to protein discovery centers on shifting a protein to the left of the graph. This has been achieved by a combination of *sample processing, instrument design,* and the generation of *sequence databases*.

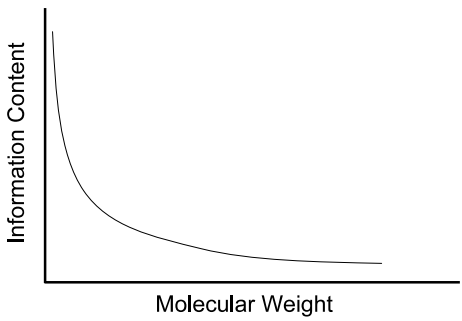


Fig. 7.6 Generalized plot of compositional information as a function of molecular weight. Compositional information diminishes rapidly with increasing molecular weight of the molecule being analyzed. A fundamental limit exists where the number of possible combinations of elements prevents identification based on molecular weight determination.

7.4.1 Sample Processing

The insufficiency of protein moleclar weight is circumvented by controlled degradation of the protein into lower molecular weight peptide fragments. These fragments are then analyzed by mass spectrometry. Controlling the mechanism of degradation is very important, to limit the number of peptide masses and to "add" structural information. The most common proteomic process for the generation of peptides is enzymatic hydrolysis with the endoprotease trypsin. Trypsin is highly specific for cleavage C-terminal to lysine and arginine, thus all peptides generated through tryptic digestion can be expected to contain a C-terminal lysine or arginine (except for the C-terminus of the protein). This extra compositional information is useful in the downstream identification process. In tryptic digestion, K and R-terminated peptides also support

Table 7.2 List of protein-degradation methods commonly used in proteomics.

Hydrolysis option	*Specificity*
Trypsin	C-terminal to K/R
Lys-C	C-terminal to K
Arg-C	C-terminal to R
Asp-N	C-terminal to D
CNBr	C-terminal to M
Glu-C	C-terminal to D/E
Chymotrypsin	C-terminal to F/Y/W/L
Pepsin	C-terminal to F/L

sensitive mass spectrometric detection, by providing readily protonated basic residues [62]. Other enzymes can be used, and chemical dissociation processes; common options are listed in Tab. 7.2. When fully dissociated, the association between peptides and the protein of origin is lost. In other words, MS analysis of the peptides does not determine whether peptides arise from the same protein or from a mixture of proteins. To preserve this association and maximize compositional information, complete isolation of the protein must be achieved before digestion and analysis.

7.4.2 Instrumentation

The peptides generated as described above are assessed by mass spectrometry to determine molecular weight. Accurate mass measurement of peptides requires sufficient resolution to separate the isotopic envelope within a peptide and to discriminate peptides from each other. The isotopic content of peptides is fairly straightforward to predict, thus selection of the monoisotopic peak (i.e. all common isotopes) can be quickly ascertained. This process is somewhat more challenging for electrospray-generated mass spectra than for MALDI spectra, because the multiple charging of peptides leads to m/z values that compress the isotopic envelope and increase the need for high resolution (Fig. 7.7). The mixture of peptides that results from a protein digest can lead to peptides with almost equivalent molecular weights, with the probability of this occurrence increasing if digests of protein mixtures are analyzed directly. In these situations a higher resolution instrument might be needed or the peptide mixture must be prefractionated by chromatography to minimize the possibility of peak overlap. Some popular mass spectrometers used in peptide analysis are listed in Tab. 7.3.

One of the key strengths of modern mass spectrometry is the ability to select ions with specific m/z values and fragment these ions in a compositionally informative manner. We usually describe tandem mass spectrometers as "tandem in space" and "tandem in time" (Fig. 7.8). Tandem-in-space instruments are serial mass analyzers – for example the triple quadrupole. Ions are selected for dissociation in one mass analyzer, fragmented, and the fragmentation products mass analyzed in a separate analyzer. The mass analyzers need not be of the same type, and many different hybrids have been designed for different purposes. For example, the triple quadrupole is an ideal instrument for high-sensitivity monitoring of known compounds (e.g., drugs in biological fluids) whereas the quadrupole time-of-flight is more appropriate for identifying and characterizing unknown compounds.

In tandem-in-time mass spectrometers, typically ion-trapping instruments, ion selection and dissociation are conducted in the same mass analyzer. All ions but those of a specific m/z are ejected from the trap and thereafter the selected ions are dissociated. The resulting fragments remain trapped until scanned out of the analyzer to

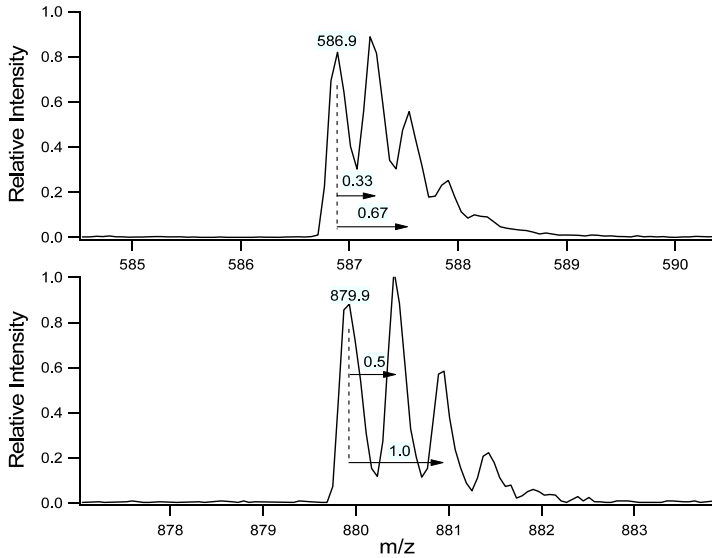


Fig. 7.7 Expanded region of the electrospray-generated spectrum of a peptide standard (renin substrate, monoisotopic mass of 1758.9 u) showing isotopic resolution for the doubly and triply charged ions.

Table 7.3 The mass spectrometers commonly used in proteomics experiments and their figures of merit. Note that resolution and mass accuracy values are not the maximum attainable but rather operating values commonly used during standard proteomics experiments.

Mass spectrometer	*Resolution ($m/\Delta m$)*	*Mass accuracy (ppm)*	*Stages of MS*
Ion trap	5000	150	<12
Reflectron time-of-flight (TOF)	15,000	20	1
Triple quadrupole	2000	150	2
Fourier transform ion cyclotron resonance	100,000	1	<12
Quadrupole TOF	15,000	10	2
TOF–TOF	10,000	50	2

generate the tandem mass spectrum. In both types of tandem instrument ion fragmentation is achieved by a collisionally induced dissociation (CID) mechanism. In CID the ions selected for fragmentation are energized by inelastic collisions with neutral gas molecules. Multiple collisions lead to sufficient energy deposition and unimolecular ion dissociation processes lead to the fragmentation patterns observed. Fragmentation of the peptides produces informative MS–MS spectra containing sequence and compositional information, although these are not always readily interpretable. Current automated sequencing algorithms return error-prone partial sequence information, frequently requiring manual oversight for validation purposes.

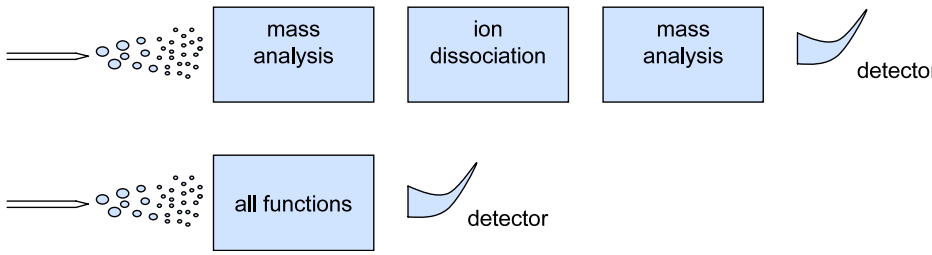


Fig. 7.8 Tandem-in-space (top) and tandem-in-time (bottom) mass spectrometers for obtaining product ion spectra. For tandem-in-space, ion selection is segregated from ion dissociation whereas for tandem-in-time, all functions are performed in the same volume element, separated by time.

The sensibly named "ion trap" mass spectrometer remains the workhorse instrument for proteomics, partly because of its lower cost and low limits of detection. Configured for electrospray ionization, bench-top ion traps interface readily with chromatographic equipment for peptide enrichment. A chromatographically separated peptide can be detected and sequenced in under 3 s with such an instrument, equivalent to a sequencing rate of up to 1200 peptides h^{-1}. This is rarely achieved in practice because the peptides arising from protein digestion cannot be perfectly separated chromatographically [63]. Nevertheless, this is a remarkable improvement in peptide sequencing technology compared with the Edman-based sequencing instrumentation previously used (~1 peptide day^{-1}).

7.4.3 MS Bioinformatics/Sequence Databases

The information content of a peptide mass spectrum is increased if the list of possible proteins is constrained. Intuitively, there is a greater chance of identifying a molecule from an accurate mass measurement if all the possible molecules are known and random compositions are not permitted. Even though the proteome is large, the list of proteins that occur in nature represents but a small sub-set of all theoretically possible protein sequences. The first successful application of mass spectrometry to protein identification involved comparison of mass lists of peptides from a protein digest with computer-generated digests of all known proteins in existing databases [64–66]. This is referred to as a *peptide mass fingerprint*, in which the bioinformatics exercise involves identifying a theoretical protein digest that provides a best-fit with the dataset. Clearly the specificity of the chosen enzyme or reagent for protein hydrolysis plays a large role in constraining the size and scope of the searching exercise. Intensity data are not used in these fingerprint analyses.

In the MOWSE-based approach – the initial algorithm behind the popular Mascot database searching tools – once a list of peptide masses is generated, each measured value is weighted according to its frequency of occurrence within the database [67]. A score is assigned to each protein in the database based on the number of matched peptides. In this basic method the absolute value of the scores has little meaning. It is the separation of the score from that of random hits that bears significance. This has led to couching the score in probabilistic terms, by generating probability-based val-

ues indicating the likelihood of a determination arising as a result of a random event [68]. By assigning probability to a hit, the user can determine the confidence level that should be applied. There are numerous variants of this approach, but all such fingerprinting methods share the following stringent requirements to ensure a high success rate:

High database quality (complete protein list, low error rate, full knowledge of post-translational modifications). Most databases are populated with translated genomic data of different levels of maintenance and quality. Because these are in a constant state of improvement, the score reflects the quality of the database as much as it reflects the quality of the dataset. It is not surprising that the peptide mass fingerprinting approach is best suited to bacterial studies in which the genome is small and complete [69]. Identification success rates approach 100%. Searches based on human databases, however, deliver significantly lower success rates.

High mass-measurement accuracy and sequence coverage for the protein digest. There is a strong correlation between mass accuracy and the ability to determine an identity with high confidence [70]. Fig. 7.9 indicates the importance of mass accuracy in this context, thus instruments with the ability to generate <20 ppm levels are strongly recommended.

Isolated protein. The presence of two or more proteins will lead to a list of peptides with association to the protein of origin erased. Although it is possible to search for the best x proteins in the database to explain the data, the success rate of the search drops. Proteins should be well separated before such experiments to improve the success of the search.

The peptide mass fingerprinting approach is biased against small proteins, and although correction factors can be applied to avoid overweighting large proteins there is still a minimum requirement for the length of the mass list (approximately five peptides or more). Many small proteins will not, on digestion, produce a list of this size.

Tandem MS data sets are richer in that peptide sequence can be empirically determined. Manual or computer-assisted inter-

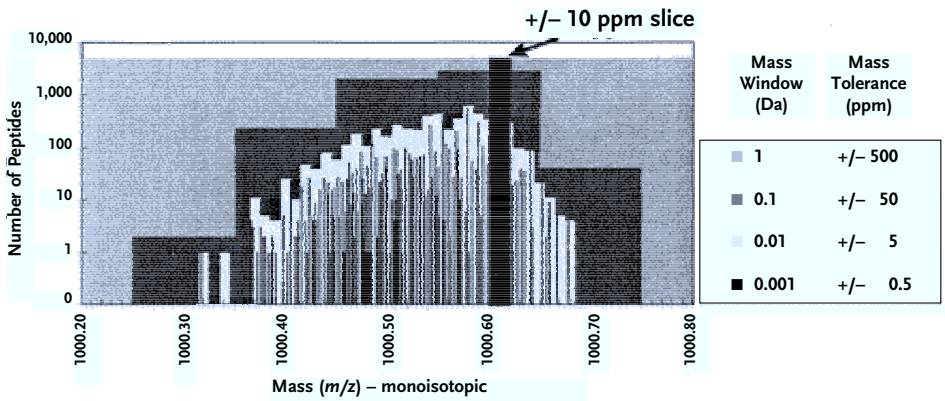


Fig. 7.9 The number of tryptic peptides of nominal mass 1000 present in the Genpept database release 98 (12/96), which contains ~210,000 entries. To facilitate assessment of the discriminating power of mass accuracy, the number of peptides was counted in mass windows of four different widths. One missed tryptic cleavage was allowed (from Ref. [70]). With permission from American Chemical Society.

pretation of product-ion spectra from tandem MS experiments conducted on peptides provides short sections of sequence that can be used in a conventional Blast search [71]. This rather laborious *de-novo* procedure still remains the most reliable approach for identifying new proteins, particularly when forced to interrogate incompletely annotated genomic databases and/or EST collections. It is also necessary when mutations, alternative splicing, and post-translational modifications are encountered. New tools are emerging to assist in the *de-novo* sequencing of peptides (e.g. Lutefisk, Sherenga), but the inherent ambiguities and low signal-to-noise ratios of typical product-ion spectra will hamper the complete success of such ventures.

When databases are of sufficient quality, a correlation approach not unlike peptide mass fingerprinting can be applied [71]. Uninterpreted product ion spectra are compared with expected product ion spectra for peptides generated from all the database entries. The size and complexity of such a set of peptides depends on the constraints supplied by the user, and includes as a partial list the number of missed cleavage points, post-translational modifications, and fragment types. Comparisons can be at the level of simple matching of mass lists (MS-Tag), or more complex numerical cross-correlation of filtered empirical data sets with theoretical spectra that preserve some measure of intensity information (SeQuest). Whatever the method, the database-searching strategy is better than the *de novo* approach from the standpoint of speed, but it must be used cautiously. The appropriateness of the selected database must be carefully evaluated, as well as the quality of the product-ion spectra. The methods are designed to return a hit regardless of accuracy, thus the top-ranked peptide does not imply identity. Validation of the result by manual inspection of the data or an independent experiment is always advisable. Other peptides from the same protein will usually be present, and their product-ion spectra might help confirm identity. Search tools such as Mascot can assemble all such data into a combined probability-based score for enhanced identification power [68].

The concept of identity is used rather loosely in MS-driven sequencing. Given that determinations might be based on as little as a single peptide from a protein, identity is achieved at a basic level only. A single peptide is usually insufficient to discriminate between isoforms, splice variants and variable states of post-translational modification. Such identifications are best viewed as entry points into additional experimentation that might include further sequencing efforts and/or biochemical experimentation (generating antibodies for Western blots, for example).

7.5 Sample-driven Proteomics Processes

The growing array of instruments and hyphenated analytical techniques in proteomics poses difficulties for the uninitiated to decide on an appropriate course of action when embarking on a protein-discovery project. The following sections are designed to offer brief descriptions of the key discovery approaches, in a manner that helps the prospective user successfully undertake a proteomics project. Fig. 7.10 outlines the popular approaches and attempts to quantitate both the complexity of the sample for which the methods are appropriate, and the effort involved. Naturally, treatment such as this is somewhat subjective and based on personal experience, and so is best regarded as a guide. Each description is accompanied

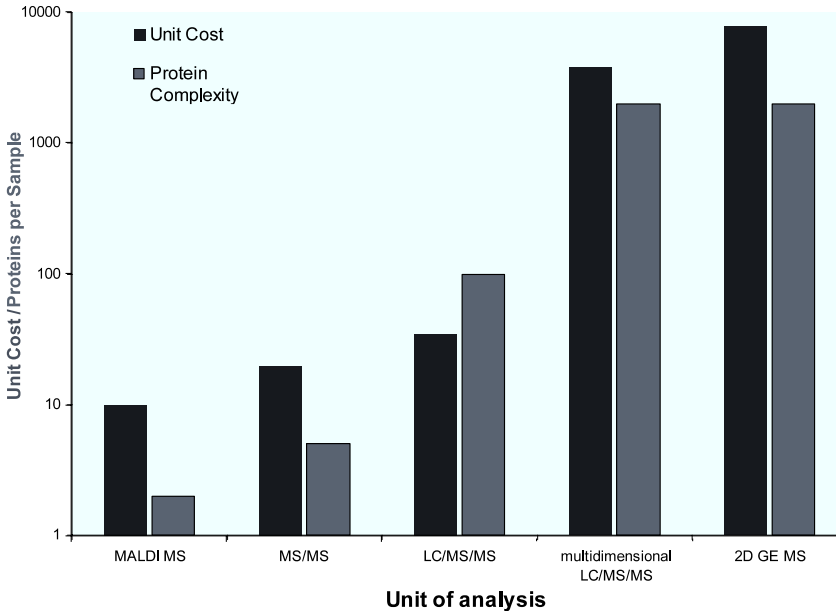


Fig. 7.10 Relative quantitation of the costs associated with key proteomics methodology and the return evaluated in the number of proteins identified. The cost in arbitrary units reflects reagents, consumables, and analysis time.

by a discussion of strengths, weaknesses, applications, and new developments.

7.5.1 Direct MS Analysis of a Protein Digest

Direct processing of a protein preparation with an enzyme such as trypsin or Lys-C generates a pool of peptides that can be sampled directly for MALDI–TOF analysis, without fractionation of the digest. This process is the simplest and quickest of all proteomics experiments and, with the exception of the digestion step, analysis can be completed in minutes. It delivers a peptide mass fingerprint that can be very quickly searched against available databases. The principle weakness relates to its inability to process mixtures of proteins. As mentioned above, the database-searching routine works best when all detected peptides arise from the same protein. The MS spectrum arising from a digest of mixed proteins will also suffer from *peak suppression* and thus the resulting spectrum does not reflect all the peptides present in the sample. In general the method is very sensitive to spectral intensity (Fig. 7.11) thus any procedure for enriching the protein and/or its digest usually improves the score.

The classical use of this approach is in the analysis of gel-separated proteins, principally by 2D gel electrophoresis. The gel-isolated proteins are fixed/precipitated before excision from the rest of the gel. An in-gel digestion procedure is used to dissociate the protein and at this stage the peptides can be readily extracted for mass analysis [72–74]. This procedure is quite successful – the gel matrix provides the opportunity to wash and treat the protein without sample loss and the gel electrophoresis step pro-

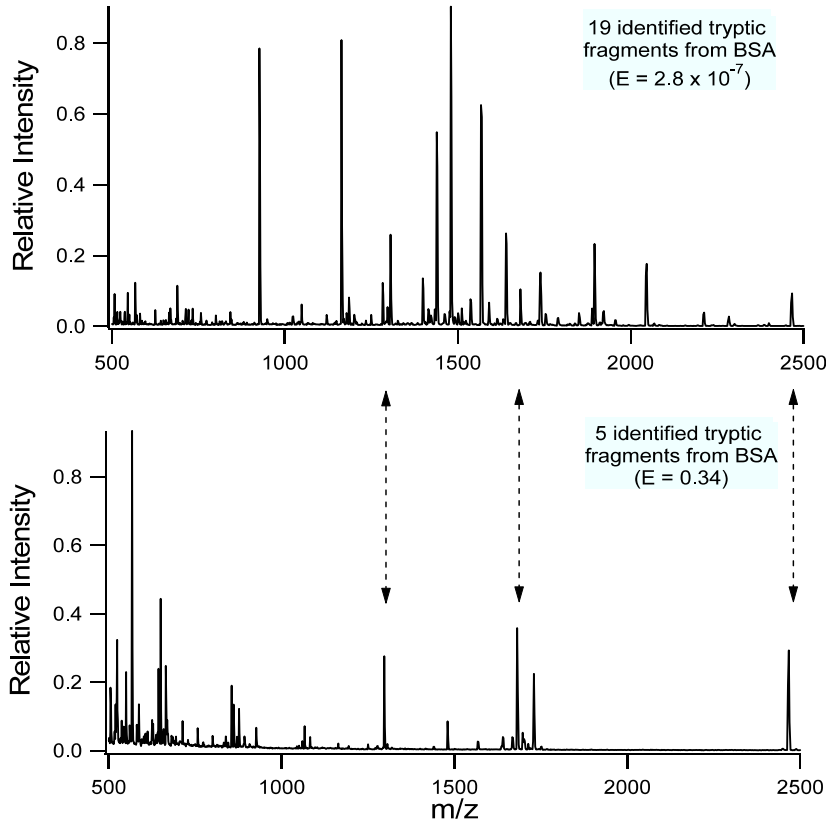


Fig. 7.11 Peptide mass map, with associated database search results, arising from in-gel digestion of bovine serum albumin (BSA) at two load levels, 2.5 pmol (A) and 0.5 pmol (B). BSA was reduced pre-gel and modified by residual acrylamide in the gel. Dashed arrows mark the internal calibrants (doped at 20 fmol in each analysis).

vides protein enrichment suitable for efficient enzymatic processing. The process can be applied to gel spots visualized by numerous methods, including Coomassie, silver, and sypro-ruby stains. The silver staining method must be conducted very carefully to obtain good results. An MS-friendly version has been developed [75]; nevertheless overexposure of the protein to the stain can generate an intractable spot. Most service labs continue to recommend avoiding silver staining altogether. The 2D gel electrophoresis step with subsequent spot processing before MS analysis is labor intensive. Robotic systems for imaging, spot picking, digesting and MALDI spotting have been developed to lighten the load. Frequently, there is expectation that a visualized gel spot will lead to a detectable protein digest. While this is usually true with the universal protein stains, it is not true for specifically labeled proteins. Gels visualized by autoradiography or Western blotting can reflect lower limits of detection that this basic MS method simply cannot achieve.

As indicated in Fig. 7.9, high measurement accuracy increases the information

content of mass determinations. Although MALDI–TOF instruments are limited to ~20 ppm, obtaining peptide mass fingerprints by using some hybrid instruments enables tolerances as tight as 2 ppm [76]. This high mass-measurement accuracy occasionally enables unambiguous identification from a single peptide [77]. This, in turn, implies that protein identities from digests of more complex protein mixtures can be accommodated. In general, the excellent S/N ratio achievable with modern FT-ICR also improves the sequence coverage obtained from digests [78, 79]. Tryptic digestion of proteins in solution can now be achieved rapidly, at concentrations below conventional solution digests. Reactor-based digestion in the presence of accelerants (acetonitrile, methanol) has enabled complete digestion in seconds of proteins at low nanomolar concentrations [80]. The in-gel tryptic digestion procedure has also been accelerated to approximately 30 min [74].

7.5.2 Direct MS–MS Analysis of a Digest

In this mode, peptides are mass selected from a sample digest in the first stage of a tandem mass spectrometer. The most common form of analysis involves electrospraying the digest at low flow-rates (sub-microliter min^{-1}), the so-called static nanospray experiment. One or two microliters of sample can usually sustain a 15–30-min experiment and at a 3-s cycle time per product for ion acquisition, a large number of peptides can be sequenced in such an experiment. Samples can be preconcentrated before analysis to maximize S/N, but the technique does not fractionate the digest. This is an excellent method for confirming protein identities obtained by MALDI–TOF acquisition, although it requires a separate instrument and method. The MALDI–TOF experiment rarely uses the entire digest, thus sufficient sample is almost always available for such an experiment.

Because protein identification is possible on the basis of a single quality product-ion spectrum, such an analysis can process a digest from an unfractionated mixture of proteins. As a general tool for the biological researcher, MS–MS analyses such as these are useful for extensive sequencing of an isolated protein. It remains one of the best methods for discovering post-translational modifications, for two reasons. The MS–MS mode generates sequence information and thus supports the identification of modified amino acids from the product ion spectrum. Continuously infusing the digest in the nanospray experiment results in sufficient time to scan all detected peptides for certain modifications. In recent work by Mann et al. a hybrid quad-TOF instrument in product-ion scan mode was used to detect all phosphotyrosine-containing peptides [81] (Fig. 7.12). This particular experiment takes several minutes to complete, thus requiring the continuous infusion mode. Such modifications are usually low abundance and not straightforward to detect from a simple MS scan, therefore the combination of high sensitivity nanospray with long integration times is an excellent method for these analyses.

Recent developments in tandem mass spectrometry have improved our ability to extract useful sequence information from MS–MS experiments and streamline the analytical effort. The MALDI process can now be successfully implemented on MS–MS machines [82–84]. Therefore, one can generate product ion spectra from peptides selected after an initial peptide mass fingerprinting experiment. Typical electrospray-based tandem mass spectrometers select peptide ions for sequencing on the

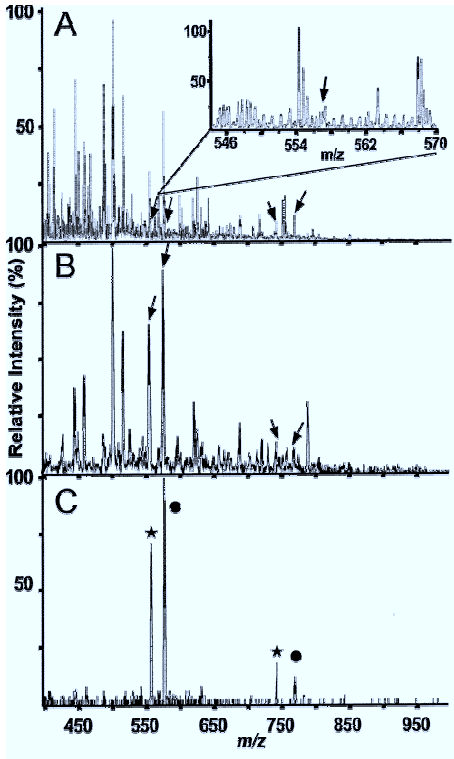


Fig. 7.12 Analysis of a mixture of four proteins – single-strand DNA binding protein (*E. coli*), human transferrin, His-tagged RrmA (*E. coli*), and human MAP-kinase 2. (A) Nanoelectrospray quadrupole-TOF mass spectrum of the unseparated peptide mixture. Arrows indicate the position of the expected triply and quadruply charged phosphopeptide signals. The insert shows an expanded view of the quadruply charged phosphopeptide at m/z 556.7 (marked by an arrow). (B) Precursor ion scan for 216.1 (± 0.3 Da). This spectrum is comparable with that from a similar experiment on a well-tuned triple-quadrupole mass spectrometer. Apart from the phosphotyrosine-containing peptides (marked with arrows), signals corresponding to peptides giving rise to interfering a- and b-type fragment ions are observed. Thus, no substantial claim regarding the presence of phosphopeptides in this mixture can be made. (C) Precursor ion scan for *m/z* 216.04 (± 0.02 Da). All signals that are observed correspond to the triply and quadruply charged phosphopeptide derived from the MAPK in its doubly (stars) or singly (circles) phosphorylated state [81]. With permission from John Wiley & Sons.

basis of intensity, and while this generates the best quality MS–MS data, it might ignore lower-intensity peptide ions that would have a high information content. MALDI MS–MS enables greater operator control of the list of peptides chosen for sequencing, by enabling a database search with the peptide masses before ion selection. MALDI can now be successfully implemented on ion traps, quad-TOF instruments and TOF–TOF instruments, although the sequence data are not always of high quality. MALDI generates singly charged peptides and these tend not to fragment as efficiently and informatively as doubly charged ions produced by electrospray [85]. Labeling chemistries have been developed to improve the quality of MALDI MS–MS data, which should prove valuable [86].

Chemical noise can frequently hide the peptide ion signal, thus preventing the critical first step in peptide sequencing, i.e. peptide selection. Electrospray-based direct MS–MS experiments have recently been adapted in some instruments to enhance the signal of multiply-charged ions [86]. Because the chemical noise tends to be largely singly charged in nature, peptide ion signals can be enhanced by suppressing singly-charged ion signals. This results in improved capability to select ions for MS–MS, thus maximizing the sequence information extracted from a single nanospray experiment.

7.5.3
LC–MS–MS of a Protein Digest

Combining the tandem MS experiment with chromatographic separation of protein digests provides the opportunity to achieve the lowest limits-of-detection and efficiently use the available sequencing capacity of the chosen mass spectrometer. The most com-

mon configuration involves small-bore reversed phase LC columns with electrospray-based ion-trap or quad-TOF mass spectrometers. To a first approximation, peak concentration scales with the inverse of column diameter. Columns with i.d. as low as 30 µm are being used, although 75 µm is more common. Such systems require flow rates in the sub-µL min^{-1} range and a nanospray configuration. This combination is not the most rugged of systems and currently requires considerable operator intervention. The key benefits of these combinations include enrichment and the ability to sequence hundreds of peptides per run, translating into a capacity for low hundreds of protein identifications per LC–MS–MS experiment. Licklider et al. have used the approach to identify proteins from an outer membrane preparation from *Pseudomonas aeruginosa* [87]. A single LC–MS–MS run was able to identify more than one hundred proteins. Such capability minimizes the need for exhaustive protein separation before analysis – the key strength of the approach. It achieves a substantial increase in identification power over the direct MS–MS approach by providing the mass spectrometer additional time for sequencing. These experiments are conducted in a data-dependent fashion; a survey MS scan is first used to identify a peptide ion m/z and this is followed by an MS–MS experiment on this ion. By concentrating a given peptide in a band several seconds wide and separating it from others, the instrument has sufficient time to generate a quality MS–MS spectrum. Thus the efficiency of the chromatography used is determined by the MS–MS requirements of the mass spectrometer. Coeluting peptides can be sequenced only if the chromatographic bandwidth is sufficiently broad to enable for more than one sequencing event per cycle. Ultrahigh resolution separation techniques with bandwidths in the low to sub-second timeframe currently cannot be supported by MS–MS instrumentation.

The range of applications of such a system is quite enormous, although they center principally on protein identification [88]. Such analyses are not exhaustive and, because of the finite sequencing capacity mentioned above, additional runs of a given sample should be performed. New developments in this area center on maximizing the peak capacity of LC–MS–MS systems and improving their ruggedness. Ultra-long chromatographic gradients and long columns for highest separation efficiency have recently been applied, for example to the analysis of *D. radiodurans* [89, 90]. Runs as long as five hours achieve peak capacities greater than 1000 at flow rates that support very sensitive detection. New monolithic columns have been developed that can achieve high efficiency separations at flow rates higher than with conventional columns; this should lead to shorter analysis times [91]. Better quality low-flow pumps have been developed to support reproducible chromatography; this has led to the concept of using predicted chromatographic retention times as an aid in the determination of peptide sequence [92]. With the advent of MALDI MS–MS, systems are being developed that couple microLC with MALDI spotters [93]. Because MS–MS can be implemented without the time constraints of electrospray LC–MS–MS, the hyphenated technique of LC-MALDI MS–MS should prove quite useful. In general, attention to high-performance LC–MS–MS system development for peptide separation and analysis affords the best option for improving the dynamic range of analysis. Recent work by Smith et al. suggests that FT-ICR-based systems can deliver up to six orders of magnitude [89, 90].

7.5.4 Multidimensional LC–MS–MS of a Digest (Top-down vs Bottom-up Proteomics)

Two dimensional gel electrophoresis provides the opportunity to separate large numbers of proteins, and their intercalation within the gel matrix enables sample treatment and cleanup before protein extraction. Gel-based approaches for biomolecule separation are but a subset of the available fractionation techniques that can be used to process protein samples before to mass spectrometric analysis. A range of chromatographic methods have been applied both to intact protein separation and to separation of protein digests. Compared with gel electrophoresis these methods enable more selective retention behavior and ease of automation. Protein separation by chromatographic means can be achieved by size exclusion, hydrophobic retention, ion exchange, and a variety of affinity-based procedures. These are but a few approaches that share with gel electrophoresis the means to separate mixtures of proteins for the purpose of preparing simplified samples for subsequent MS analysis. As with gel methods, column separation supports clean-up, enrichment, and maximum dynamic range in the analysis. Additional dimensions of separation have been applied at the protein and/or peptide level; this has led to a variety of multidimensional systems. Typically, protein-based separations (e.g. SEC) would be considered for the first dimension followed by protein digestion and subsequent reversed-phase separation of peptides for the second dimension (this latter dimension is the LC–MS–MS discussed above).

The approach has been extended to include multidimensional systems for the separation of wholesale digests of complex protein mixtures. This "shotgun proteomics" involves predigestion of unseparated protein mixtures followed by two dimensions of peptide chromatography, usually strong cation-exchange followed by reverse-phase [94]. The benefits of this approach over conventional gel-to-MALDI methods are several: fully automated systems can be assembled for analysis of complex protein mixtures; more informative sequence data are generated, and limits of detection are usually lower. The key weakness of the approach is the scrambling of the proteins during digestion. This results in loss of information relating to protein characteristics such as intact molecular weight, isoelectric point and relative quantitation – data provided by 2D gel protein separation. Digestion of protein mixtures disallows the assumption that any two peptides arise from the same protein, requiring a probability-based assessment of common origin. Dispensing with protein separation makes it difficult to "drill-down" and obtain additional protein-specific data (e.g. greater sequence coverage, analysis of post-translational modifications), thus these multidimensional systems are best viewed as tools for discovery. Relative quantitation can be achieved by use of the method, but specific isotopic labeling steps to generate internal standards are required to overcome the low quantitative power of the mass spectrometer.

A brief survey of some of the applications of multidimensional LC–MS–MS suggests the power of the general approach. The MudPIT (MUltiDimensional Protein Identification Technology) method developed by Yates et al. has been applied to full proteomic analysis of *S. cerevisiae* lysate, divided into three crude fractions [95]. An integrated strong cation-exchange and C_{18}-based reversed phase column was used as a front-end to an ion trap mass spectrometer. Fifteen fractions from the SCX column

were processed through the reversed-phase portion and all three fractions were processed similarly (Fig. 7.13). Overall, these analyses led to the detection of 5540 unique peptides in approximately 105 h of data-collection time, corresponding to 1484 unique proteins or nearly 15 identifications per hour (not including data analysis time). From a strict efficiency perspective, the rate of identification is not significantly different from the conventional gel-to-MALDI approach. It does, however, impart less compositional bias than the gel approach and limits of detection are lower. The system was operated at or near maximum sample complexity and under these conditions could be used to detect proteins present in low fmol amounts. This translates into a within-sample dynamic range of four orders of magnitude, or the potential to detect proteins at levels as low as 100 copies/cell.

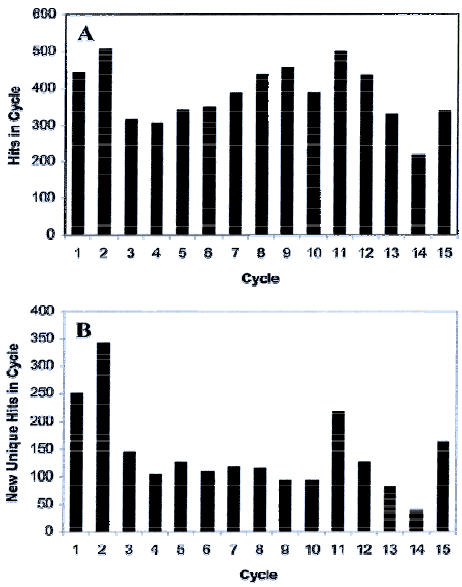


Fig. 7.13 Number of peptide identifications from each cycle of a 15-cycle MudPIT analysis of a heavily washed insoluble fraction from S. cerevisiae. (A) shows the total number of peptide identifications in each cycle. These peptide identifications are not necessarily unique, because a large number of peptides are identified several times during a typical MudPIT analysis. The total number of peptides identified in this sample was 5738. (B) shows the total number of new unique peptide identifications from each part of the cycle. The total number of unique peptides identified in this sample was 2114 [95]. With permission from American Chemical Society.

The greater ease with which multidimensional LC–MS–MS can be implemented using modern systems suggests these chromatographic approaches can be viewed as an extension of the MS detector – they do not preclude the incorporation of upstream protein-separation methods. For example, Wienkoop et al. have shown that anion-exchange prefractionation of proteins isolated from *Arabidopsis thaliana* followed by a shortened version of MudPIT led to identification of significantly more proteins (>1000) than a conventional MudPIT approach (~300) [96]. This method has the benefit of extending the dynamic range of the overall analysis by preventing high-abundance protein such as RUBISCO from appearing in all peptide fractions and preserving some measure of protein-specific information (i.e. detection of protein isoforms). This general concept underpins "pathway proteomics", in which interacting proteins are affinity purified by immobilization of a tagged bait protein plus all of its associated binding partners. All peptides arising from the digestion of the affinity-purified complex are assumed to originate from the complex thus implying a first-order assumption of functional relevance. Graumann et al. have rigorously evaluated the performance of MudPIT in conjunction with a revised tandem affinity purification procedure [7]. The general procedure involves isolating the bait protein via two independent affinity tags as a means of minimizing the isolation of nonspecifically

bound proteins. In this example the authors constructed bait proteins containing myc-epitopes separated from nine histidines by PreScission protease cleavage sites. Isolation of the bait and associated proteins could be achieved via pulldown assays using immobilized anti-myc antibody. The cleanup of the pulldown is achieved by proteolytic cleavage of the complex from the immobilized antibody, then identification based on nickel chelation to histidine. A MudPIT approach was applied to tryptic digests of the isolated complexes and the authors found the method to be both reproducible and capable of validating previously known interactions in addition to identifying new ones. As shown in Fig. 7.14, complete coverage of all known physical interactors is not achieved by this approach, and thus the method is best viewed as an iterative one in which newly discovered interactors are used as the next bait protein [97].

There has been a return of late to the analysis of intact proteins by use of multidimensional column-based systems. Conceptually similar to 2D gel electrophoresis in this regard, these systems present intact protein to the mass spectrometer for analysis [98]. Digestion and analysis of protein fractions remain an option with this system but, most importantly, it enables high resolution and accurate measurement of protein molecular weight. These data are useful in profiling protein heterogeneity, and the methods are essential for capitalizing on newer procedures in "top-down" proteomics. MudPIT and its variants are

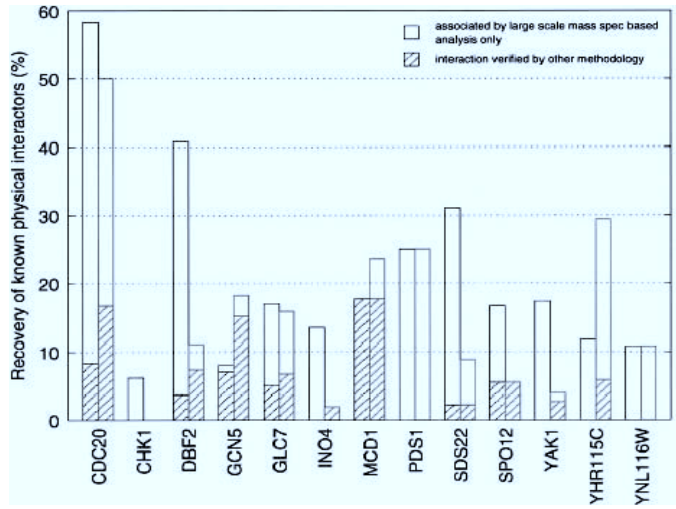


Fig. 7.14 Comparing yeast pathway proteomics experiments, using datasets from Graumann et al. [7] (grey) and Ho et al. [6] (black). The open reading frames listed were used as baits in both studies. Bars represent the percentage of previously known interacting partners (as reported in the MIPS, CYGD, GRID,and YPD databases) that were recovered in each immunoprecipitate experiment. Empty bars represent the percentage of gene products reported as interactors only by large-scale mass spectrometric analysis, whereas hatched bars represent interactions established or verified by other methods. Figure 7 from Ref. [7]. With permission from The American Society for Biochemistry and Molecular Biology.

increasingly referred to as "bottom-up" proteomics in which the first step of the analysis involves digestion of protein mixtures. In top-down, protein-specific information is generated first by analysis of the intact protein, followed by dissociation of the protein [99]. Gel-to-MALDI can be considered the earliest form of top-down. Novel methods for gas-phase dissociation of protein ions such as electron-capture dissociation and in-source decay hold some promise for dispensing with a digestion step and achieving completely MS-driven identification [100, 101].

7.6 Conclusions

Proteomics methods and technology are developing at a rapid pace, with each new innovation providing a new benchmark in detection limit, speed, and reliability. The drive to implement these approaches in a fashion usable by the non-expert in analytical sciences continues. An increasing number of mass spectrometers and associated systems are finding their way into conventional molecular biology laboratories, implying that some headway is being made. Many proteomics researchers are now turning their attention to issues of data quality and validation [7, 102]. Quantitating the value of the data generated is a serious issue, because proteomic data is error prone. Current and emerging databases need to reflect this quality issue to avoid the possibility of spurious biochemical inferences. This is a natural progression of the field that finds its parallel in genomics. Finally, proteomics researchers are pushing the technology beyond simple identification. Methods for protein quantitation and characterization of the myriad of post-translational modifications are also under development, with the sheer enormity of the protein world suggesting room for many new approaches.

References

1 Dunn, M. J. (2001). *Proteomics Reviews 2001*, IM Publications.

2 Figeys, D. (2003). "Proteomics in 2002: a year of technical development and wide-ranging applications." *Anal Chem* **75**(12):2891–905.

3 Yaneva, M. and P. Tempst (2003). "Affinity capture of specific DNA-binding proteins for mass spectrometric identification." *Anal Chem* **75**(23):6437–48.

4 Camacho-Carvajal, M. M., B. Wollscheid, et al. (2004). "Two-dimensional Blue native/SDS gel electrophoresis of multi-protein complexes from whole cellular lysates: a proteomics approach." *Mol Cell Proteomics* **3**(2):176–82.

5 Himeda, C. L., J. A. Ranish, et al. (2004). "Quantitative proteomic identification of six4 as the trex-binding factor in the muscle creatine kinase enhancer." *Mol Cell Biol* **24**(5):2132–43.

6 Ho, Y., A. Gruhler, et al. (2002). "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry." *Nature* **415**(6868):180–3.

7 Graumann, J., L. A. Dunipace, et al. (2004). "Applicability of tandem affinity purification MudPIT to pathway proteomics in yeast." *Mol Cell Proteomics* **3**(3):226–37.

8 Zhang, R., T. L. Tremblay, et al. (2003). "Identification of differentially expressed proteins in human glioblastoma cell lines and tumors." *Glia* **42**(2):194–208.

9 Stannard, C., V. Soskic, et al. (2003). "Rapid changes in the phosphoproteome show diverse cellular responses following stimulation of human lung fibroblasts with endothelin-1." *Biochemistry* **42**(47):13919–28.

10 Sinha, P., J. Poland, et al. (2003). "Study of the development of chemoresistance in melanoma cell lines using proteome analysis." *Electrophoresis* **24**(14):2386–404.

11 Hirabayashi, J., Y. Arata, et al. (2001). "Glycome project: concept, strategy and preliminary application to *Caenorhabditis elegans*." *Proteomics* **1**(2):295–303.

12 Kaji, H., H. Saito, et al. (2003). "Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins." *Nat Biotechnol* **21**(6):667–72.

13 Patton, W. F. (2002). "Detection technologies in proteome analysis." *J Chromatogr B Analyt Technol Biomed Life Sci* **771**(1/2):3–31.

14 Melle, C., R. Kaufmann, et al. (2004). "Proteomic profiling in microdissected hepatocellular carcinoma tissue using ProteinChip technology." *Int J Oncol* **24**(4):885–91.

15 Petricoin, E. F. and L. A. Liotta (2004). "SELDI–TOF-based serum proteomic pattern diagnostics for early detection of cancer." *Curr Opin Biotechnol* **15**(1):24–30.

16 Patterson, S. D. and R. H. Aebersold (2003). "Proteomics: the first decade and beyond." *Nat Genet* **33** Suppl:311–23.

17 Corthals, G. L., V. C. Wasinger, et al. (2000). "The dynamic range of protein expression: a challenge for proteomic research." *Electrophoresis* **21**(6):1104–15.

18 Patterson, S. D. (2003). "Data analysis – the Achilles heel of proteomics." *Nat Biotechnol* **21**(3):221–2.

19 Ramstrom, M. and J. Bergquist (2004). "Miniaturized proteomics and peptidomics using capillary liquid separation and high resolution mass spectrometry." *FEBS Lett* **567**(1):92–5.

20 Lee, C. L., H. H. Hsiao, et al. (2003). "Strategic shotgun proteomics approach for

efficient construction of an expression map of targeted protein families in hepatoma cell lines." *Proteomics* **3**(12):2472–86.
21 Zhou, M., D. A. Lucas, et al. (2004). "An investigation into the human serum "interactome"." *Electrophoresis* **25**(9):1289–98.
22 O'Donovan, C., M. J. Martin, et al. (2002). "High-quality protein knowledge resource: SWISS-PROT and TrEMBL." *Brief Bioinform* **3**(3):275–84.
23 Linial, M. (2003). "How incorrect annotations evolve – the case of short ORFs." *Trends Biotechnol* **21**(7):298–300.
24 Baker, C. J. www.sams.ucalgary.ca.
25 Feng, Z., L. Kachnic, et al. (2004). "DNA Damage Induces p53-dependent BRCA1 Nuclear Export." *J Biol Chem* **279**(27):28574–84.
26 Iwadate, Y., T. Sakaida, et al. (2004). "Molecular classification and survival prediction in human gliomas based on proteome analysis." *Cancer Res* **64**(7):2496–501.
27 Bisca, A., C. D'Ambrosio, et al. (2004). "Proteomic evaluation of core biopsy specimens from breast lesions." *Cancer Lett* **204**(1):79–86.
28 Craven, R. A. and R. E. Banks (2001). "Laser capture microdissection and proteomics: possibilities and limitation." *Proteomics* **1**(10):1200–4.
29 Rautajoki, K., T. A. Nyman, et al. (2004). "Proteome characterization of human T helper 1 and 2 cells." *Proteomics* **4**(1):84–92.
30 Cordwell, S. J. (2002). "Acquisition and archiving of information for bacterial proteomics: from sample preparation to database." *Methods Enzymol* **358**:207–27.
31 Nilsson, C. L. (2002). "Bacterial proteomics and vaccine development." *Am J Pharmacogenomics* **2**(1):59–65.
32 Cash, P. (2003). "Proteomics of bacterial pathogens." *Adv Biochem Eng Biotechnol* **83**:93–115.
33 Ghaemmaghami, S., W. K. Huh, et al. (2003). "Global analysis of protein expression in yeast." *Nature* **425**(6959):737–41.
34 Wohlschlegel, J. A. and J. R. Yates (2003). "Proteomics: where's Waldo in yeast?" *Nature* **425**(6959):671–2.
35 Schirmer, E. C., L. Florens, et al. (2003). "Nuclear membrane proteins with potential disease links found by subtractive proteomics." *Science* **301**(5638):1380–2.
36 Zhang, W., G. Zhou, et al. (2003). "Affinity enrichment of plasma membrane for proteomics analysis." *Electrophoresis* **24**(16):2855–63.
37 Rabilloud, T., J. M. Strub, et al. (2002). "Comparative proteomics as a new tool for exploring human mitochondrial tRNA disorders." *Biochemistry* **41**(1):144–50.
38 McDonald, T. G. and J. E. Van Eyk (2003). "Mitochondrial proteomics. Undercover in the lipid bilayer." *Basic Res Cardiol* **98**(4):219–27.
39 Ospina, J. K. and A. G. Matera (2002). "Proteomics: the nucleolus weighs in." *Curr Biol* **12**(1):R29–31.
40 Schuldt, A. (2002). "Proteomics of the nucleolus." *Nat Cell Biol* **4**(2):E35.
41 Desjardins, M. (2003). "ER-mediated phagocytosis: a new membrane for new functions." *Nat Rev Immunol* **3**(4):280–91.
42 Taylor, S. W., E. Fahy, et al. (2003). "Global organellar proteomics." *Trends Biotechnol* **21**(2):82–8.
43 Foster, L. J., C. L. De Hoog, et al. (2003). "Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors." *Proc Natl Acad Sci U S A* **100**(10):5813–8.
44 Wu, S. L., H. Amato, et al. (2002). "Targeted proteomics of low-level proteins in human plasma by LC/MSn: using human growth hormone as a model system." *J Proteome Res* **1**(5):459–65.
45 Haney, P. J., C. Draveling, et al. (2003). "SwellGel: a sample preparation affinity chromatography technology for high throughput proteomic applications." *Protein Expr Purif* **28**(2):270–9.
46 Pieper, R., Q. Su, et al. (2003). "Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome." *Proteomics* **3**(4):422–32.
47 Wang, Y. Y., P. Cheng, et al. (2003). "A simple affinity spin tube filter method for removing high-abundant common proteins or enriching low-abundant biomarkers for serum proteomic analysis." *Proteomics* **3**(3):243–8.
48 Xu, Q. and K. S. Lam (2003). "Protein and Chemical Microarrays-Powerful Tools for Proteomics." *J Biomed Biotechnol* **2003**(5):257–266.

49 Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." *Nature* **415**(6868):141–7.

50 Gavin, A. C. and G. Superti-Furga (2003). "Protein complexes and proteome organization from yeast to man." *Curr Opin Chem Biol* **7**(1):21–7.

51 Stuart, J. N., A. B. Hummon, et al. (2004). "The chemistry of thought: neurotransmitters in the brain." *Anal Chem* **76**(7):121A–128A.

52 Chaurand, P., S. A. Schwartz, et al. (2004). "Integrating histology and imaging mass spectrometry." *Anal Chem* **76**(4):1145–55.

53 Diamandis, E. P. (2004). "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations." *Mol Cell Proteomics* **3**(4):367–78.

54 Mehta, A. I., S. Ross, et al. (2003). "Biomarker amplification by serum carrier protein binding." *Dis Markers* **19**(1):1–10.

55 Hu, S., Z. Le, et al. (2003). "Cell cycle-dependent protein fingerprint from a single cancer cell: image cytometry coupled with single-cell capillary sieving electrophoresis." *Anal Chem* **75**(14):3495–501.

56 Hu, S., Z. Le, et al. (2003). "Identification of proteins in single-cell capillary electrophoresis fingerprints based on comigration with standard proteins." *Anal Chem* **75**(14):3502–5.

57 Karas, M. and F. Hillenkamp (1988). "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons." *Anal Chem* **60**(20):2299–301.

58 Fenn, J. B. (2003). "Electrospray wings for molecular elephants (Nobel lecture)." *Angew Chem Int Ed Engl* **42**(33):3871–94.

59 Frankevich, V. E., J. Zhang, et al. (2003). "Role of electrons in laser desorption/ionization mass spectrometry." *Anal Chem* **75**(22):6063–7.

60 Keller, B. O. and L. Li. (2001). "Detection of 25,000 molecules of substance P by MALDI–TOF mass spectrometry and investigations into the fundamental limits of detection in MALDI." *J Am Soc Mass Spectrom* **12**(9):1055–1063.

61 Mora, J. F., G. J. Van Berkel, et al. (2000). "Electrochemical processes in electrospray ionization mass spectrometry." *J Mass Spectrom* **35**(8):939–52.

62 Covey, T. R., E. C. Huang, et al. (1991). "Structural characterization of protein tryptic peptides via liquid chromatography/mass spectrometry and collision-induced dissociation of their doubly charged molecular ions." *Anal Chem* **63**(13):1193–200.

63 Peng, J., J. E. Elias, et al. (2003). "Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC–LC–MS–MS) for large-scale protein analysis: the yeast proteome." *J Proteome Res* **2**(1):43–50.

64 Pucci, P., C. Carestia, et al. (1985). "Protein fingerprint by fast atom bombardment mass spectrometry: characterization of normal and variant human haemoglobins." *Biochem Biophys Res Commun* **130**(1):84–90.

65 Cottrell, J. S. (1994). "Protein identification by peptide mass fingerprinting." *Peptide Res* **7**(3):115–24.

66 Henzel, W. J., C. Watanabe, et al. (2003). "Protein identification: the origins of peptide mass fingerprinting." *J Am Soc Mass Spectrom* **14**(9):931–42.

67 Pappin, D. J. C., P. Hojrup, et al. (1993). "Rapid identification of proteins by peptide-mass fingerprinting." *Curr Biol* **3**(6):327–332.

68 Perkins, D. N., D. J. Pappin, et al. (1999). "Probability-based protein identification by searching sequence databases using mass spectrometry data." *Electrophoresis* **20**(18):3551–67.

69 Lim, H., J. Eng, et al. (2003). "Identification of 2D-Gel Proteins: a comparison of MALDI/TOF Peptide Mass Mapping to microLC–ESI Tandem Mass Spectrometry." *J Am Soc Mass Spectrom* **14**:957–970.

70 Clauser, K. R., P. Baker, et al. (1999). "Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching." *Anal Chem* **71**(14):2871–82.

71 Kinter, M. and N. E. Sherman (2000). *Protein Sequencing and Identification using Tandem Mass Spectrometry*. New York, John Wiley & Sons.

72 Aitken, A. and M. Learmonth (2002). "Protein identification by in-gel digestion and mass spectrometric analysis." *Mol Biotechnol* **20**(1):95–7.

73 Courchesne, P. L., R. Luethy, et al. (1997). "Comparison of in-gel and on-membrane digestion methods at low to sub-pmol level for subsequent peptide and fragment-ion mass analysis using matrix-assisted laser-desorption/ionization mass spectrometry." *Electrophoresis* **18**(3/4):369–81.

74 Havlis, J., H. Thomas, et al. (2003). "Fast-response proteomics by accelerated in-gel digestion of proteins." *Anal Chem* **75**(6):1300–6.

75 Shevchenko, A., M. Wilm, et al. (1996). "Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels." *Anal Chem* **68**(5):850–8.

76 Loboda, A. V., S. Ackloo, et al. (2003). "A high-performance matrix-assisted laser desorption/ionization orthogonal time-of-flight mass spectrometer with collisional cooling." *Rapid Commun Mass Spectrom* **17**(22):2508–16.

77 Brock, A., D. M. Horn, et al. (2003). "An automated matrix-assisted laser desorption/ionization quadrupole Fourier transform ion cyclotron resonance mass spectrometer for "bottom-up" proteomics." *Anal Chem* **75**(14):3419–28.

78 Bruce, J. E., G. A. Anderson, et al. (1999). "High-mass-measurement accuracy and 100% sequence coverage of enzymatically digested bovine serum albumin from an ESI–FTICR mass spectrum." *Anal Chem* **71**(14):2595–9.

79 Smith, R. D., G. A. Anderson, et al. (2002). "An accurate mass tag strategy for quantitative and high-throughput proteome measurements." *Proteomics* **2**(5):513–23.

80 Slysz, G. W. and D. C. Schriemer (2003). "On-column digestion of proteins in aqueous-organic solvents." *Rapid Commun Mass Spectrom* **17**(10):1044–50.

81 Steen, H., B. Kuster, et al. (2001). "Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode." *Anal Chem* **73**(7):1440–8.

82 Laiko, V. V., S. C. Moyer, et al. (2000). "Atmospheric pressure MALDI/ion trap mass spectrometry." *Anal Chem* **72**(21):5239–43.

83 Yergey, A. L., J. R. Coorssen, et al. (2002). "De novo sequencing of peptides using MALDI/TOF–TOF." *J Am Soc Mass Spectrom* **13**(7):784–91.

84 Zhu, X. and I. A. Papayannopoulos (2003). "Improvement in the detection of low concentration protein digests on a MALDI TOF/TOF workstation by reducing alpha-cyano-4-hydroxycinnamic acid adduct ions." *J Biomol Tech* **14**(4):298–307.

85 Wysocki, V. H., G. Tsaprailis, et al. (2000). "Mobile and localized protons: a framework for understanding peptide dissociation." *J Mass Spectrom* **35**(12):1399–406.

86 Keough, T., M. P. Lacey, et al. (2001). "Atmospheric pressure matrix-assisted laser desorption/ionization ion trap mass spectrometry of sulfonic acid derivatized tryptic peptides." *Rapid Commun Mass Spectrom* **15**(23):2227–39.

87 Licklider, L. J., C. C. Thoreen, et al. (2002). "Automation of nanoscale microcapillary liquid chromatography–tandem mass spectrometry with a vented column." *Anal Chem* **74**(13):3076–83.

88 Peng, J. and S. P. Gygi (2001). "Proteomics: the move to mixtures." *J Mass Spectrom* **36**(10):1083–91.

89 Shen, Y., N. Tolic, et al. (2004). "Ultrasensitive proteomics using high-efficiency on-line micro-SPE-nanoLC-nanoESI MS and MS/MS." *Anal Chem* **76**(1):144–54.

90 Shen, Y., N. Tolic, et al. (2004). "Nanoscale proteomics." *Anal Bioanal Chem* **378**(4):1037–45.

91 Barroso, B., D. Lubda, et al. (2003). "Applications of monolithic silica capillary columns in proteomics." *J Proteome Res* **2**(6):633–42.

92 Strittmatter, E. F., P. L. Ferguson, et al. (2003). "Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry." *J Am Soc Mass Spectrom* **14**(9):980–91.

93 Zhen, Y., N. Xu, et al. (2004). "Development of an LC–MALDI method for the analysis of protein complexes." *J Am Soc Mass Spectrom* **15**(6):803–22.

94 MacCoss, M. J., W. H. McDonald, et al. (2002). "Shotgun identification of protein modifications from protein complexes and lens tissue." *Proc Natl Acad Sci U S A* **99**(12):7900–5.

95 Wolters, D. A., M. P. Washburn, et al. (2001). "An automated multidimensional protein identification technology for shotgun proteomics." *Anal Chem* **73**(23):5683–90.

96 Wienkoop, S., M. Glinski, et al. (2004). "Linking protein fractionation with multidimensional monolithic reversed-phase peptide chromatography/mass spectrometry enhances protein identification from complex mixtures even in the presence of abundant

proteins." *Rapid Commun Mass Spectrom* **18**(6):643–50.

97 Seol, J. H., A. Shevchenko, et al. (2001). "Skp1 forms multiple protein complexes, including RAVE, a regulator of V-ATPase assembly." *Nat Cell Biol* **3**(4):384–91.

98 Yan, F., A. Sreekumar, et al. (2003). "Protein microarrays using liquid phase fractionation of cell lysates." *Proteomics* **3**(7):1228–35.

99 Nemeth-Cawley, J. F., B. S. Tangarone, et al. (2003). "Top Down" characterization is a complementary technique to peptide sequencing for identifying protein species in complex mixtures." *J Proteome Res* **2**(5): 495–505.

100 Ge, Y., B. G. Lawhorn, et al. (2002). "Top down characterization of larger proteins (45 kDa) by electron capture dissociation mass spectrometry." *J Am Chem Soc* **124**(4): 672–8.

101 Kelleher, N. L. (2004). "Top-down proteomics." *Anal Chem* **76**(11):197A–203A.

102 Venable, J. D. and J. R. Yates, 3rd (2004). "Impact of ion trap tandem mass spectra variability on the identification of peptides." *Anal Chem* **76**(10):2928–37.

8
Proteome Analysis by Capillary Electrophoresis

Md Abul Fazal, David Michels, James Kraly, and Norman J. Dovichi

8.1 Introduction

The human genome contains 3×10^9 bases with ~30,000 genes. Alternative splicing of exons results in perhaps 100,000 different proteins, which undergo post-translational modification to generate an unknown number of products. This set of proteins is called the proteome [1]. Unlike the genome, which is essentially static and identical in each cell of an organism, the proteome varies from cell-to-cell and changes in response to the environment and during development and disease.

Two-dimensional electrophoresis, followed by mass-spectrometric analysis of isolated proteins, has been the workhorse tool for proteomics research [2]. The protein extract is first separated by isoelectric focusing, which is performed in either strip or tube format. The isoelectric focusing gel is then placed at the top of an SDS/PAGE gel and proteins are separated on the basis of size. The proteins are detected by a staining procedure that creates a two-dimensional set of spots. The location of each spot is determined by the protein's mass and isoelectric point. The intensity of each spot is related to the amount of the protein present in the cell extract. Those proteins with interesting expression patterns are cut from the gel, digested by a protease, and identified by mass spectrometry and database searching.

More recently, methods have been developed on the basis of digestion of the protein homogenate with a protease then liquid-chromatographic separation and mass-spectrometric identification of the peptide fragments. This automated method is particularly useful both for surveying protein expression in a sample, as in MudPIT experiments, and in determining changes in protein expression between two states of the organism, as in ICAT experiments [3, 4]. Neither approach is particularly useful in determining either the absolute amount of a protein in a sample or any post-translational modification of the expressed proteins, although efforts are in progress to determine such information. In most experiments proteins are extracted from the lysate generated from several million cells and perhaps 1000 different proteins are identified.

Proteomic research differs from genomic research in one important way. DNA research benefits greatly from PCR technolo-

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

gy, which enables amplification of specific genomic regions so that a few copies of interesting mRNA or genomic DNA can be amplified to levels that are easily handled. Protein research does not have an analogous tool. As a result, protein research requires the use of extremely high sensitivity analytical tools to monitor proteins expressed at low levels.

8.2 Capillary Electrophoresis

Capillary array electrophoresis has proven to be a powerful tool for DNA analysis [5]. Automated instruments have been commercialized and used to sequence the human genome. We believe capillary electrophoresis can also play a useful role in proteome analysis.

8.2.1 Instrumentation

Capillaries are typically 10 to 50 µm inner diameter, 150 µm outer diameter, and 35 cm long. The outside of the capillary is coated with a thin layer of polymer, usually polyimide, which gives the capillaries great strength and flexibility. The total volume of the capillary is usually less than 1 µL and sample volume is typically a few nanoliters.

Instrumentation for capillary electrophoresis is very simple. It consists of an injector to introduce sample into the capillary, the capillary, a detector, and a high-voltage power supply to drive the separation. A computer controls injection and automatically records data. Because of the high voltage employed for separation, the operator must be protected with a safety interlock system.

8.2.2 Injection

Analyte is usually held in a disposable microcentrifuge tube. Injection is performed by dipping the capillary into the sample. In electrokinetic injection, a high voltage electrode is also placed in contact with the sample, and an electric field is applied to the sample for a few seconds, drawing the protein into the capillary. Electrokinetic injection is biased; the amount of sample injected is proportional to the sample's velocity during electrophoresis [6]. More of the faster moving analytes are injected than of the slower moving components.

Alternatively, sample can be injected hydrodynamically by applying pressure to the sample or vacuum to the distal end of the capillary. Hydrodynamic injection is unbiased; the amount injected is independent of the physical properties of the analyte. Irrespective of which injection method is used, the sample volume should be less than ~0.1 % of the capillary volume to preserve separation efficiency.

Conventionally, a few nanoliters of sample is injected on to the capillary. Fortunately, stacking methods developed for capillary electrophoresis enable injection of larger volumes of sample without overloading the separation [7, 8].

8.2.3 Electroosmosis

Electroosmosis is particularly important when separations are performed in uncoated capillaries. Residual silanol groups on the capillary surface take a negative charge at neutral or basic pH. These silanol groups are fixed and do not move on application of an electric field to the capillary. Electrical

neutrality requires that cationic counter ions be present in a diffuse cloud very near the capillary wall. These cations migrate in an electric field toward the negative electrode. The cations draw solvent with them, transporting bulk fluid to the negative electrode. This bulk flow is called electroosmosis. Unlike pressure-driven flow in chromatography, electroosmotic flow velocity is uniform across the capillary lumen, which minimizes peak broadening so that the separation efficiency is much higher than in liquid chromatography.

During analysis, electrophoresis and electroosmosis occur simultaneously. The overall mobility of an ion is the sum of the electrophoretic and electroosmotic mobilities. Electroosmosis is usually stronger than electrophoresis, and all components of a sample will migrate through the capillary to the detector, with cations migrating first, neutrals second, and anions last.

Electroosmosis can be reduced to negligible levels by coating the capillary. Neutral coating can be chemically bound to the wall by use of either silane or Grignard reactions [9]. Dynamic coating relies on the physical adsorption of, usually, a polymer by the capillary wall; dynamic coatings are not as robust as chemical coatings but can be regenerated as needed by successive treatment with the coating reagent.

8.2.4 Separation

Specific separation methods for proteins are discussed in Sect. 8.3. Here we present general information on capillary electrophoresis separations. The capillary is filled with a separation buffer, the nature of which determines the separation mechanism. The buffer is usually a few millimolar in concentration, although low-ionic-strength zwitterionic buffers can be used at higher concentration; it is important not to employ a high ionic strength buffer, which suffer from excessive Joule heating.

When the sample has been injected into the capillary, the sample vial is replaced with a buffer-filled vial, and an electric field is used to separate the components within the sample. The narrow diameter of the capillary minimizes the temperature rise associated with the electric field. As a result, very high potentials can be applied across the capillary without degradation of the separation as a result of heating. Potentials up to 30,000 V and electric fields exceeding 450 V cm^{-1} are routinely employed. The use of high fields results in rapid and efficient separations.

Unlike slab gels, on which proteins are visualized after a fixed run time, capillary electrophoresis is a finish-line technique – the proteins are detected after traversing a fixed distance. The time necessary for a protein to migrate through the capillary is given by:

$$t = \frac{IL}{(\mu_{eof} + \mu_{ep})\, V}$$

where L is the length of the capillary, l is the distance from the injection end of the capillary to the detector, μ_{ep} is the electrophoretic mobility of the analyte, μ_{eof} is the electroosmotic mobility of the buffer, and V is the applied potential [10]. The separation time is inversely proportional to the applied potential. For post-column detectors, $l = L$ and the separation time is proportional to the length of the capillary squared. Separation time decreases dramatically as capillary length is shortened. Typical protein separations are completed in less than 15 min, and separations as fast as 60 s have been demonstrated.

8.2.5 Detection

UV absorbance, laser-induced fluorescence, and mass spectrometry are used for protein detection. UV absorbance measurements use the capillary itself as the detection cuvette. Unfortunately, the relatively short optical pathlength across the capillary tends to result in poor sensitivity in absorbance measurements, which must be performed with relatively high-concentration samples.

Mass spectrometry is usually better suited to analysis of peptides rather than proteins; sensitivity and mass resolution fall for the higher-molecular-weight species. Electrospray ionization is a common interface between capillary electrophoresis and mass spectrometry [11]. Electrospray mass spectrometry has been used with impressive results for detection of peptides separated by isoelectric focusing. Unfortunately, electrospray is incompatible with capillary separations based on a surfactant. The surfactant disrupts electrospray ionization and the surfactant ions tend to contaminate the detector.

Fluorescence detection from native proteins requires the use of expensive and temperamental lasers that operate in the ultraviolet portion of the spectrum. Instead, proteins are usually labeled with a fluorescent reagent that might be excited with lower cost lasers that operate in the visible or near infrared. We find that classic fluorescent reagents, for example fluorescein isothiocyanate, are not convenient for protein labeling in capillary electrophoresis. Unreacted reagent and fluorescent impurities generate an intense background signal that can saturate the fluorescence detector and obscure the signal from labeled proteins. Instead, we prefer the use of fluorogenic reagents [12, 13]. These reagents are nonfluorescent themselves but produce a highly fluorescent product on reaction with a primary amine, such as the ε-amine of lysine residues. Because the reagents are nonfluorescent until they react with the amine, the background signal from unreacted reagent does not interfere in the measurement. In particular, we employ the reagent 3-(2-furoyl)quinoline-2-carboxaldehyde (FQ) for protein labeling. Reaction products are excited efficiently in the blue portion of the spectrum. In the work discussed in this chapter we employed either an air-cooled argon ion laser or a solid-state, diode pumped laser, both of which operate at 488 nm with 5–10 mW power. Fluorescence is detected with either a photomultiplier tube or a single-photon counting-avalanche photodiode.

There is a subtle point concerning labeling chemistry. Lysine is a common amino acid and there usually are several lysine residues in each protein. The labeling reaction seldom goes to completion and a complex a mixture of reaction products is formed. A single protein with n primary amines can produce 2^n–1 possible fluorescent products [14]. For example, ovalbumin has 20 lysine residues and a blocked N-terminus, so 1,048,575 different products can be generated on reaction with a labeling reagent. These products have different mobilities and result in a broad and complex electrophoresis pattern. Much effort has been devoted to developing reaction procedures and buffer systems that minimize the effects of multiple labeling. As we show below, the use of an appropriate separation buffer eliminates the effects of multiple labeling in SDS/gel and free solution electrophoresis. Capillary isoelectric focusing remains incompatible with fluorescence labeling technology, however [15].

8.3 Capillary Electrophoresis for Protein Analysis

Three types of electrophoresis are used for protein analysis. Two – isoelectric focusing and SDS/gel electrophoresis – are similar to classic protein electrophoresis techniques. The third is unique to capillary electrophoresis and is based on migration in a simple buffer.

8.3.1 Capillary Isoelectric Focusing

Stellan Hjerten performed the first work on capillary isoelectric focusing [16]. The sample was mixed with ampholytes and used to fill the capillary. A high electric field was applied, focusing the proteins at their isoelectric point and rapidly generating a high-resolution electropherogram [17]. Typically, the column was coated to minimize electroosmosis and reduce protein adsorption so that a stationary isoelectric focusing profile is formed in the capillary. When the separation is complete, the proteins must be detected. Although the proteins can be visualized with an imaging detector [18], it is much more common to mobilize the proteins so that they flow through a detector at one end of the capillary. The isoelectric focusing pattern is recorded as the proteins pass through the detector.

Protein mobilization can be performed in several ways. If an uncoated capillary is used for the separation, residual electroosmosis will force the proteins to migrate past the detector. Coated capillaries are, however, frequently used to improve the resolution of the separation and these capillaries have negligible electroosmosis. In this case, several methods can be used to drive the contents of the capillary through the detector.

Electrokinetic mobilization of proteins can be performed by addition of a neutral salt such as sodium chloride to either the cathode or anode buffer to drive proteins from the capillary to the detector [19]. This salt migrates into the capillary under the influence of an electric field, driving the focused proteins to the opposite end of the capillary.

Alternatively, one end of the capillary can be pressurized to drive the contents through the detector. The pressure must be very low to avoid peak broadening due to formation of a parabolic flow profile. The pressure can be introduced most simply by lowering the detector end of the capillary relative to the injection end to form a siphon which draws the contents of the capillary through the detector.

UV absorbance and mass spectrometry are used as detection techniques for isoelectric focusing. Mass spectrometry is particularly important, because it can enable extremely high-accuracy molecular mass determination for proteins. The resulting data resemble two-dimensional gels with much higher mass resolution than can be produced by SDS/PAGE [10]. Unfortunately, Fourier-transform mass spectrometers and electrospray ionization are required for analysis of the high-molecular-weight proteins. This technology is expensive and not routinely available.

Fluorescence detection is not useful for labeled proteins; multiple labeling produces a complex mixture of reaction products that results in a complex electropherogram [15].

8.3.2 SDS/Capillary Sieving Electrophoresis

SDS/PAGE is a commonly used slab electrophoresis method for separation of pro-

teins on the basis of their size. Unfortunately, the crosslinked polyacrylamide used in slab-gels is not appropriate for capillary-based separations. The capillary containing the crosslinked polymer must be discarded after each run, which is expensive. Instead, non-crosslinked polymers are usually employed for size-based separation of proteins. These materials have relatively low viscosity and can be pumped into the capillary after each run in an automated system. Because neither polyacrylamide nor a gel is used for the separation, the term PAGE is not appropriate to describe this separation. Instead, IUPAC has recommended the term capillary sieving electrophoresis (CSE) to describe all separations based on migration through a polymer or other sieving matrix.

Several different polymers have been used to separate proteins in capillary electrophoresis [20]. Polyacrylamide is not particularly useful. It has low shelf life and must be produced on demand. Unfortunately, the free redial polymerization is not highly reproducible, which is undesirable when comparing separations of samples with standards.

Instead, we have found that several polymers, including poly(ethylene oxide), pullulan, and dextran, can be used for size-based protein separations. As an example, Fig. 8.1 depicts the CSE separation of proteins extracted from the adrenal gland cell line, AtT 20. Quite a few peaks are resolved in this separation. To aid identification of these proteins we performed a detergent fractionation. Figure 8.1 also shows traces generated from the cytosolic, membrane, nuclear, and cytoskeleton fraction of proteins from this cell line. It is important to note that the membrane-fraction generates several components, which suggests that capillary electrophoresis, with appropriate buffers, can be used to characterize this important class of proteins. Finally, the separation is reasonably fast – complete in 15 min.

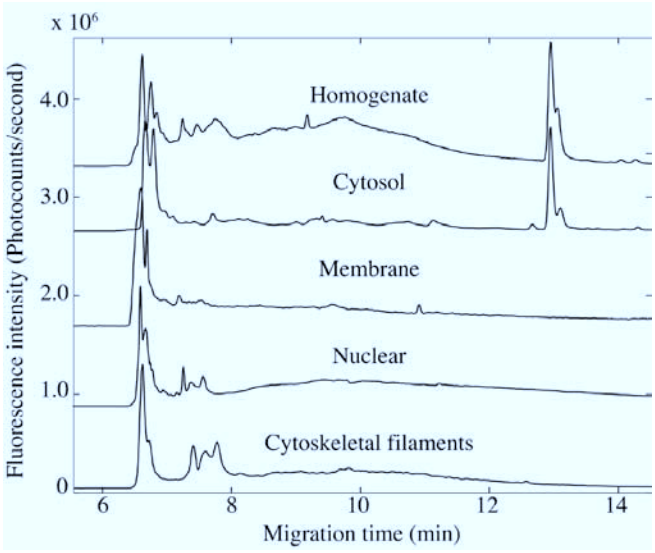


Fig. 8.1 CSE separation of proteins from the AtT20 cell line. Proteins were labeled with the fluorogenic reagent FQ and detected by laser-induced fluorescence.

8.3.3 Free Solution Electrophoresis

Capillary free solution electrophoresis does not have an analog in slab electrophoresis. A capillary is filled with a millimolar buffer and proteins are separated on the basis of their mobility in the buffer.

To reduce interaction of the proteins with the capillary wall, a modest amount of anionic surfactant is often added to the separation buffer. This buffer tends to ion-pair with cationic amino acid residues that would otherwise interact with silanol groups on the capillary wall. This interaction would lead to adsorption of the protein by the capillary, which leads to band broadening and reduced resolution. More importantly, the surfactant reduces the effects of multiple labeling to negligible levels [12]. Surfactants also interact with the hydrophobic portions of the protein, again reducing interaction with the capillary walls. Surfactant concentrations above and below the critical micelle concentration (CMC) have been used. If the surfactant concentration is above the CMC proteins can partition into the micelle and the separation is called micellar electrokinetic capillary chromatography (MECC). Although the separation mechanism is complex, free solution electrophoresis in the presence of a surfactant will be based, in part, on the hydrophobicity of the protein.

A micellar separation of a protein homogenate from the AtT20 cell line is depicted in Fig. 8.2. As in CSE, the free solution separation resolves several of the components of this mixture. We are curious about the identity of these components, and whether the analysis generates a representative sample of the cellular components and the separation was studied further. Figure 8.2 also shows free-solution separations of the cytosolic, membrane, nuclear, and cytoskeleton fractions isolated from the protein by detergent fractionation. The separation of the cytosolic protein contains several sharp peaks indicative of high efficiency, Fig. 8.3. The

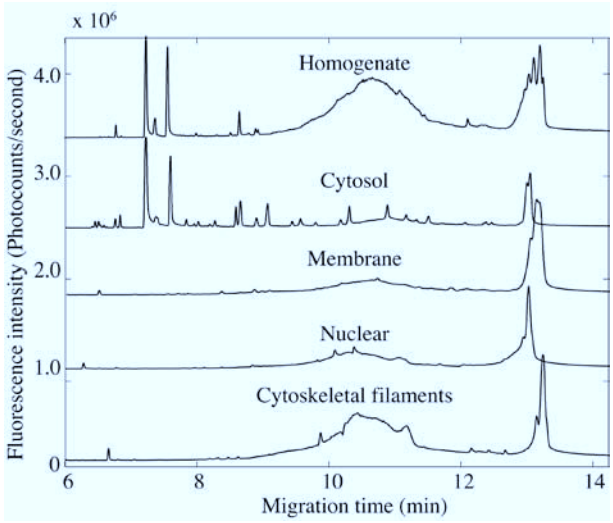


Fig. 8.2 MECC separation of proteins from the AtT20 cell line. Proteins were labeled with the fluorogenic reagent FQ and detected with laser-induced fluorescence.

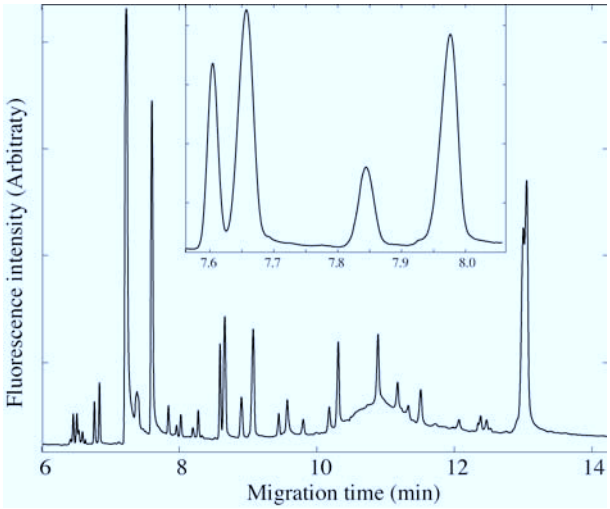


Fig. 8.3 MECC separation of the cytosolic protein fraction from AtT20 cells.

insert shows several peaks in greater detail; the first peak produces 1.2 million theoretical plates, or 3 million plates per meter. We believe that this is the highest efficiency separation ever reported for a protein homogenate.

8.4 Single-cell Analysis

Capillary electrophoresis with laser-induced fluorescence detection is an extraordinarily sensitive analytical technique. We have obtained subpicomolar concentration detection limits for proteins; only a few zeptomoles of protein are used for the analysis. This extraordinary sensitivity provides opportunities to perform proteome analysis of single somatic cells [21]. These cells are typically ~10 µm in diameter and 0.5 pL in volume. Assuming the cell is 10 % protein by weight, it contains 500 ng protein. The average molecular weight of the protein is ~30 kDa and the cell contains approximately 2 femtomole protein. A typical cell may express ~10,000 proteins; the average protein would be present as 100,000 copies or 200 zeptomoles. Of course, proteins are not expressed uniformly. Structural proteins would constitute most of the mass of proteins. The vast majority of proteins would be expressed at relatively low copy number. Nevertheless, capillary electrophoresis with laser-induced fluorescence should detect a large number of proteins from a single cell. We have published electropherograms from single cells from a number of different sources, including a colon cancer cell line, a *C. elegans* zygote, an osteoprecursor cell line, and a breast cancer cell line [21–26].

Figure 8.4 depicts results from CSE analysis of the proteins in a single AtT20 adrenal gland cell. In this experiment, the cell was injected into a 30 µm i.d., 50 cm long fused-silica capillary. The capillary had been coated with polyacrylamide to reduce electroosmosis and filled with a 50 mM Tris, 50 mM CHES, 5 mM SDS, and 2.5 % dextran (513 kDa) buffer, pH 8.7. The cell was

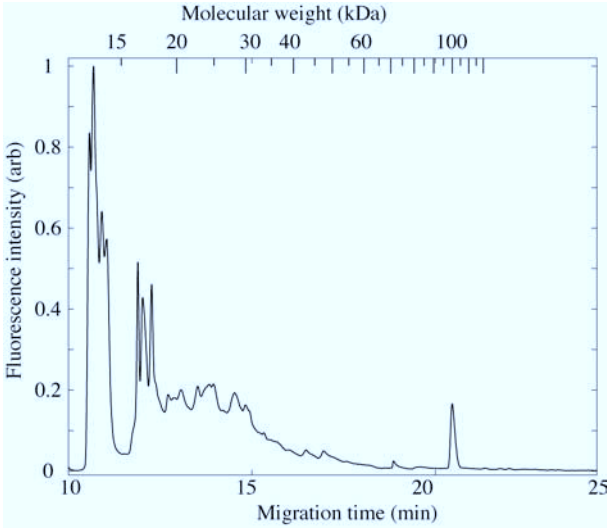


Fig. 8.4 CSE separation of the protein from a single AtT20 cell. The single cell was aspirated into the separation capillary, lysed during contact with the SDS-containing buffer, labeled with the fluorogenic reagent FQ, separated by CSE, and detected by laser-induced fluorescence. The molecular weight scale was determined from analysis of standard proteins.

lysed by contact with SDS within the separation medium. The freed proteins were fluorescently labeled with FQ, separated by CSE at 15 kV, and detected by laser-induced fluorescence in a sheath-flow cuvette detector. Approximately 25 components are resolved in this electropherogram; resolution was similar to that produced by conventional SDS/PAGE analysis of the cell extract from several million cells. A slightly longer capillary was used in this experiment compared with that used to study sub-cellular fractions, and the separation required slightly longer time to complete.

8.5 Two-dimensional Separations

Capillary electrophoresis versions of isoelectric focusing, SDS/PAGE, and free solution electrophoresis have been demonstrated for protein analysis. While these techniques produce quite impressive separation power, they are one-dimensional analyses, and their ability to separate complex mixtures is limited. Jorgenson demonstrated comprehensive two-dimensional analysis by combining capillary size-exclusion chromatography with free solution electrophoresis for analysis of peptides [27].

We are modifying Jorgenson's technology for two-dimensional capillary electrophoresis of proteins [28]. In our method a protein mixture is injected into the first capillary in which components are separated by capillary sieving electrophoresis [29]. As components migrate from the first capillary, successive fractions are transferred to a second column, where co-migrating components are separated by free solution electrophoresis. Our technology differs from

Jorgen-son's in an important way. In Jorgenson's publications, high-performance liquid chromatography is employed in the first dimension. Because pressure released slowly, analyte flows continually from the first dimension capillary, even during the second dimension separation. In contrast, we reduce the voltage across the first capillary to zero during the second dimension separation. In this way, analytes are stationary in the first capillary during the second dimension separation, which greatly simplifies the comprehensive analysis of the components in the mixture. This procedure of transfer and analysis is repeated to build a two-dimensional electropherogram as a raster image from successive separations of the first capillary's fractions.

Figure 8.5 depicts the two-dimensional separation of proteins obtained from a homogenate of *Deinococcus radiodurans*. This prokaryote is the most radiation-hardy organism known and has potential application in the bioremediation of nuclear waste sites. In this figure successive CSE fractions are transferred to the MECC capillary for further separation. The data are presented in the form of a landscape in which signal height is proportional to fluorescence intensity. A sea of peaks is observed, corresponding to the complex proteome of this prokaryote.

This analysis used fluorescently labeled proteins and laser-induced fluorescence detection. The detector produces signal-to-noise ratios greater than 10^4, which means that both major and minor components can be quantified in the same electropherogram. Furthermore, extremely small amounts of protein are required for the analysis; attomoles of protein were used to generate the data of Fig. 8.5. In this separation, roughly 50 fractions were transferred from the CSE capillary to the MECC capillary, and relatively long separation was used in the second dimension to resolve overlapping components.

Figure 8.6 shows the two-dimensional separation of proteins from a human esophageal cell line derived from a patient with the precancerous condition Barrett's esophagus. This separation differs from the previous example. Here, approximately 250 fractions are transferred from the CSE capillary and a very fast MECC separation, 20 s, is used in the second dimension.

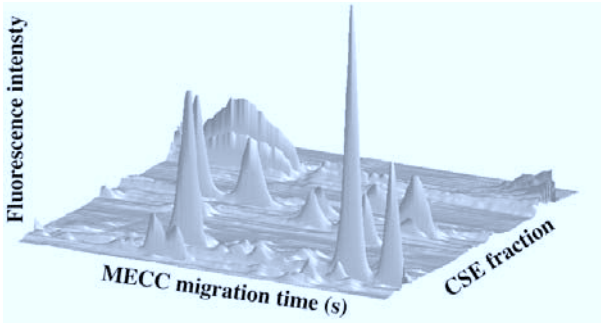


Fig. 8.5 Two-dimensional capillary electrophoresis separation of a protein homogenate prepared from the bacterium *Deinococcus radiodurans*. Proteins were separated by CSE in the first dimension and MECC in the second. Roughly 50 fractions were transferred to the second dimension capillary.

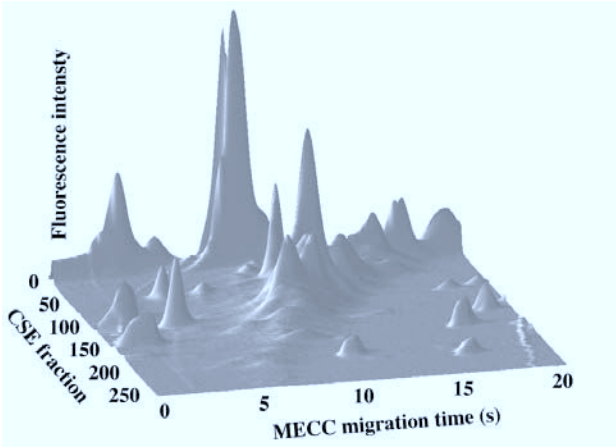


Fig. 8.6 Two-dimensional capillary electrophoresis separation of a protein homogenate prepared from a cell line generated from a Barrett's esophagus patient. Proteins were separated by CSE in the first dimension and MECC in the second. Over 250 fractions were transferred to the second-dimension capillary.

8.6 Conclusions

Capillary electrophoresis enables rapid and sensitive separation of proteins, often producing resolution similar to that produced by conventional slab-gel based electrophoresis methods. Unlike conventional methods, capillary electrophoresis is easily automated for analysis of large numbers of samples without requiring operator attention.

Protein identification is a challenge. The amount of protein in a single cell is orders of magnitude smaller than that which can be detected by mass spectrometry. Instead, we rely on spiking of the sample with identified proteins isolated from an electrophoresis gel. Co-migration during capillary electrophoresis is taken as evidence of the identity of the peak [30].

Two-dimensional separations are being developed by coupling two capillaries; the second capillary is used to separate fractions that migrate from the first capillary. This separation method will produce separations that are similar to those of IEF-SDS/PAGE in a fully automated system. Furthermore, when combined with laser-induced fluorescence detection the separation will provide orders of magnitude higher sensitivity and dynamic range than current IEF-SDS/PAGE analysis.

When coupled with laser-induced fluorescence detection, capillary electrophoresis is a very powerful analytical tool for the study of minute samples, including single human cancer cells. The cell-to-cell variation in protein expression will be of interest in cancer prognosis, in developmental biology, and in gene expression studies.

Acknowledgments

This work was funded by grants 1R21 CA1171, 1R21DK070317, and 1R01GM 071666 from the National Institutes of Health, grant DE-FG02-04ER63937 from Department of Energy, and a contact from MDS-Sciex.

References

1 V.C. Wasinger, S.J. Cordwell, A. Cerpa-Poljak, J.X. Yan, A.A. Gooley, M.R. Wilkins, M.W. Duncan, R. Harris, K.L. Williams, I. Humphery-Smith, *Electrophoresis* 1995, 16, 1090–1094.

2 S.C. Hall, S.M. Smith, F.R. Masiarz, V.W. Soo, H.M. Tran, L.B. Epstein, A.L. Burlingame, *Proc. Nat. Acad. Sci. (USA)* 1993, 90, 1927–1931.

3 S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold, *Nat. Biotechnol.* 1999, 17, 994–999.

4 M.P. Washburn, D. Wolters, J.R. Yates, *Nat. Biotechnol.* 2001, 19, 242–247.

5 J.Z. Zhang, K.O. Voss, D.F. Shaw, K.P. Roos, D.F. Lewis, J. Yan, R. Jiang, H. Ren, J.Y. Hou, Y. Fang, X. Puyang, H. Ahmadzadeh, N.J. Dovichi, *Nucleic Acids Res.* 1999, 27, e36.

6 X. Huang, M.J. Gordon, R.N. Zare, *Anal. Chem.* 1988, 60, 375–377.

7 D.S. Burgi, R.L. Chien, *Methods Mol. Biol.* 1996, 52, 211–226.

8 D.S. Burgi, *Anal. Chem.* 1993, 65, 3726–3729.

9 C. Gelfi, M. Curcio, P.G. Righetti, R. Sebastiano, A. Citterio, H. Ahmadzadeh, N.J. Dovichi, *Electrophoresis* 1998, 19, 1677–1682.

10 JW. Jorgenson, K.D. Lukacs, *Anal. Chem.* 1981, 53, 1298–1302.

11 P.K. Jensen, L. Pasa-Tolic, G.A. Anderson, J.A. Horner, M.S. Lipton, J.E. Bruce, R.D. Smith, *Anal. Chem.* 1999, 71, 2076–2084.

12 D.M. Pinto, E.A. Arriaga, D. Craig, J. Angelova, N. Sharma, H. Ahmadzadeh, N.J. Dovichi, C.A. Boulet, *Anal Chem.* 1997, 69, 3015–3021.

13 I.H. Lee, D. Pinto, E.A. Arriaga, Z. Zhang, N.J. Dovichi, *Anal. Chem.* 1998, 70, 4546–4548.

14 J.Y. Zhao, K.C. Waldron, J. Miller, J.Z. Zhang, H.R. Harke, N.J. Dovichi, *J. Chromatogr.* 1992, 608, 239–242.

15 D. Richards, C. Stathakis, R. Polakowski, H. Ahmadzadeh, N.J. Dovichi, *J. Chromatogr. A* 1999, 853, 21–25.

16 S. Hjerten, M.D. Zhu, *J. Chromatogr.* 1985, 346, 265–270.

17 R. Rodriquez-Diaz, T. Wehr, M. Zhu, *Electrophoresis* 1997, 18, 2134–2144.

18 J. Wu, S.C. Li, A. Watson, *J. Chromatogr. A* 1998, 817, 163–171.

19 X.W. Yao, F.E. Regnier, *J. Chromatogr.* 1993, 632, 185–193.

20 S. Hu, Z. Zhang, L.M. Cook, E. Carpenter, N.J. Dovichi, *J. Chromatogr. A.* 2000, 895, 291–296.

21 Z. Zhang, S. Krylov, E.A. Arriaga, R. Polakowski, N.J. Dovichi, *Anal. Chem.* 2000, 72, 318–322.

22 S. Hu, R. Lee, Z. Zhang, S.N. Krylov, N.J. Dovichi *J. Chromatogr. B* 2001, 752, 307–310.

23 S. Hu, L. Zhang, L.M. Cook, N.J. Dovichi, *Electrophoresis* 2001, 22, 3677–3682.

24 S. Hu, J. Jiang, L. Cook, D.P. Richards, L. Horlick, B. Wong, and N.J. Dovichi, *Electrophoresis* 2002, 23, 3136–3142.

25 S. Hu, L. Zhang, S.N. Krylov, N.J. Dovichi, *Anal. Chem.* 2003, 75, 3495–3501.

26 S. Hu, D. Michels, M. A. Fazal, C. Ratisoontorn, M.L. Cunningham, N.J. Dovichi, *Anal. Chem.* 2004, 76, 4044–4049.

27 M.M. Bushey, J.W. Jorgenson, *Anal. Chem.* 1990, 62, 161–167.

28 D. Michels, S. Hu, R. Schoenherr, M.J. Eggertson, N.J. Dovichi, *Mol. Cell. Proteomics* 2002, 1, 69–74.

29 D.A. Michels, S. Hu, K.A. Dambrowitz, M.J. Eggertson, K. Lauterbach, N.J. Dovichi, *Electrophoresis* 2004, 25, 3098–3105.

30 S. Hu, L. Zhang, R. Newitt, R. Aebersold, J.R. Kraly, M. Jones, N.J. Dovichi, *Anal. Chem.* 2003, 75, 3502–3505.

9
A DNA Microarray Fabrication Strategy for Research Laboratories

Daniel C. Tessier, Mélanie Arbour, François Benoit, Hervé Hogues, and Tracey Rigby

9.1 Introduction

DNA microarrays (also known as DNA chips or gene chips) are a powerful new tool for the study of gene expression and genetic variation. With the availability of increasing numbers of completely sequenced genomes, it is now possible to make DNA microarrays, on which all the genes of an organism are represented, enabling simultaneous assessment of the expression of all these genes. This technology has lead biomedical research into a new era of "discovery research" that is complementing the type of "hypothesis-driven research" that has marked the phenomenal success of molecular biology in the past four decades [1].

Pioneering studies of the chemistry of nucleic acids some four decades ago [2] showed that mRNA could be measured by hybridization to total bacteriophage DNA fixed on nitrocellulose filters by Coulombic forces. The characteristics of hybridization of RNA–DNA and DNA–DNA were established by extensive experimentation both in liquid and on solid supports (typically nitrocellulose membranes). The advent of restriction enzymes enabled the electrophoretic separation of DNA fragments and localization of mRNA to specific regions – Southern blots [3]. The concept of DNA chips arose from the coincidence of a number of technologies, including rapid oligonucleotide synthesis and genome sequencing, understanding of the enzymology of DNA synthesis, and the availability of such enzymes, robotic methods of arraying, refinement of DNA chemistry for creating fluorescent hybridization probes, and the availability of completely sequenced genomes. Thus these developments combined with the evolution of nucleic acid hybridization methods led to the realization that individual genes arranged separately and in order on a solid substrate could be used to monitor all the genes of an organism.

Essentially two techniques are used for production of DNA chips. The first employs direct spatially ordered synthesis of oligonucleotides on a solid support (silica) by a process called photolithography. This technology, patented by Affymetrix, is reminis-

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

cent of the technology used in the electronic industry for the fabrication of semiconductors. The density of features displayed on their GeneChips can now exceed 500,000 on an area ~1.28 cm^2. The other technology uses post-synthesis arraying of oligonucleotides or DNA on solid supports, generally optically flat glass slides, using either contact printing or ink-jet spotting. There are numerous variations in the details of both techniques but the methods based on post-synthesis arraying are clearly the most accessible to the smaller research lab. Although initial experiments used arrayed cDNA and genomic DNA clones on nitrocellulose or nylon supports and detection using ^{32}P or ^{33}P-labeled probes, the recent technological improvements mentioned above and the accessibility of precise arraying robots have made glass microscope slides increasingly popular substrates. For those familiar with conventional hybridization literature, the term "probe" was used for the species of nucleic acid that carried the detection system (usually radioactivity) present in the solution phase. For the sake of uniformity, this conventional terminology will be used throughout this chapter.

The main applications of DNA chips and microarrays are in gene expression profiling [4–8], mutation analysis [9, 10], detection of single-nucleotide polymorphisms (SNP) [11], pharmacogenomics [12], validation of drug targets, identification of tagged biological strains, and monitoring of microbial flora in soil and wastewater. Its potential does not need to be emphasized but this technology is not an end in itself, it is, first and foremost, a quantitative high-throughput method of screening using genetic material as a target.

Researchers face many hurdles in the custom fabrication and use of microarrays. The costs of microarrays are still high, the technology for their fabrication is precise, and the conditions for their use and analysis are numerous and demanding. Several cost-effective strategies can, however, be used by either large core facilities or academic centers to make high quality microarrays available to researchers at reasonable cost.

In many cases, fabrication of microarrays requires a source of available cloned cDNA or expressed sequenced tags (EST) for an organism. These are typically available from central clone repositories or commercial sources (a list of clone providers is given in Tab. 9.1). Plasmid DNA can be arrayed directly, but most frequently cloned DNA inserts are amplified by PCR using common oligonucleotide primers immediately flanking the multiple cloning cassette [13, 15]. In this way, although the initial cost of acquiring the clones might be high, the cost of oligonucleotides is low. An alternative strategy is to identify coding regions within genomic DNA, open reading frames (ORF), or exons and to amplify the selected DNA fragments using PCR and specific oligonucleotide primers. The advantage of this strategy is that for any sequenced genome all identified coding or potentially coding sequences can be arrayed. This also avoids a problem inherent to EST-based strategies whereby some transcripts may be extremely rare and absent from cloned libraries. All gene copies are present equally in genomic DNA. The use of a specific primer pair to amplify a selected gene segment also prevents the undesirable amplification of low-complexity sequence stretches (repeated sequences and poly(A) tails) from the 3′ untranslated regions (3′UTR) of cloned cDNA, which will invariably contribute to cross-hybridization of spots on the microarray [16, 17].

Table 9.1 List of suppliers*.

Manufacturers of high-throughput DNA synthesis instrumentation	
Applied Biosystems	www.appliedbiosystems.com
BioAutomation Corporation	www.bioautomation.com
Gene Machines	www.genemachines.com
Polygen	www.polygen.com
Manufacturers of liquid handlers	
Beckman–Coulter	www.beckmancoulter.com
Packard Bioscience	www.packardbiochip.com
Qiagen	www.qiagen.com
Tecan	www.tecan.com
Tomtec	www.tomtec.com
Zymark	www.zymark.com
Providers of oligonucleotides, clone sets or oligonucleotide sets	
BD Biosciences	www.bd.com
I.M.A.G.E.	www.image.com
Incyte Genomics	www.incyte.com
Integrated DNA Technologies (IDT)	www.idtdna.com
Invitrogen	www.invitrogen.com
Genset Oligos	www.genset.com
MWG Biotech	www.mwg-biotech.com
Operon Technologies	www.operon.com
Pierce	www.piercenet.com
Proligo	www.proligo.com
Qiagen	www.qiagen.com
Research Genetics	www.resgen.com
Sigma Genosys	www.genosys.com
Stratagene	www.stratagene.com
Tm Bioscience Corporation	www.tmbioscience.com
Manufacturers of arraying robots	
Amersham Biosciences	www.amershambiosciences.com
BioGenex	www.biogenex.com
BioRad	www.bio-rad.com
BioRobotics	www.biorobotics.com
Cartesian Technologies	www.cartesiantech.com
Gene Machines	www.genemachines.com
GeneScan	www.genescan.com
Genetic Microsystems	www.geneticmicro.com
Genetix	www.genetix.com
Genomic Solutions	www.genomicsolutions.com
GeSiM	www.gesim.com

Hitachi Genetic Systems	www.miraibio.com
Intelligent Automation Systems	www.ias.com
Perkin Elmer Life Sciences	www.perkinelmer.com/lifesciences
TeleChem International (Quill pins)	www.arrayit.com
V&P Scientific	www.vp-scientific.com

Suppliers of glass and other substrates for microarrays

Amersham Biosciences	www.amershambiosciences.com
Apogent	www.nuncbrand.com
Corning	www.corning.com/cmt
Erie Scientific Company	www.eriesci.com
Exiqon	www.exiqon.com
Full Moon BioSystems	www.fullmoonbiosystems.com
NoAb Diagnostics	www.noabdiagnostics.com
Quantifoil	www.quantifoil.com
Schleicher and Schuell	www.s-and-s.com
Schott Nexterion	www.schott.com/nexterion
Sigma–Aldrich	www.sigma-aldrich.com
SurModics	www.surmodics.com
TeleChem International	www.arrayit.com

Providers of premade or custom microarrays

Affymetrix	www.affymetrix.com
Agilent Technologies	www.agilent.com
Amersham Biosciences	www.amershambiosciences.com
BD Biosciences	www.bd.com
Corning Life Sciences	www.corning.com
Genmed Biotechnologies	www.genmed.com
Genomic Solutions	www.genomicsolutions.com
Incyte Genomics	www.incyte.com
MetriGenix	www.metrigenix.com
Operon	www.operon.com
Perkin Elmer Life Sciences	www.perkinelmer.com/lifesciences
Spectral Genomics	www.spectralgenomics.com
Xeotron	www.xeotron.com

Suppliers of high-throughput automated slide-handling systems

Advalytix AG	www.advalytics.com
Amersham Biosciences	www.amershambiosciences.com
Gene Machines	www.genemachines.com
Genomics Solutions	www.genomicsolutions.com
Perkin Elmer Life Sciences	www.perkinelmer.com/lifesciences
Thermo Hybaid	www.thermohybaid.com
Ventana	www.ventanadiscovery.com

Suppliers of microarray scanners and data analysis packages

Affymetrix	www.affymetrix.com
Agilent Technologies	www.agilent.com
Alpha Innotech	www.alphainnotech.com
Applied Precision	www.appliedprecision.com
Axon Instruments	www.axon.com
BioDiscovery	www.biodiscovery.com
BioGenex	www.biogenex.com
BioRad	www.bio-rad.com
GeneFocus	www.genefocus.com
Genetic Microsystems	www.geneticmicro.com
Genomics Solutions	www.genomicsolutions.com
Hitachi Genetic Systems	www.miraibio.com
Imaging Research	www.imagingresearch.com
Incyte Genomics	www.incyte.com
Informax	www.informaxinc.com
Iobion Informatics	www.iobion.com
Lynx Therapeutics	www.lynxgen.com
Media Cybernetics	www.mediacy.com
Molecular Dynamics	www.mdyn.com
Molecularware	www.molecularware.com
NetGenics	www.netgenics.com
OmniViz	www.omniviz.com
Perkin Elmer Life Sciences	www.perkinelmer.com/lifesciences
Research Genetics	www.resgen.com
Rosetta Biosoftware	www.rosettabio.com
Silicon Genetics	www.sigenetics.com
Spectral Genomics	www.spectralgenomics.com
Spotfire	www.spotfire.com
Stanford University	rana.lbl.gov
The Institute for Genomic Research	www.tigr.org/tdb/microarray
Thermo Hybaid	www.thermohybaid.com
Vysis	www.vysis.com

Procedures for microarrays

MicroArray Lab at BRI	www.bri.nrc.ca/microarraylab
National Human Genome Research Institute	www.nhgri.nih.gov/DIR/LCG/15K/HTML/protocol.html
Stanford University	www.cmgm.stanford.edu/pbrown
TeleChem International	www.arrayit.com
The Institute for Genomic Research	www.tigr.org/tdb/microarray
University of Toronto	www.oci.utoronto.ca/services/microarray

* Please note that this list of suppliers is far from complete but should provide the reader with a few options enabling familiarization with some of the tools available

In another strategy for microarray production, one or more unmodified or amino-modified oligonucleotide sequences of 60–80 nucleotides representative of each coding region can be arrayed directly on to glass slides. The immediate disadvantage of this strategy is the higher cost of oligonucleotide synthesis. The growth of large-scale genomics efforts including microarrays has fortunately driven the cost of oligonucleotides down in recent years. One advantage, however, of oligonucleotide arrays over amplicon arrays is that the target sequences are short and can be more easily selected to minimize cross-hybridization to multiple probe sequences.

The implementation of microarray technology also requires the installation of high-throughput, high-precision instrumentation to carry out the numerous tasks involved in the production of the microarrays, and a comprehensive bioinformatics platform to support the design, clone tracking, data collection, and analysis aspects of microarrays. In this chapter, we will not address the issues related to analysis of microarray data. We will describe instead the practical implementation of a microarray facility (www.bri.nrc.ca/microarraylab) using as an example the fabrication of *Candida albicans* cDNA microarrays.

9.2 The Database

Candida albicans is an opportunistic human fungal pathogen causing systemic and very often fatal infections in immuno-compromised individuals. The application of microarray technology is a powerful new tool in the study of the mechanism of pathogenesis, and we were particularly interested in the molecular events involved in the dimorphic transition from yeast to hyphae when *Candida albicans* is exposed to changes in medium composition or environmental conditions (see reviews by Whiteway [18], Berman and Sudbery [19], Cowen et al. [20], Nantel et al. [21], and Enjalbert et al. [22]). The genome of *Candida albicans* (strain SC5314) contains approximately 20 million bases organized in eight pairs of chromosomes. The latest assembly of the genome, originally sequenced by Ron Davis (www. sequence.stanford.edu/-group/candida, see acknowledgement in References), has now been reduced to a mere 266 contigs and 6354 genes (http://genome-www.stanford.edu/fungi/Candida/ and Nantel et al., personal communication, www.candida.bri.nrc.ca). Open reading frames (ORF) were originally identified using an automated genome annotation system called MAGPIE (multipurpose automated genome project investigation environment) [23], but many other programs can now be used to do this. Essentially, software tools scan for any of the three termination codons or STOP codons (TAA, TAG, and TGA) in all six possible reading frames and then read the sequence backwards until they reach an in-frame "ATG" initiation codon. These open-reading frames, which represent potential coding regions, were automatically searched against GenBank to identify DNA and/or protein similarities.

One of the challenges of chip fabrication is data handling. For this we developed an integrated bioinformatics platform (www. bri.nrc.ca/bridna) and a set of tools to perform primer design, assessment of amplicon quality, tracking of the position of PCR products on the chips, and links to data analysis packages (Fig. 9.1). The first step in our fabrication process was to select those ORFs greater than 250 base pairs (>80 amino acids) and to identify short unique primer sequences (20 nucleotides) within each ORF. For *C. albicans* only 46 % of the original 14,400 ORFs identified were greater

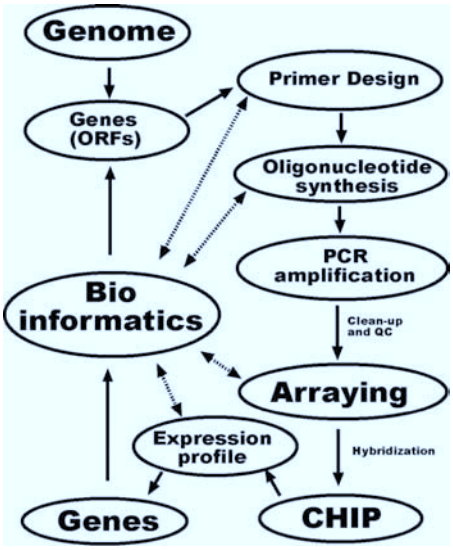


Fig. 9.1 Flowchart demonstrating the pivotal position of an integrated bioinformatics database in tracking sequence information from a genome all the way to microarrays and linking expression-profiling data back to individual genes (www.bri.nrc.ca/bridna).

than 250 base pairs. The primer identification tool takes into account the length of the primer, its location around the initiation and termination codons, and melting temperature (T_m) while considering potential secondary binding sites, hairpin formation, and primer dimerization energies [14]. The most important aspect of primer design was to limit the length of the amplicons to 250–800 base pairs to minimize potential homologies between each individual amplicon and the bulk of the genes in the organism, and to reduce the variance in the number of molecules when arraying equal masses of DNA. Sequence motifs are present and abundant among gene families and often result in labeled cDNA hybridizing non-specifically to other targets on the microarray thus biasing expression levels and profiles [16].

Hughes and coworkers [24] have performed a quite systematic assessment of the relationship between the length of the target sequence and the specificity/sensitivity of hybridization under different experimental conditions. Using *in-situ* synthesized oligonucleotide arrays they concluded that regions exhibiting perfect homology over more than 40 nucleotides in the target sequences immobilized on the microarray lead to significant cross-hybridization of labeled cDNA probes under classical hybridization conditions [24]. We have applied their conclusion to our fabrication process to increase the quality and reliability of our microarrays. We ran BLAST searches for each candidate amplicon to the entire ORF database to eliminate or minimize the number of hits showing perfect matches greater than 40 bases. When an amplicon was rejected, a new amplicon was designed for that gene once again to reduce the amount of homology to other ORF. It is sometimes impossible to find regions that will discriminate between highly conserved gene family members, even with oligonucleotide arrays. This rational chip design (RCD) strategy in our hands has significantly increased the value of the experimental data obtained with our arrays (Hervé Hogues, manuscript in preparation).

For ORF longer than 1 kb the primer closest to the START codon was positioned within ~800 base pairs of the STOP codon before applying the rules of RCD, thus favoring the 3′ end of the gene. The rationale was that cDNA tend to be over-represented in 3′ sequences as a consequence of using oligo(dT) priming at the 3′ end of mRNA during the probe amplification process. When appropriate amplicons were chosen to represent each ORF, primers closest to the START codon within each ORF were classified as Forward primers and primers closest to the STOP codon were classified as

Reverse primers. The production database was designed to assign batches of 96 Forward primers directly to 96-well plates for oligonucleotide synthesis. The Reverse primers corresponding to the same 96 ORF were assigned to a second 96-well plate such that the Forward/Reverse primer pairs would superimpose well for well for every ORF. This was to facilitate the automation of all the subsequent liquid handling steps in the production process.

In addition to ORF identification and PCR primer design, the microarray production database automatically redesigns primers in the case of PCR failures or weak bands. It systematically assigns and tracks the origin of oligonucleotides and/or amplicon products from 96-well plates to 384-well plates all the way to their position on the microarrays and generates a key (XY coordinates for every spot) of the genes based on the spotting method used by the arrayer (number of pins, size of the array, number of samples, etc.). One key component in any tracking scheme is the use of bar-coding. The large number of samples involved in any genomic/microarray project makes this procedure absolutely essential and simplifies the quality-control procedures.

9.3 High-throughput DNA Synthesis

Accessibility to high-quality oligonucleotides at a reasonable cost is an important consideration for a microarray fabrication facility. A preferred approach is to use synthetic oligonucleotides to amplify ORF directly from genomic DNA. This latter approach, however, necessitates the synthesis of thousands of oligonucleotides per genome, creating a demand for a large number of oligonucleotides.

There are also situations in which one might prefer to use oligonucleotides directly immobilized on the microarray in preference to amplicons. For example, when developing diagnostic microarrays for a large number of pathogens it is much easier, faster, and safer to design oligonucleotides from available genomic sequences than to obtain the various organisms, culture them to prepare genomic DNA under appropriate biosafety conditions, and prepare the desired amplicons (a list of providers of oligonucleotide sets is given in Tab. 9.1). Another situation is when creating microarrays intended to distinguish hybridization events between very closely related genes or organisms. Specificity is then of utmost importance to discriminate subtle polymorphisms.

9.3.1 Scale and Cost of Synthesis

The current technology of oligonucleotide synthesis, based on automated synthesizers using solid-phase phosphoramidite chemistry [25], offers a reasonable solution for the preparation of large numbers of oligonucleotides at a reasonable cost. The scale of synthesis is not a problem, inasmuch as current oligonucleotide synthesis technology generally prepares far more material than is normally required for PCR amplifications. As an example, a 5-nmol-scale synthesis, small by modern standards, produces enough material for 200 PCR reactions, given the usual quantity of 25 pmol per reaction. Even smaller scale could be envisaged if automated mechanical and fluidic devices were capable of efficiently delivering very small volumes of reagents and excluding oxygen and moisture from the synthesis process.

A 5-nmol synthesis scale is also adequate for many applications using immobilized oligonucleotide microarrays, unless large production runs are envisaged. Spotting concentration for amino-linked oligonu-

cleotides is in the range 10–25 nmol mL^{-1} (SurModics/Motorola), so a 5-nmol synthesis would provide 200–300 µL sample, enough to print several hundred slides.

Reagent costs alone for oligonucleotides produced on modern high-throughput DNA synthesizers currently lie in the range of US $0.05–0.10 per base. This translates into a cost per oligonucleotide of between US $1.00–2.00 (for 20 bases). By way of an example, the oligonucleotide cost for a microarray covering a full bacterial genome (~3000 ORF, ~6000 oligonucleotides) would be between US $7000 and $14,000, assuming a 15 % failure rate in the PCR.

9.3.2 Operational Constraints

The operation of a high-throughput DNA synthesizer capable of producing >500 oligonucleotides per day involves handling significant quantities of flammable, corrosive, highly reactive, and carcinogenic materials. These chemicals are hazardous and should be handled by a trained chemist. Several of the reagents are also of limited stability once installed on the synthesizer; a chemist must pay careful attention to their age and condition to maintain the highest quality while simultaneously keeping costs under control.

An important advantage lies with those synthesizers which accept the use of universal synthesis support (controlled-pore glass, CPG). Conventional DNA synthesizers use membranes or CPG columns bearing one of the four bases A, C, G, or T. This is fine for low numbers of syntheses but would be intolerable in a 96-well plate format. The time saved and the avoidance of possible errors fully justifies the use of the universal supports, in which the first base is added automatically by the synthesizer as opposed to being present on the solid phase.

One commercially available high-throughput oligonucleotide synthesizer is the LCDR/MerMade (BioAutomation Corporation, www.bioautomation.com, a list of high-throughput DNA synthesizers currently commercially available is available in Tab. 9.1). The MerMade is a Liquid Chemical Dispensing Robot that was adapted to perform all the operations of DNA synthesis using classical phosphoramidite chemistry [25] in a fully automated fashion. The MerMade is equipped with a motorized XY table within a closed argon chamber providing the inert atmosphere necessary for synthesis. The XY table can hold up to two filter plates (96-well format) in which a universal control pore glass (CPG) support is loaded to enable solid-phase synthesis of the oligonucleotides. The instrument is equipped with computer-controlled valves to deliver the reagents from the bottles to the injection heads. The newer generations of instruments can be set to operate on any scale of synthesis with minimum adjustment. The setup time for synthesis is routinely about 1 h and, depending on the synthesis procedure, the MerMade will synthesize 192 oligonucleotides (two 96-well plates of 20mers) in as little as 8 h without operator intervention. The yield of fully deprotected oligonucleotides is >80 % for 20mer products. Oligonucleotide products up to 70 bases can be routinely obtained by making minor adjustments to the synthesis procedures.

Our microarray facility has operated two LCDR/MerMade synthesizers with good success since March 2000. Using these synthesizers, one operator can produce up to 700 oligonucleotides per day with a success rate better than 98 %, determined by electrospray mass spectrometry. The instruments have no major flaws and provide a good level of user serviceability, convenience, and performance. As alluded to

above, the economics of high-throughput in-house DNA synthesis compared with out-sourcing to commercial oligonucleotide providers are becoming more difficult to justify.

9.3.3 Quality-control Issues

Quality control of oligonucleotides has always been a difficult problem. Polyacrylamide gel electrophoresis (PAGE) is inexpensive and adequate for small numbers of syntheses, but far too labor-intensive for daily productions of hundreds of oligonucleotides. High-performance liquid chromatography (HPLC) techniques suffer from a throughput problem. To achieve the throughput the entire procedure would have to be completed in less than 2 min per oligo, which is virtually impossible with ion-exchange chromatography. On the positive side, HPLC still remains the most suitable method for resolution and purification of $n-1$ failure sequences. A similar throughput problem also affects capillary electrophoresis (CE) methods, although equilibration is more rapid in CE and automatic sampling could possibly fulfill the requirement. Many core facilities and commercial providers of oligonucleotides are now using mass spectrometry (MALDI–TOF) to determine the purity and assess the quantity of their oligonucleotides; unfortunately, this method requires a desalting step, because the salt ions and the abundance of impurities affect the efficacy of electrospray ionization. We have recently tested a coupled LC–MS system (Agilent 1100 Series LC–MSD) and have come to realize that because of the orthogonal spray geometry of the instrument, there is no need to desalt the MerMade-synthesized products before MS. Furthermore, we have been able to conduct analyses in less than 1 min per sample thus comfortably achieving the desired throughput of 700 samples per day with minimum operator intervention. The products can be sampled directly from 96-well plates and the instrument has precision better than 1 Dalton. Oligonucleotide standards are run in parallel for the purpose of quantitation. We realize, however, that the last three QC methods (HPLC, CE, and MS) require important infrastructure investments. Here again, the economics of out-sourcing to commercial providers require careful consideration.

9.4 Amplicon Generation

Plates (96-wells) containing Forward and Reverse PCR primers have been used to set up PCR reactions using Biomek2000 and/or Biomek F/X workstations (Beckman–Coulter; other manufacturers of liquid handlers are listed in Tab. 9.1). *Candida albicans* ORFs were amplified directly from 100 ng genomic DNA and the amplicons were purified on 96-well ArrayIt SuperFilter plates (TeleChem/ArrayIt) and/or MultiScreen FB plates (Millipore) to eliminate unincorporated triphosphates, salts, and primers. This purification step, although not absolutely necessary [26], significantly improves the subsequent binding of the amplified DNA to the glass slides. After purification, the products were analyzed by agarose gel electrophoresis.

We have developed BandCheck, a unique bioinformatics tool, to assess the quality of PCR products after agarose gel electrophoresis. It creates a virtual band pattern predicted from the database and this is superimposed on a digitized TIFF image of the actual gel. The tool enables annotation of 96 PCR amplifications in less than 5 min (Fig. 9.2). The tool compensates for anoma-

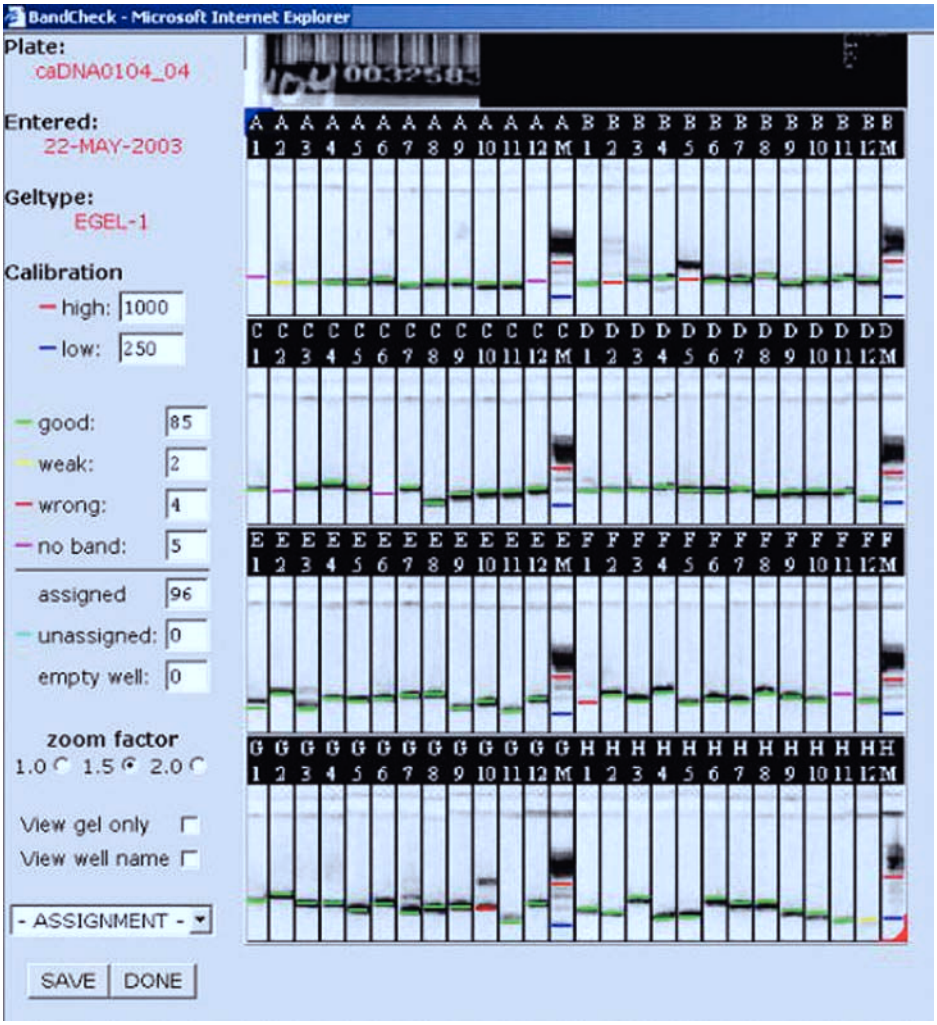


Fig. 9.2 BandCheck, the gel annotation tool of our microarray production database, was developed to assess the quality of PCR products after agarose gel electrophoresis. A TIFF image of the gel is uploaded into the database, the expected band pattern is superimposed on to the gel, and markers are positioned over designated bands in the marker lanes for calibration. Clicking the virtual objects superimposed on the PCR bands on the gel defines these products as either good, weak, wrong, or absent. These annotations are transferred back to the database and tracked all the way to the spotting on the microarrays or redirect a failed PCR for synthesis of a new pair of gene-specific primers.

lous migrations and uses any molecular weight standards as a reference. PCR amplifications were scored and an overall success rate of ~85 % was obtained for *Candida* genomic PCR. New oligonucleotide primers were automatically redesigned and synthesized to cover the 15 % PCR failures or weak bands.

Each genomic DNA amplicon was quantified by $O.D._{260nm}$ and yielded an average of 4 µg DNA per 100 µL PCR reaction. Amplified DNA were lyophilized in 96-well V-bottom plates and reconstituted in spotting buffer at a concentration such that >90 % of all the products were between 0.1–0.2 µg μL^{-1}. Product from four 96-well plates was then transferred to 384-well V-bottom plates for spotting. With such yields, a single PCR amplification would be sufficient to print a few thousand slides.

9.5 Microarraying

Many microarraying robots are now available on the market. Contact printing via a variety of pins (solid pins, split pins, pin and ring system, glass or quartz tubes) and non-contact ink-jet devices (crystal activated piezoelectric or syringe pump and solenoid valve tips) are among the most popular technology used for applying small volumes of target molecules on the surface of the glass substrate in an ordered fashion. One contact-printing instrument was developed in a collaboration between engineers at the University of Toronto and biologists at the Ontario Cancer Institute and is available as the SDDC-2, now known as the ChipWriterPro (Engineering Services, now BioRad; www.esit.com, www.biorad.com; other manufacturers of arraying robots are listed in Tab. 9.1). This particular arrayer is equipped with a print head capable of holding up to 48 pins (TeleChem/ArrayIt, quill pins) and the surface platen enables simultaneous printing of up to 75 slides. Integrated inside the spotting chamber are a circulating and sonicating waterbath, and a vacuum station to clean the pins between samplings. The chamber is under a slight positive pressure of HEPA-filtered air to minimize dust and particulate matter. It is also temperature- and humidity-controlled to optimize spot morphology and minimize sample evaporation in the 96-well or 384-well source plates and on the slides. The pins enable delivery of sub-nanoliter volumes of solution at densities exceeding 2500 spots cm^{-2}. This translates into ~30,000 spots on 25×75 mm^2 microscope slides. Spotting buffer and slide surface chemistry are key aspects of successful printing and later hybridization of microarrays. We have found that 50 % DMSO (occasionally with 0.05 % SDS for smoother spot morphology) contributes to some denaturation of the DNA thus increasing the number of single-stranded molecules available for hybridization [13]. DNA solutions stored in DMSO buffer are, moreover, less prone to evaporation, because of the intrinsic hygroscopic properties of this solvent. Although there are now many vendors of derivatized glass and other substrates for microarrays (Tab. 9.1), we opted for the CMT-GAPS slides (Corning) for our cDNA microarrays. These aminopropyltrimethoxysilane-coated glass slides in concert with a DMSO spotting buffer have been the most consistent in our hands giving more uniform spot morphology, better signal intensity and lower background.

9.6 Probing and Scanning Microarrays

A typical transcription profiling microarray experiment is designed to compare a test condition and an experimental condition. To illustrate a classical example, total or mRNA is extracted and purified from biological samples, and cDNA are synthesized in vitro using reverse transcriptase to incorporate a specific fluorescently labeled nucleotide analogue [27] corresponding to the

test condition (e.g. Cyanine3) and another for the experimental condition (e.g. Cyanine5). The use of multi-color fluorescent labels enables simultaneous analysis of two or more biological samples or states in a single experiment. After the reverse transcriptase reaction, fluorescently labeled cDNA are purified to eliminate unincorporated fluorescent dye and left to hybridize with the surface of the microarray in a volume rarely exceeding 100 µL. Cover slips or, especially, sealable contraptions are used to prevent dehydration of the probe and the entire arrangement is placed in a humid chamber for the duration of the hybridization. When hybridization is complete, unreacted probe is washed from the surface of the glass slide and the hybridized probe molecules are visualized by fluorescence scanning. Detailed procedures for probe preparation and clean-up, prehybridization, hybridization, and washes have been given by Nelson and Denny [28] and Hegde et al. [13].

Commercial kits for indirect labeling of cDNA using aminoallyl derivatives of nucleotides (Stratagene) or tyramide signal amplification (TSA) (Perkin–Elmer Life Sciences [29]) and chemical labeling (Amersham Biosciences, Kreatech, Panvera) have provided solutions to obtaining probes with higher specific activity and thus more sensitive detection of low-abundance mRNA. Linear amplification of mRNA based on the Eberwine method is also very efficient, because researchers are constantly reducing the number of cells in their starting biological material (Refs. [30] and [31] and Arcturus).

CCD cameras and confocal scanning devices from different instrument companies are currently being used for microarray scanning (Tab. 9.1). Light emitted by the excited fluorophores on the surface of the slide is converted into an electrical signal by a photomultiplier tube and captured by a detector. Confocal scanning devices such as the ScanArray from Perkin–Elmer Life Sciences (www.perkinelmer.com; other manufacturers of scanning instruments are listed in Tab. 9.1) are capable of scanning at 5-µm resolution. All these instruments enable quantitation of fluorescent emissions from the different spots on the microarray. Irrespective of the scanning instrument used, four settings should be adjusted to control signal intensity and enable the detection of low-intensity spots: focus, scanning speed, laser power, and sensitivity of the photomultiplier tube (PMT). For each spot in the microarray corresponding to a different gene or coding region the quantitation software automatically compares and identifies the induced and repressed genes between the control and experimental conditions. Linking the data obtained from microarray scans to gene IDs is probably the most crucial aspect of the bioinformatics platform we have developed. Data collection, normalization and analysis are topics of their own and are beyond the scope of this chapter. Complete discussions and review articles are available elsewhere [13, 32, 33].

9.7 Conclusion

Microarrays have had a significant impact on the way genome research is organized and performed. The wider impact of this technology is limited by the cost of the commercial microarrays currently available and the relatively few species for which arrays are available. We show here that it is a feasible and cost-effective for even small laboratories to set up and produce high-quality custom microarrays in-house for their favorite genome. The keys to this flexibility are the ability to synthesize or have access to large numbers of high-quality oligonu-

cleotides at reasonable cost and to have an integrated informatics platform to track samples throughout the quality-control steps.

Acknowledgements

The authors would like to acknowledge the contributions of Drs Roland Brousseau, Bill Crosby, and David Y. Thomas for their vision and guidance in the original set-up of the MicroArray Lab at BRI in 1999. This paper was published as NRCC publication number 42991. Sequence data for *Candida albicans* were obtained from the Stanford DNA Sequencing and Technology Center website at www-sequence.stanford.edu/-group/candida. Sequencing of *Candida albicans* was accomplished with the support of the NIDR and the Burroughs Wellcome Fund.

References

1 Ramsay, G. (1998). DNA chips: State-of-the-art. Nat. Biotechnol. 16: 40–53.

2 Nygaard, A.P. and Hall, B.D. (1964). Formation and properties of RNA–DNA complexes. J. Mol. Biol. 9:125–142.

3 Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503–517.

4 Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470.

5 Schena, M. and Shalon, D. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. USA 93:10614–10619.

6 Lockhart, D.J., Dong, H., Bynre, M.C. Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. 14:1675–1680.

7 DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M. (1996). Use of a cDNA microarray to analyze gene expression patterns in human cancer. Nat. Genet. 14:457–460.

8 Heller, R.A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D.E. and Davis, R.W. (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc. Natl. Acad. Sci. USA 94:2150–2155.

9 Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S. and Fodor, S.P. (1996). Accessing genetic information with high-density DNA arrays. Science 274:610–614.

10 Cronin, M.T., Fucini, R.V., Kim, S.M., Masino, R.S., Wespi, R.M. and Miyada, C.G. (1996). Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum. Mutat. 7:244–255.

11 Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lander, E.S. et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082.

12 Schena, M. and Davis, R.W. (1999). Genes, genomes and chips. In DNA Microarrays: A Practical Approach. Ed. M. Schena. Oxford University Press, New York. pp. 1–16.

13 Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Earle-Hughes, J., Snesrud, E., Lee, N. and Quackenbush, J. (2000). A concise guide to cDNA microarray analysis. Biotechniques 29:548–562.

14 Gordon, P.M. and Sensen, C.W. (2004). Osprey: a comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequency and microarrays. Nucleic Acids Res. e133.

15 Gaspard, R., Dharap, S., Malek, J., Qi, R. and Quackenbush, J. (2001). Optimized growth conditions for direct amplification of cDNA clone inserts from culture. Biotechniques 31:35–36.

16 Evertsz, E.M. Au-Young, J., Ruvolo, M.V. Lim, A.C. and Reynolds, M.A. (2001). Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. Biotechniques 31:1182–1192.
17 Murray, A.E., Lies, D., Nealson, K., Zhou, J. and Tiedje, J.M. (2001). DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc. Natl. Acad. Sci. USA 98:9853–9858.
18 Whiteway, M. (2000). Transcriptional control of cell type and morphogenesis in *Candida albicans*. Curr. Opin. Microbiol. 3:582–588.
19 Berman, J. and Sudbery, P.E. (2002). *Candida albicans*: A molecular revolution built on lessons from budding yeast. Nat. Rev. Genet. 3:918–930.
20 Cowen, L.E., Nantel, A., Whiteway, M.S., Thomas, D.Y., Tessier, D.C., Kohn, L.M. and Anderson, J.B. (2002). Population genomics of drug resistance in *Candida albicans*. Proc. Natl. Acad. Sci. USA 99:9284–9289.
21 Nantel, A., Dignard, D., Bachewich, C., Harcus, D., Marcil, A., Bouin, A.-P., Sensen, C.W., Hogues, H., van het Hoog, M., Gordon, P., Rigby, T., Benoit, F., Tessier, D.C., Thomas, D.Y. and Whiteway, M. (2002). Transcription profiling of *Candida albicans* cells undergoing the yeast to hyphal transition. Mol. Biol. Cell 13:3452–3465.
22 Enjalbert, B., Nantel, A. and Whiteway, M. (2003). Stress-induced gene expression in *Candida albicans*: Absence of a general stress response. Mol. Biol. Cell 14:1460–1467.
23 Gaasterland, T. and Sensen, C.W. (1996). MAGPIE: Using multiple tools for automated genome interpretation. Trends Genet. 12:76–78
24 Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H. and Linsley, P.S. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat. Biotechnol. 19:342–347.
25 Caruthers, M.H., Barone, A.D., Beaucage, S.L., Dodds, D.R., Fisher, E.F., McBride, L.J., Matteucci, M., Stabinsky, Z. and Tang, J.Y. (1987). Chemical synthesis of deoxyoligonucleotides by the phosphoramidite method. Methods Enzymol. 154:287–313.
26 Diehl, F., Beckmann, B., Kellner, N., Hauser, N.C., Diehl, S. and Hoheisel, J.D. (2002). Manufacturing DNA microarrays from unpurified PCR products. Nucleic Acids Res. 30:e79.
27 Mujumdar, R.B., Ernst, L.A., Mujumdar, S.R., Lewis, C.J. and Waggoner, A.S. (1993). Cyanine dye labeling reagents: sulfoindocyanine succinimidyl esters. Bioconj. Chem. 4:105–111.
28 Nelson, S.F. and Denny, C.T. (1999). Representational differences analysis and microarray hybridization for efficient cloning and screening of differentially expressed genes. In DNA Microarrays: A Practical Approach. Ed. M. Schena. Oxford University Press, New York. pp. 43–58.
29 Heiskanen, M.A., Bittner, M.L., Chen, Y., Khan, J., Adler, K.E., Trent, J.M. and Meltzer, P.S. (2000). Detection of gene amplification by genomic hybridization to cDNA microarrays. Cancer Res. 15:799–802.
30 Phillips, J. and Eberwine, J.H. (1996). Antisense RNA Amplification: A linear amplification method for analyzing the mRNA population from single living cells. Methods 10:283–288.
31 Wang, E., Miller, L.D., Ohnmacht, G.A., Liu, E.T. and Marincola, F.M. (2000). High-fidelity mRNA amplification for gene profiling. 2000. Nat. Biotechnol. 18:457–459.
32 Eisen, M.B., Spellman, P.T., Brown, P.O. and Bostein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95:14863–14868.
33 Eisen, M.B. and Brown, P.O. (1999). DNA arrays for analysis of gene expression. Methods Enzymol. 303:179–205.

10
Principles of Application of DNA Microarrays

Mayi Arcellana-Panlilio

10.1
Introduction

The completion of the Human Genome Project with the release of "... >98 % of the gene-containing part of the human sequence finished to 99.99 % accuracy ... " in April 2003 [1] poses the challenge of putting this newfound information to work, to define how these sequences function. Although the traditional approach to biological questions has been to survey a relatively small number of genes at a time, it has long been recognized that genes act in concert with other genes, and often in separate dimensions of time and space. These molecular interactions are therefore best studied within the framework of the entire genome. The logistical challenges of this global approach are being overcome with the development of new tools, such as those provided by microarray technology, which can enable the analysis of thousands of genes in parallel by specific hybridization to a miniaturized, orderly array of DNA fragments. Because the human genome might comprise fewer than 40,000 protein-coding genes, well within the current capacity of a single array, elucidating the functional roles of all these genes becomes surprisingly attainable.

Since their first reported application in 1995 to analyze gene expression in *Arabidopsis* [2], microarrays have become a major investigative tool in life-science research. The numbers of publications retrieved by a "microarrays" keyword search of PubMed from 1995 to 2004 are plotted in Fig. 10.1; the figure shows the publication boom that took off in 1999 and continues apace. In late 2001, papers on protein (133 papers total by mid-2004), tissue (54 papers total), whole cell (12 papers total), and carbohydrate (7 papers total) arrays started to appear, suggesting diversification and broadening of the basic technology to encompass the interrogation of array elements other than nucleic acids. DNA microarrays, which constitute >95 % of all array papers to date, are the focus of this review.

The main objectives of this chapter are:

1. to discuss the principles of DNA microarray technology;
2. to describe how arrays are made and how they are used, focusing on the two-color interrogation of printed arrays and their analysis; and
3. to cite examples of how the technology has been used to address issues relevant to cancer research.

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

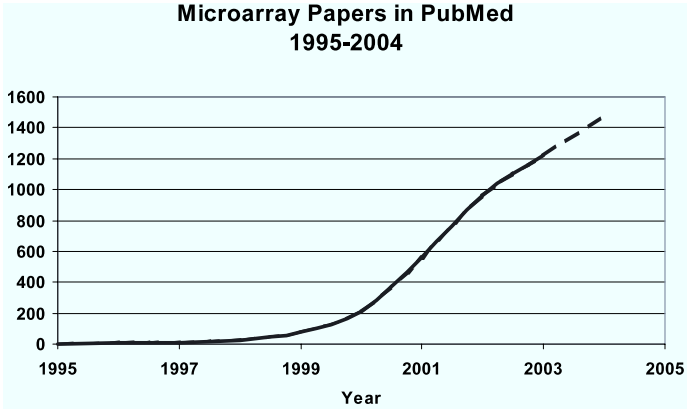


Fig. 10.1 Graphical representation of the number of publications retrieved by a "microarrays" keyword search of PubMed 1995–2004. The dotted line to the projected number of publications for 2004 is based on data at the end of June 2004.

10.2 Definitions

By definition, microarrays contain grids of up to tens of thousands of array elements presented in a miniaturized format. For DNA microarrays, those array elements or spots comprise minute amounts of DNA that have been either laid down robotically or synthesized *in situ* at precise locations on a solid support. These arrays are interrogated by allowing their immobilized sequences to hybridize by Watson–Crick base-pairing with labeled nucleic acids derived from the samples of study. The intensity of hybridization over individual spots is a measure of the amount of homologous sequence in the sample. Stated simplistically, when labeled cDNA is used to interrogate arrays, signal intensity can be directly related to the abundance of the RNA transcript, and when labeled genomic DNA is used, signal intensity can be related to gene copy number.

The nomenclature adopted by the microarray research community of calling the arrayed DNA the "probe" and the labeled nucleic acid population the "target" [3] will be used throughout this chapter.

10.3 Types of Array

DNA microarrays can be categorized according to their content, and so a microarray might represent the entire genome or a defined subset of it. The development of genome-wide arrays has mirrored the progress in annotating the respective genome databases, and the technical improvements that have enabled the production of increasingly higher-density arrays. Genome-wide arrays for the study of gene expression are usually designed to include all known sequences likely to code for protein. Early on, much effort went into ensuring that sets were nonredundant, and the UniGene database proved invaluable to this cause. Latter sets have been designed to incorporate other information, for example, alternative splicing to pick up transcript variants that might be expressed differentially. For example, the array-ready human genome sets from Operon have gone through several versions from a set of approximately 14,000 oligonucleotides designed against UniGene released in 2001, to 22,000 released in 2002, to its current set of 34,000 oligonucleotides based on the latest ENSEMBL database build. Genome-wide arrays for the study of genomic DNA might be designed to include promoter re-

gions and intervening sequences to address issues of gene regulation. The established method of chromatin immunoprecipitation to investigate DNA protein interactions, for example, has been combined with microarrays to identify novel targets of transcription factors [4].

Focused arrays are not designed to include the entire genome, and the genes or sequences included largely depends on the question being addressed. Thus, there are arrays to study specific pathways or gene networks, such as for apoptosis, oncogenes, and tumor suppressors (evaluated in Ref. [5]), and arrays of a specific cell type or tissue, such as the Colonochip designed to study colorectal cancer [6]. Focused arrays might also be designed to be part of a multitiered approach to studying gene expression, where their function becomes less as a means of discovery as an efficient means of validating initial results obtained on larger, more comprehensive arrays.

10.4 Production of Arrays

DNA microarrays are often categorized according to the technology platform used to produce them. The underlying concept of DNA microarrays has been, and continues to be, interpreted and realized in myriad ways, leading to the development of different platforms, each with its own strengths and weaknesses (reviewed in Ref. [7]).

10.4.1 Sources of Arrays

The DNA spots on a microarray are produced either by synthesis *in situ* or by deposition of pre-synthesized product. DNA synthesis *in situ* methods (e.g. Affymetrix's photolithography [8] and Agilent's inkjet technology [9]) have largely been within the purview of commercial companies. Rigorous production and quality assurance procedures ensure delivery of commercial arrays of consistently high quality. Ongoing R&D efforts ensure optimum design of the DNA content, and continued technological advancement enables the production of increasingly higher-density arrays. All of these come at a cost, however, which is not lost on prospective users. Predictably, academic pricing packages are wrought to woo the reluctant researcher. A highly successful strategy has been to involve the research community in the development of new arrays, giving birth to various consortia of laboratories that generate potentially useful information for themselves while providing companies with invaluable input.

For academic or government laboratories wishing to produce their own arrays, an option is to adopt the array printing methodology pioneered by the Brown laboratory of Stanford University (http://cmgm.stanford.edu/pbrown/mguide/index.html). The robotic deposition of nano amounts of presynthesized 60–70mer oligonucleotides or DNA fragments generated by PCR or by plasmid insert purification produces arrays that can be of very high quality at relatively low cost. Compared with commercially available arrays, perhaps the greatest advantage of spotted arrays is their flexibility, which can facilitate the rapid iteration of experiments, especially when testing specific hypotheses [10]. The sequences spotted and the number, whether or not there are replicate spots, and how the array is laid out are some of the design features that can be modified to suit experimental objectives. Production volume can be tailored to a specific demand, and can easily be adjusted to respond to changes. In the Southern Alberta Microarray Facility at the University of Calgary, for instance, we have produced

custom arrays according to the specification that the oligos be arranged on the array according to their genomic location, and the timing of delivery of those printed arrays has matched the researcher's needs.

Although hardly ever intimately involved with array production *per se*, users do well to be acquainted with the process, especially how production conditions can affect the quality of arrays. Details of the manufacture of spotted microarrays are discussed elsewhere in this volume (Chapt. 9) and what follows is a discussion of pertinent aspects.

10.4.2 Array Content

The choice of the type of DNA to print is fundamental. Longer stretches of DNA such as obtained from PCR amplification of cDNA clones produce robust hybridization signals. Because of their length, however, they are more forgiving of mismatches, which can affect specificity. When the finest discrimination is required, as when assessing single-nucleotide changes, short oligonucleotides (24–30 nt) are in order. Somewhat paradoxically, specificity can sometimes be problematic in very short oligos, and designing several to represent a single gene has been a means of circumventing this problem. Long oligonucleotides (50–70 nt) afford an excellent compromise between signal strength and specificity [11–13] and their use has increased in popularity among academic core facilities.

Needless to say, the design of the oligonucleotides is extremely important, and requires a fair amount of expertise and capital outlay for proper validation. Choosing oligos corresponding to the 3′ untranslated regions (3′ UTR) increases the likelihood of their being specific [14]. For a different reason, designing oligos close to the 3′ end might also boost signal intensity. This has to do with how labeling can be affected by the efficiency of RNA reverse transcription primed by oligo dT. Because the oligonucleotides are meant to be assayed in parallel on the array, under identical conditions of time, temperature, and salt concentration, their hybridization characteristics (e.g. their T_m, or melting temperature) should be as similar as possible. They should not have secondary structures or runs of identical nucleotides that could attenuate hybrid-formation with their labeled complement. Laboratories may design and validate oligonucleotides themselves or purchase pre-designed sets from commercial sources.

10.4.3 Slide Substrates

Glass microscope slides are the solid support of choice, and they should be coated with a substrate that favors binding of the DNA. The earliest substrates used were poly-lysine or aminopropylethoxysilane, established slide coatings for *in-situ* hybridization studies of gene expression. The need to address data quality has led to the development of new substrates on atomically flat (and sometimes even mirrored) slide surfaces with the promise of better DNA retention and minimum background for higher signal-to-noise ratios. Different versions of silane, amine, epoxy, and aldehyde substrates, which attach DNA by either ionic interaction or covalent bond formation, have become commercially available (e.g. see www.arrayit.com). The commercial arrays industry has seen the development of proprietary substrates to provide unique advantages, such as the highly efficient hybridization kinetics afforded by Amersham's three-dimensional Code Link surface, which the company uses for their own ar-

rays and offers for sale as a slide substrate (http://www5.amershambiosciences.com/). Laboratories that print arrays inevitably make their choices based on empirical data of how well substrates perform in their own hands.

10.4.4
Arrayers and Spotting Pins

The physical process of delivering the DNA to pre-determined coordinates on the array involves spotting pens or pins carried on a print head that is controlled in three dimensions by gantry robots with sub-micron precision. Although 30,000 features of ~90 μm diameter can easily be spotted on a 25 mm×75 mm slide, specification sheets from pin manufacturers (e.g. www.array-it.com/ and www.genetix.com/) report maximum spotting densities of over 100,000 features per slide for their newest pin designs. Improvements in arraying systems have included shorter printing times and longer periods of "walk-away" operation. Arrayers are invariably installed within controlled-humidity cabinets to maintain an optimum environment for printing.

From the user's perspective, arrays must be reliably consistent in quality from the outset. One has enough to deal with without having to worry about whether a particular spot on an array is what the array lay-out file says it is, or that today's array is missing several grids. It is ridiculously obvious that these should never become an issue for the user, because if they do, confidence in the supplier (and in extreme cases, even trust in the technology) is significantly eroded. Having said all this, however, the responsibility for making good use of well-made arrays rests ultimately on the user.

10.5
Interrogation of Arrays

Soon after arriving on the scientific scene, microarrays were recognized as a potentially powerful tool for the post-genomic era, and it became a priority of every major research institution to have the capability of running microarray experiments. At the most basic level, this entails gaining access to pre-made arrays and an array scanner to acquire data from the arrays after hybridization. Although this model works for individual laboratories, even this is not an inexpensive proposition, and it does not provide the option to custom print arrays for specific needs. The core technology laboratory model, with the ability to print arrays and to run experiments and acquire data, maximizes the use of resources, making arrays, materials, and specialized equipment, and technical and analytical support, available to researchers. At the University of Calgary, scientists in the Faculty of Medicine led the microarray initiative, and with establishment funds from a private donor, the Alberta Cancer Board, and the Alberta Heritage Foundation for Medical Research, the Southern Alberta Microarray Facility (SAMF) opened its doors in 2001. Core technology laboratories in Canada are listed on the website of the Woodgett lab at the University of Toronto (http://kinase.uhnres.utoronto.ca/CanArrays.html). Although there are sites such as BioChipNet (http://www.biochipnet.de) that seek to maintain a worldwide directory of companies and institutions involved in microarrays, the most current listing of microarray facilities with a web presence can be retrieved by using a global search engine.

Given the burgeoning supply of pre-made arrays and technical services that will even do the microarray experiments for clients,

becoming a user of the technology is easier than ever. Whether it is wiser than ever is the important issue, which every prospective user must think about and consider advisedly with regard to his or her specific research objectives. The question becomes: Are microarrays the best way to find out what I want to know?

10.5.1 Experimental Design

Well before the first animal is killed the experimental design for a microarray study has to be in place. In practical terms, the goal is to determine the type of array to run, how many, and which sample(s) will be hybridized to each slide to obtain meaningful data amenable to statistical analysis, upon which sound conclusions can be drawn. Unless one is well versed in statistics and the recent developments in statistical methods to deal with microarrays, it would be wise to consult a statistician with this expertise, and to do so early on (i.e. not to wait until after the data deluge) [15]. Smyth and coauthors [16] offer an excellent overview of the many statistical issues in microarrays.

An important aspect of experimental design is deciding how to minimize variation, which can be thought of as occurring in three layers: biological variation, technical variation, and measurement error [17]. The easy answer to dealing with variation is replication, but to make the best use of available resources it is important to know what to replicate and how many replicates to apply, which will depend on the primary objectives of the experiment. Thus, if the question pertains to determining the effects of a treatment on gene expression in a type of tumor in mice, for instance, it would be more valuable to test many mice (biological replicates) a few times than a few mice many times (technical replicates). In the latter experiment precision would be very high, and probably *true* for the few mice being tested. It is less certain how true that information would be of mouse tumors in general.

Usually, arrays containing inserts of cDNA clones or long oligonucleotides (e.g., from Agilent) are amenable to hybridization with two differently labeled samples at the same time to the same array, whereas short oligo (25–30 nt) arrays, including those from Affymetrix and Amersham, are hybridized to single samples. Exceptions to this are the 60mer Expression Arrays from Applied Biosystems (www.appliedbiosystems.com/) to which single samples are hybridized.

Hybridization of two samples to the same slide is made possible by labeling each sample with a chemically distinct fluorescent tag. The ideal tags to use for these two-color microarray experiments would behave the same in every way, except that they would fluoresce at wavelengths well separated on the spectrum to prevent cross talk. In the real world, however, the fluorescent tags currently in widest use (i.e. Cy 3 and Cy5), behave unequally in ways besides their optimum excitation wavelength, introducing a dye bias that should be dealt with in the experimental design. Although running dye-swap experiments has become routine, doing so immediately doubles the number of arrays required. There might be more efficient ways of dealing with dye bias, such as estimating and correcting for bias by extracting data from a pair of split-control microarrays, where control RNA is split, labeled with Cy3 and Cy5, and combined on the same array [18]. Dobbin and coauthors [19] suggest that balanced experimental designs obviate the need to run dye swaps for every sample pair. Labeling procedures are discussed in greater detail in the next section.

Although having two samples hybridizing to a slide might add a level of complexity

to the experimental design, it also provides the opportunity to make direct comparisons between samples of primary interest. If the aim is to identify genes that are differentially expressed in one sample versus another, it is more efficient to directly compare these two samples on one array than to make indirect comparisons through a reference sample and use two arrays in the process [16]. Yang and Speed [20] explain that the main difference between direct and indirect comparisons is the higher variance seen in indirect designs, basically from having to run two arrays instead of one, and therefore, whenever possible, direct comparisons are preferred. When there are more than two samples to compare and the pairings are of equal importance, direct comparisons can still be made between the samples, in so-called saturated designs [16]. Diagrams to illustrate these design options are shown in Fig. 10.2.

When looking at a large number of samples, however, comparison with a common reference becomes more efficient. In the now classic study by Alizadeh and coworkers [21] that identified sub-types of diffuse large B-cell lymphoma, 96 normal and malignant lymphocytes were analyzed on 128 microarrays and the reference RNA used was prepared from a pool of nine different lymphoma cell lines.

When an experiment is testing the effect(s) of multiple factors, a well-thought out design is extremely critical so that resources are not wasted on eventually useless comparisons. Discussion of these multifactorial designs is beyond the scope of this review; for details on these and the scenarios covered above, however, the reader is directed to the references already cited in this section plus several others [19, 22–26].

The next two sections focus on the preparation and labeling of sample pairs for hybridization to a gene expression array. Gene

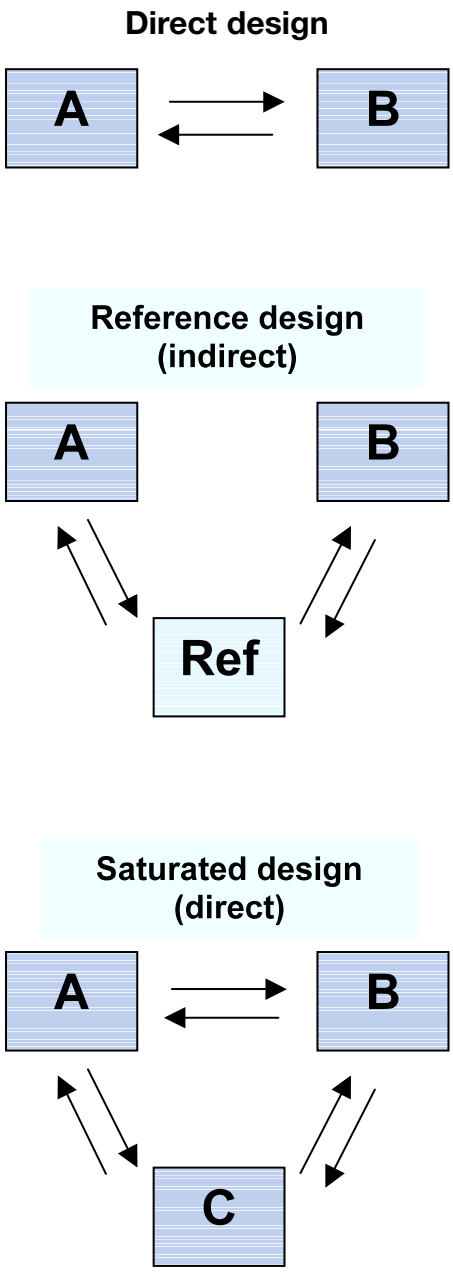


Fig. 10.2 Basic experimental designs for microarrays. Boxes identify samples being tested. The arrows indicate direction of labeling. Thus, double arrows in opposing directions indicate dye-swap experiments.

expression is measured in terms of the relative abundance of RNA, and reverse transcription of the RNA into cDNA enables the incorporation of fluorescent label whose intensity can be measured and related to transcript abundance.

10.5.2 Sample Preparation

In its allocation of resources, the experimental design will identify the sources of RNA for the samples that will be hybridized to a microarray. Those sources may be tissue culture cells, biopsies, or tissues of certain pathology. Each of these presents its own set of challenges where the isolation of RNA is concerned. Obtaining pure, intact RNA in the sense of its being free from DNA or protein contamination is certainly important and should never be discounted. There is, however, another layer of purity that must be emphasized, and that is the homogeneity of the RNA source itself as defined by the biological question being asked. When an experiment seeks to compare normal tissue and tumor, the normal RNA should be prepared from normal tissue, and the tumor RNA from tumor tissue. But tissue can be comprised of many different cell types and be very heterogeneous. Thus, unless cell type is the essence of the question, it should not be the reason for differences picked up by an experiment. Even so-called tumor samples can contain normal cells, and it becomes important to know how much of a tumor is really tumor, and to decide whether to go with the bulk tissue in hand or to isolate portions of interest with a scalpel or specific cells with laser capture microdissection [27, 28].

Although the preparation of RNA has been facilitated in recent years by the development of RNA isolation kits, basic principles and caveats set forth in early literature (for example Ref. [29]) remain true. Because of its chemical nature, RNA is notoriously prone to degradation by ribonuclease (-RNase) activity, and so a main concern of any laboratory doing RNA work is keeping these RNases at bay. The pancreatic RNases are particularly problematic, because they are ubiquitous and also extremely stable, working within a wide range of pH values and needing no cofactors. Their hardiness is legendary – a method of ensuring that RNase A for digestion of bacterial RNA in DNA plasmid preps is free from DNase activity involves boiling for 10 to 30 min [30]. The RNase A simply renatures after heating and so regains its activity.

The most widely used method of isolating RNA is based on the guanidinium thiocyanate–phenol–chloroform extraction procedure developed by Chomczynski and Sacchi [31]. Cultured cells instantly lyse due to the chaotropic action of the "TRI reagent" and undegraded RNA of high purity is extracted with relative ease. RNA from tissue samples is best obtained by quickly homogenizing the sample in the presence of at least a tenfold volume excess of reagent before proceeding to the separation of the aqueous and the organic phases. The RNA partitions into the aqueous phase, and because of the acidic pH of the extraction mixture, the DNA preferentially goes to the interphase. The quality of the RNA can be assessed by gel electrophoresis to visualize the ribosomal bands, which should migrate as sharp bands with intensities that vary relative to each other according to their expected abundance ratio. Aberrant intensity ratios, laddering, and/or smearing of the ribosomal bands suggest that the RNA is degraded and therefore ill-suited for use in an experiment. With the development of lab-on-a-chip technology, instruments, for example the Agilent 2100 bioanalyzer, capable of assessing quality and quantity of RNA

(and DNA, proteins, and other materials) have become available as a viable alternative to gel electrophoresis.

To ascertain that the RNA in a sample is what is being tested in a microarray experiment, any genomic DNA that might have come through the RNA-isolation process should be removed enzymatically with DNase that is free of RNase contamination. The DNase itself must then be inactivated or it will begin to degrade the cDNA product of reverse transcription in the succeeding steps. Although phenol–chloroform extraction reliably removes DNase [32], it reduces the overall yield of RNA.

The amount of RNA required per hybridization ranges from as little as 2–5 μg total RNA for short oligonucleotide arrays to 10–25 μg total RNA for spotted cDNA and long oligonucleotide arrays [7]. This requirement for microgram amounts of starting RNA can sometimes be difficult to meet as when dealing with tissue biopsies or single cells harvested from laser-capture microdissection.

In these circumstances it becomes necessary to amplify the RNA in the sample to obtain adequate amounts for labeling and hybridization to an array. In the linear amplification method of Eberwine and colleagues [33–35], the RNA is reverse transcribed in the presence of oligo-dT to which the T7 RNA polymerase promoter sequence has been added. The second strand is then synthesized and the double stranded molecule becomes amenable to in-vitro transcription by T7 RNA polymerase to produce large amounts of complementary RNA (cRNA) that can be labeled directly or further amplified. The earliest paper [33] reports an 80-fold molar amplification after one round; the latest procedures claim microgram yields from starting amounts of as little as 1–5 ng [36]. Valid concerns about fidelity and reproducibility of RNA amplification [37–40] have led to entirely rational recommendations that samples for comparison be amplified using identical procedures.

10.5.3 Labeling

Although there are ways of labeling the RNA itself, such as by attaching fluorophores to the N-7 position of guanine residues [41], reverse transcription into cDNA provides stable material to work with that, in principle, is a faithful representation of the RNA population in the sample. The cDNA product can then be labeled either directly or indirectly.

In the direct labeling procedure, fluorescently labeled nucleotide is incorporated into the cDNA product as it is being synthesized [2, 42]. As expected, the efficiency of incorporation for the labeled nucleotide is lower than for the unlabeled nucleotide, and labeling frequency cannot be predicted simply from concentration ratios. More pertinent to the issue of running two-color experiments, however, is that a difference in the steric hindrance conferred by different label moieties causes some labeled nucleotides to be more efficiently used than others, producing a dye bias in which one sample is labeled at a higher level overall than the other. Thus, Cy3-nucleotide tends to be incorporated at a higher frequency than Cy5. Surprisingly, this does not necessarily translate into a "better" labeled target. Indeed, Cy3-labeled cDNA tends to give higher background on an array compared to the Cy5, which might be indicative of prematurely terminated products sticking nonspecifically to the slide. The reported observation that Taq polymerase terminates chain extension next to a Cy-labeled nucleotide, so that a higher frequency of label incorporation results in shorter products [43], lends

some credence to the idea that reverse transcriptase might be similarly affected.

To prevent the dye bias introduced by direct incorporation of labeled nucleotide, the indirect labeling approach was developed [9, 44]. Here, RNA is reverse transcribed in the presence of a much less bulky aminoallyl-modified nucleotide that, more importantly, enables the chemical coupling of fluorescent label after the cDNA is synthesized. If the coupling reaction goes to completion, the frequency of labeling becomes independent of the fluorophore. In practice, dye bias is very much reduced by the indirect labeling method, but it is still not completely eradicated. Figure 10.3 shows parallel sche-

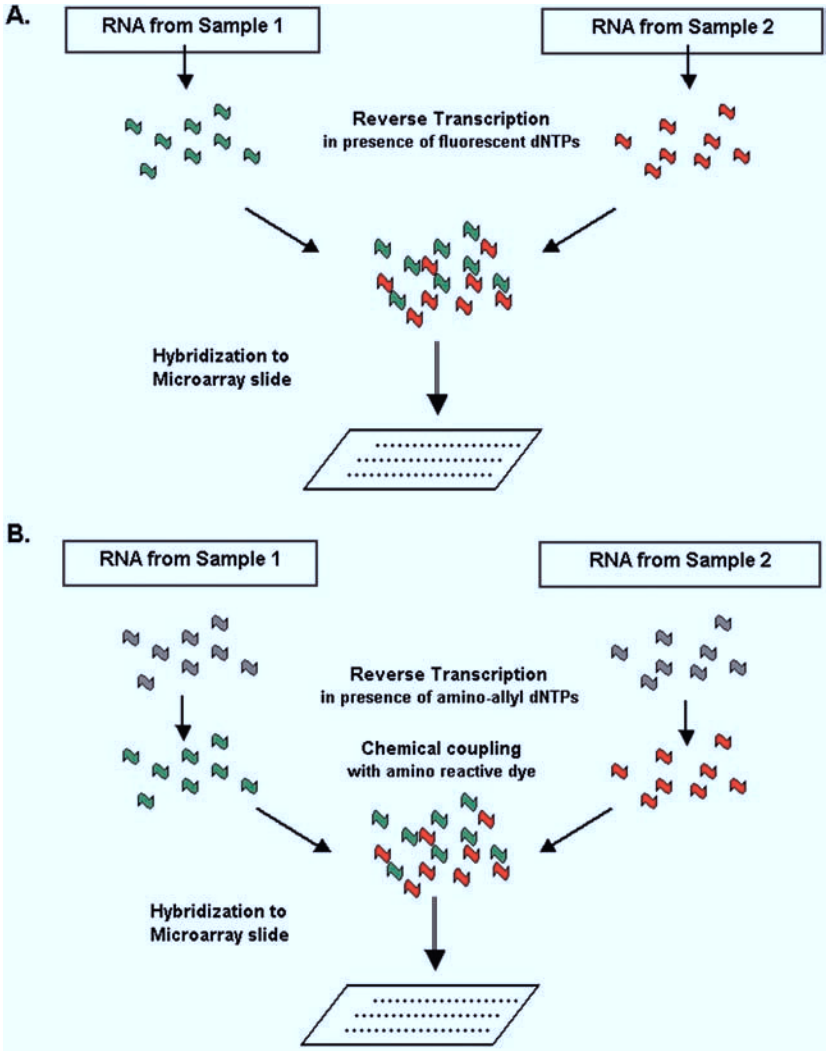


Fig. 10.3 Labeling schemes. (A) Direct labeling of cDNA by incorporation of fluorescent dNTP during reverse transcription. (B) Indirect labeling by incorporation of amino-modified dNTP at reverse transcription followed by chemical coupling of fluorescent dye.

ma for direct and indirect labeling of target cDNA for hybridization.

cRNA from RNA amplification procedures can be labeled with biotinylated ribonucleotides during *in-vitro* transcription of the double-stranded cDNA with T7 RNA polymerase [45]. Interestingly, Cy-nucleotides are not used because they are not good substrates for the RNA polymerase. To limit the effects of secondary structure it is customary to fragment the labeled cRNA before hybridization.

Exploring alternatives to Cy3 and Cy5 is worthwhile. Though historically Cy3 and Cy5 have been the fluorophores of choice, other options (such as the Alexa dyes from Molecular Probes) are becoming available that have similar excitation and emission characteristics, greater stability against photo-bleaching, and higher resistance to quenching artifacts.

10.5.4
Hybridization and Post-hybridization Washes

The hybridization step is literally where everything comes together. Printing the arrays, preparing and labeling the samples, all lead to this point where the labeled molecules find their complement sequences on the array and form double stranded hybrids strong enough to withstand stringent washes meant to put asunder nonspecific matches.

As in the hybridization of classical Southern and Northern blots, the objective is to favor the formation of hybrids and the retention of those which are specific. Thus, the labeled samples are applied to the microarray in the presence of high-salt buffers to neutralize repulsive forces between the negatively charged nucleic acid strands. Exogenous DNA (e.g. salmon sperm and C_0t-1 DNA) reduces background by blocking areas of the slide with a general affinity for nucleic acid or by titrating out labeled sequences that are nonspecific. Denhardt's reagent (containing equal parts of Ficoll, polyvinylpyrrolidone, and bovine serum albumin) is also used as blocking agent. Detergents, for example dodecyl sodium sulfate (better known as SDS), reduce surface tension and improve mixing, while helping to lower background.

Temperature is an important factor that can be manipulated during hybridization and post-hybridization washes of microarrays, and here again much can be learned from what has already been established for Northern or Southern blots. It is valuable to have an understanding of melting temperature of a duplex, or T_m, which is the temperature at which half of the molecules are double-stranded while the other half are "melted" or are in separate, single strands. It is useful even to know the ramifications of the equation defining the T_m of a duplex (longer than 50 bp) as:

$T_m = 81.5° + 16.6 \log M + 0.41$ (mole fraction G + C) $- 500/L - 0.62$ (% formamide)

showing the dependence of T_m on the monovalent cation concentration (M, mol L^{-1}), the base composition (expressed as the mole fraction of G + C), the length of the duplex (L), and the percentage of formamide present [46]. One begins to see why we hybridize at higher salt concentrations (to promote hybrid formation), and why we use formamide (to enable hybridization at lower, gentler temperatures).

Knowing that mispairing of bases tends to reduce T_m by approximately one degree for every percentage mismatch explains why increasing the temperature will cause nonspecific hybrids to fall apart first, and why washing arrays in low salt will favor the retention of true hybrids.

It should be noted, however, that these relationships were derived for systems where the probe was a single molecule of known

length and base composition, and the salt and formamide were in the soup where the probe would seek its immobilized target. In microarrays, there is not just one probe but thousands, and it is the probe that is stationary, and the target free. Thankfully, the probes, by definition, are of known length and composition, and therefore individual T_m values can be calculated for each probe.

Perhaps the greatest advantage offered by arrays of long oligonucleotides as opposed to PCR-amplified inserts of cDNA clones lies in being able to design long oligonucleotides such that they are of equal length and similar base composition, so that they have very similar T_m values. Working with what is basically a "collective T_m" maximizes the benefits of manipulating temperature, salt concentrations, and formamide. When an array is made up of probes with a wide range of T_m values, a certain low salt concentration might indeed cause mismatched hybrids to fall apart, but it may also melt true hybrids that happen to be shorter or have a higher A + T content, and so have lower T_m values.

An aspect of microarray hybridization worth clarifying is the question of which product is in excess, the labeled target or the immobilized probe, to qualify it as following pseudo-first-order kinetics. For microarrays to be useful as a means of quantifying expression, the target has to be in limiting amounts, and the probe must be in sufficient excess as to remain virtually unchanged even after hybridization. In a two-color experiment there is the added complication of having two differently labeled targets vying for the same probe. Here, it is important to ensure competitive hybridization so that the concentration ratio of hybrids accurately reflect the initial concentration ratio of the targets, that is, the initial ratio of the mRNA transcripts in the two samples. Wang and coworkers [47] show that if the hybridization kinetics of the two labeled targets are different the observed concentration ratio of the hybrids becomes a function of the amount of probe on the array, which would lead to spurious expression ratios when the amount of probe is limiting. If, however, the amount of probe is in excess, the effect of unequal rate constants is minimized.

10.5.5 Data Acquisition and Quantification

When the washes are complete and the slides have been spun dry, the "wet work" of a microarray experiment is done, and all the data that experiment is capable of giving are set. The results, as it were, are ripe for the picking. For microarrays these are 16-bit TIFF (tagged image file format) images, worth tens of millions of bytes each, acquired using array scanners typically equipped with lasers to excite the fluorophores at specific wavelengths and photomultiplier tubes (PMT) to detect the emitted light. The choice of array scanners has widened substantially since 1999, when a survey by the Association of Biomolecular Resource Facilities (http://www.abrf.org) found only two major players in the scanner market, accounting for nearly two-thirds of all scanners used by microarray laboratories [48]. A technology review from Bowtell's group [49] lists nine manufacturers and over a dozen scanner models with different resolution and sensitivity, and different laser and excitation ranges. The most basic models offer excitation and detection of the two most commonly used fluorophores (Cy3 and Cy5) whereas higher-end models might enable excitation at several wavelengths, dynamic focus, linear dynamic range over several orders of magnitude, and options for high-throughput scanning.

The objective of the scanning procedure

is to obtain the "best" image, where the best is not necessarily the brightest but is the most faithful representation of the data on the slide. The conditions within a user's control to achieve this goal are the intensity of the light used to excite the fluorophores (laser power) and the sensitivity of detection (as set by PMT gain). The scans should pick up the bright spots and those that are not so bright, and be able to tell the difference. Theoretically, a 16-bit image can have intensities ranging from 0 to 65,535 ($2^{16}-1$) [50], and one will often see such a range of raw values; what is crucial, however, is the linearity of that range of values when related to actual amounts of fluorescent label on a spot, and ultimately to gene expression.

Spot saturation, where intensity values have basically hit the ceiling, is a problem that is dealt with by reducing laser power, which in turn might lead to loss of data at the low end. Low intensity spots can be excited to emit a measurable signal by manipulating laser power and PMT. Increasing sensitivity at the level of detection, however, also tends to increase background, which can flatten signal-to-noise ratios. Thus, setting scanning conditions becomes an exercise in optimization, and the wider the linear range of the image acquisition system, the greater the latitude available for adjustment.

Scanners invariably come with software for image analysis, although stand-alone image analysis packages are available. Data extraction from the image involves several steps: (1) gridding or locating the spots on the array; (2) segmentation or assignment of pixels either to foreground (true signal) or background; and (3) intensity extraction to obtain raw values for foreground and background associated with each spot (reviewed in Ref. [51]). Subtracting the background intensity from the foreground yields the spot intensity that can be used to calculate intensity ratios that are the first approximation of relative gene expression.

When the dynamic range of the samples themselves exceeds that of the instrument, the dynamic range of the data can be expanded by using a merging algorithm to stitch together two scans taken at different sensitivities to obtain an extended, coherent set [52]. The calculations deal with the clipping of data at saturation and the setting of certain values below a threshold to zero, a process called "quantization" by Garcia de la Nava and coworkers [52], but referred to as "flooring" by others (such as in Ref. [50]).

The notion of boundaries within which the data necessarily sit by virtue of their being based on 16-bit information is developed in a rather neat fashion by the Quackenbush group in their paper exploring how the fact that intensity values must be from 2^0 to 2^{16} actually constricts the possible range of log ratios that can be obtained from an experiment [50]. An interesting upshot of their analysis was identification of an optimum range of net intensities (between 2^{10} and 2^{12}, or 1024 and 4096, respectively), where the dynamic range for fold change is highest, and where useful data capable of validation will probably be found.

10.6 Data Analysis

Because Chapt. 17 in this volume deals specifically with data analysis issues in microarrays, the discussion here will focus on essentials and rationale. Not counting what could be a long period of sample collection, as with the laser-capture microdissection of single cells from tissue sections that could take weeks or months, the labeling, hybridization, washing, and scanning of microarrays are processes that have short turnaround times. When the quantified data from the images are obtained, typically in

the form of Tab-delimited text files, the fun (or frustration?) of data analysis begins.

Even at image quantification, most software (for example QuantArray) will enable the flagging of poor spots that need never enter the analysis. At that time, dust artifacts, comet tails, and other spot anomalies can be identified.

There are many filtering strategies for pre-processing the quantified data before formal analysis [53–55], including the flagging of ambiguous spots with intensities lower than a threshold defined by the mean intensity plus two standard deviations of supposedly negative spots (no DNA, buffer and/or nonhomologous DNA controls). Analyzing these low-intensity spots, which tend to be highly variable, can lead to spuriously high expression ratios that cannot be validated by other measures of gene expression, for example by real-time PCR. Filtering the data increases the reliability of the intensity ratio, which by definition must suffer from being calculated away from any notion of the magnitude of the component intensities

Data normalization addresses systematic errors that can skew the search for biological effects [56–58]. One of the most common sources of systematic error is the dye bias introduced by the use of different fluorophores to label the target. Print-tip differences can lead to sub-grid biases within the same array, while scanner anomalies can cause one side of an array seem to be brighter than the other. A normalization method can be evaluated by looking at its effect on the graph of log2 ratio versus intensity, familiarly referred to as the M–A plot. The mean log2 ratios of well-normalized data will have a distribution that centers at zero, the log ratios themselves will be independent of intensity, and the fitted line will be parallel to the intensity axis [51]. LOWESS (LOcally WEighted Scatterplot Smoothing), originally developed by Cleveland [59] and realized for microarray data processing by Speed and Smyth [58] is a widely used normalization method that generally fulfills the criteria above. Several variants on the method have appeared in the literature [60], including a composite method developed by the Speed group [57] that utilizes a microarray sample pool (MSP). Normalization across multiple slides can be accomplished by scaling the within-slide normalized data. In practice, examining the box plots of the normalized data of individual arrays for consistency of width can usually indicate whether one needs to normalize across arrays.

The value of data visualization at different stages of the analyses cannot be over-emphasized. Spatial plots can locate background problems and extreme values. The shape and spread of scatter plots and the height and width of box plots give an overall view of data quality that can give clues about the effects of filtering and different normalization strategies.

Clustering algorithms [61] are means of organizing microarray data according to similarities in expression patterns. Hughes and Shoemaker [62] make the case for "guilt by association" which posits that co-expressed genes must also be co-regulated, and a logical follow-up to this analysis is the search for common upstream or downstream factors that may tie these co-expressed genes together. An interesting upshot of this is that previously uncharacterized genes might be assigned a putative function on the basis of their grouping [63]. In terms of being able to look at clinical data, clustering can be a valuable tool for identifying profiles characteristic of specific pathologies. Figure 10.4 shows hierarchical clustering of unpublished preliminary data on Wilms tumor, from microarray experiments performed in this author's laborato-

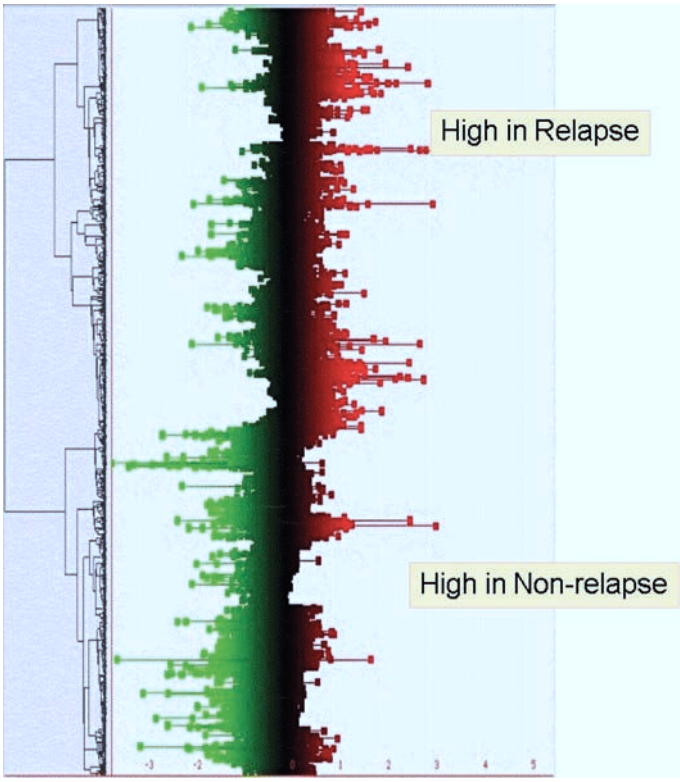


Fig. 10.4 Hierarchical clustering of microarray data showing two main clusters of gene expression profiles that can be associated with pathology.

ry, that show the clustering of genes into two main groups that might be indicative of the two disease states (relapse versus nonrelapse). Quackenbush [64] revisits the "guilt-by-association" approach in his commentary of a paper by Stuart and coworkers [65] studying co-expressed genes in an evolutionary context.

Initially, identifying differentially expressed genes simply meant finding genes with a fold change above a certain threshold, however that criterion was quickly shown to be a poor choice. As mentioned before, ratios give no information about absolute intensities; even more worrisome, however, is that when calculating mean ratios across slides, high variability within expression values for specific genes can lead to high mean ratios even if these genes are not really differentially expressed [16]. A better strategy computes *t*-statistics and corrects for multiple testing using adjusted *P*-values [66]. The *B*-statistic, derived using an empirical Bayes approach, has been shown in simulations to be far superior to either mean log ratios or the *t*-statistic for ranking differentially regulated genes [67]. The moderated *t*-statistic proposed by Smyth [68] develops the *B*-statistic of Lonnstedt and Speed [67] for application to wider experimental scenarios, remaining robust even with small numbers of arrays and allowing the presence of missing data because of filtering procedures.

Interestingly, the twofold change continues to be a benchmark for people perusing lists of microarray data who are already looking ahead to validating the data by PCR,

which is a method based on the exponential doubling of product. However, fold change has become more of a secondary criterion to select candidates for follow-up from a list of genes ranked according to more reliable measures of differential expression.

For details of the statistics of microarray data analysis, the reader is strongly advised to refer to the original literature cited above, and when strange formulae prove daunting, to consult persons with statistical expertise.

An encouraging development in this arena is the availability of open source software, three examples of which are reviewed by Dudoit and coauthors [69] – the statistical analysis tools written in R through the Bioconductor project (http://www.bioconductor.org), the Java-based TM4 software system from The Institute for Genomic Research (http://www.tigr.org/software), and BASE, the Web-based system developed at Lund University (http://base.thep.lu.se).

Referring to microarray hybridization experiments as "interrogations" sounds rather high-handed, conjuring images of arrays being cross-examined under a glaring lamp in an otherwise darkened room. And yet, as scientific jargon, it is amazingly intuitive. Microarray experiments do ask questions, with unsurpassed multiplicity, and it is not just a matter of being able to ask many questions all at once, but that the data itself can be rigorously interrogated.

After finding groups of genes with similar expression, by use of clustering and other computational techniques [70], and after differentially expressed genes are identified, further study of the data is needed if one is to make the journey back to the larger question that prompted the experiments in the first place. Thus, at this stage of the analysis, the objective is to uncover pathways, functions, or processes that might help make sense of the expression patterns and gene groupings observed. The Gene Ontology Consortium [71], which has sought to develop a common vocabulary applicable across eukaryotes to describe biological processes, molecular functions, and cellular components, has greatly facilitated the annotation of genes and therefore their assignment to specific ontologies. Tools such as Onto-Express [72] can be used to retrieve functional profiles and estimate the statistical significance of selecting certain genes from a given array as being differentially expressed.

10.7 Documentation of Microarrays

The adage that good science is documented science has never been truer or more necessary than when pertaining to microarrays. Printing arrays depend on careful and accurate documentation of the flow of materials from freezer or cupboard through the many steps of the process. Any changes to the array design, the materials, and the process, must be recorded. Fortunately, there are software packages that track DNA content through various stages of the printing process (e.g. from www.biodiscovery.com/).

Efforts to standardize the documentation of microarray hybridization experiments have been spearheaded by the Microarray Gene Expression Data Society (MGED; http://www.mged.org) with the establishment of MIAME, which describes the Minimum information about a microarray experiment that will enable unambiguous interpretation of the data, and the experiment to be repeated in other hands. The MIAME checklist at http://www.mged.org/Workgroups/MIAME/miame_checklist.html specifies the various bits of information required. In the original proposal published in 2001 [73], the minimum information was classified into six sections:

1. experimental design: the set of hybridization experiments as a whole
2. array design: each array used and each element (spot, feature) on the array
3. samples: samples used, extract preparation and labeling
4. hybridizations: procedures and conditions
5. measurements: images, quantification and specifications
6. normalization controls: types, values, and specifications

The following year, MGED wrote an open letter to the scientific journals urging the adoption of MIAME standards to microarray papers for publication [74]. Thus, microarray papers published today will invariably have links to databases that contain the MIAME information and all the raw data.

10.8 Applications of Microarrays in Cancer Research

The enormous potential of applying microarray technology to the study of human biology was first glimpsed with the publication of the paper by Schena and coworkers in 1996 [75] examining the effects of heat shock on Jurkat cells using a human cDNA array of just over a thousand genes. Later that same year the first microarray paper analyzing gene expression in human cancer appeared [42]; since then over 2500 articles on cancer and microarrays have been published. The sustained interest in using microarrays to study cancer attests to the suitability of microarrays as a tool to answer the most pressing challenges of cancer.

Although diagnosis and prognosis assessments of cancer typically rely heavily on histopathological data, no matter how fine the data they are inadequate to explain, for instance, why two children with stage I Wilms tumors that have very similar histological features and are virtually indistinguishable under the microscope will experience opposite clinical outcomes after treatment. The working hypothesis would be that these tumors must not have been as similar as previously thought. And because they are so outwardly similar, the differences between them must be subtle, though evidently far-reaching.

In the above scenario, several issues that challenge cancer management can be identified: correctly classifying or sub-classifying tumors, understanding tumorigenesis and tumor progression, predicting response to treatment, and predicting clinical outcome.

Acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) are leukemia subtypes that are managed using different treatments, and so it is important they be correctly diagnosed. In their landmark study Golub and colleagues [76] used a supervised learning approach to identify a subset of genes expressed differentially between AML and ALL that could act as a class predictor for correct classification of unknown leukemia samples with an accuracy greater than 85 %. Yeoh and coworkers [77] examined pediatric-ALL and identified distinct expression profiles characteristic of the known subtypes while gaining insight into the biology of each. More impressively, gene expression profiles of 7 and 20 genes were able to discriminate between relapse and remission in two pediatric-ALL subtypes, earmarking those patients at higher risk of relapse for more intense therapy. A study of expression profiles of diffuse large B-cell lymphoma (DLBCL) revealed two molecularly distinct types, one with a pattern reminiscent of germinal B-cells and the other of activated B-cells [21], and what is interesting is that patients with tumors

expressing genes with the germinal B-cell pattern had significantly better rates of survival over those with the activated B-cell pattern. More examples of the application of microarrays to tumor classification can be found in the primary literature and in a recent review [78].

Microarrays have also been used to identify tumor markers that can be diagnostic or prognostic, or that can act as drug targets. Gene expression profiles of bladder cancer cell lines [79] reveal that expression of cell adhesion molecules E-cadherin, zyxin, and moesin are associated with tumor stage and grade ($P < 0.05$). Interestingly, moesin expression is associated with survival with even higher significance ($P = 0.01$).

10.9 Conclusion

To say that microarrays have revolutionized science is not mere hyperbole. It is obvious from the wealth of literature that has accumulated in the short decade since their first appearance that microarrays have changed the way science is done. Never before had it been possible to even think of doing some of the experiments that have now almost become routine. Not that they were never thought of. The steep publication curve would suggest the ideas have been there all the time. All that was needed was a means to realize them. Microarrays were the door.

So what is on the other side? The publication flood has abated somewhat, for several reasons. It is not because less is being done. Rather, more is being required, which is as it should be for a field that is maturing into one that will actually make strides toward fulfilling its promise.

There are those who will say the bloom has gone from the rose, that microarrays are not all they have been cracked up to be. It is true that reality has set in for the microarray field. Far from losing heart, however, the research community has risen to the challenge. When else has science been accomplished with the input of such a diversity of expertise as to unite basic scientists, doctors, engineers, statisticians, and computer scientists? When else has it been possible (or even required) that methods, materials, and raw data be fully disclosed? Such free flow of information can only be good for the science, and the microarray community should be lauded for taking the initiative.

With the completion of the Human Genome Project (HGP), attention has rightly shifted to building on the blueprint and eventually fulfilling the project's promise of being beneficial to humankind. This lofty goal provided impetus to the Human Genome Project and motivates the natural progression beyond the sequencing of the human genome towards identifying gene functions and interactions in various contexts of time and space, in health or disease, in order to understand the workings of the human organism.

References

1 Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: lessons from large-scale biology. Science, 300, 286–290.

2 Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, 467–470.

3 Phimister B (1999) Going global. Nat Genet, 21, 1.

4 Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev, 16, 235–244.

5 Draghici S, Khatri P, Shah A, Tainsky MA (2003) Assessing the functional bias of commercial microarrays using the onto-compare database. Biotechniques, Suppl, 55–61.

6 Takemasa I, Higuchi H, Yamamoto H, Sekimoto M, Tomita N, Nakamori S, Matoba R, Monden M, Matsubara K (2001) Construction of preferential cDNA microarray specialized for human colorectal carcinoma: molecular sketch of colorectal cancer. Biochem Biophys Res Commun, 285, 1244–1249.

7 Hardiman G (2004) Microarray platforms – comparisons and contrasts. Pharmacogenomics, 5, 487–502.

8 Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D (1991) Light-directed, spatially addressable parallel chemical synthesis. Science, 251, 767–773.

9 Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol, 19, 342–347.

10 Ochs MF, Godwin AK (2003) Microarrays in cancer: research and applications. Biotechniques, Suppl, 4–15.

11 Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res, 28, 4552–4557.

12 Tiquia SM, Wu L, Chong SC, Passovets S, Xu D, Xu Y, Zhou J (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. Biotechniques, 36, 664–670, 672, 674–675.

13 Wang HY, Malek RL, Kwitek AE, Greene AS, Luu TV, Behbahani B, Frank B, Quackenbush J, Lee NH (2003) Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. Genome Biol, 4, R5.

14 Chen H, Sharp BM (2002) Oliz, a suite of Perl scripts that assist in the design of microarrays using 50mer oligonucleotides from the 3′ untranslated region. BMC Bioinformatics, 3, 27.

15 Tilstone C (2003) DNA microarrays: vital statistics. Nature, 424, 610–612.

16 Smyth G, Yang Y, Speed T (2003) Statistical issues in cDNA microarray data analysis. In: Functional Genomics: Methods and Protocols. Vol. 224 (eds. M Brownstein, A Khodursky), pp 111–136. Humana Press, Totowa, NJ.

17 Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet, 32 Suppl, 490–495.

18 Rosenzweig BA, Pine PS, Domon OE, Morris SM, Chen JJ, Sistare FD (2004) Dye bias correction in dual-labeled cDNA microarray gene expression measurements. Environ Health Perspect, 112, 480–487.

19 Dobbin K, Shih JH, Simon R (2003) Statistical design of reverse dye microarrays. Bioinformatics, 19, 803–810.

20 Yang YH, Speed T (2002) Design issues for cDNA microarray experiments. Nat Rev Genet, 3, 579–588.

21 Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403, 503–511.

22 Glonek GF, Solomon PJ (2004) Factorial and time course designs for cDNA microarray experiments. Biostatistics, 5, 89–111.

23 Townsend JP (2003) Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays. BMC Genomics, 4, 41.

24 Kerr MK (2003) Design considerations for efficient and effective microarray studies. Biometrics, 59, 822–828.

25 Simon R, Radmacher MD, Dobbin K (2002) Design of studies using DNA microarrays. Genet Epidemiol, 23, 21–36.

26 Dobbin K, Simon R (2002) Comparison of microarray designs for class comparison and class discovery. Bioinformatics, 18, 1438–1445.

27 Burgess JK, Hazelton RH (2000) New developments in the analysis of gene expression. Redox Rep, 5, 63–73.

28 Simone NL, Bonner RF, Gillespie JW, Emmert-Buck MR, Liotta LA (1998) Laser-capture microdissection: opening the microscopic frontier to molecular analysis. Trends Genet, 14, 272–276.

29 Arcellana-Panlilio MY, Schultz GA (1993) Analysis of messenger RNA. Methods Enzymol, 225, 303–328.

30 Ausubel FM (1995) Short protocols in molecular biology : a compendium of methods from Current protocols in molecular biology. Wiley, New York.

31 Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. Anal Biochem, 162, 156–159.

32 Naderi A, Ahmed AA, Barbosa-Morais NL, Aparicio S, Brenton JD, Caldas C (2004) Expression microarray reproducibility is improved by optimising purification steps in RNA amplification and labelling. BMC Genomics, 5, 9.

33 Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci USA, 87, 1663–1667.

34 Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P (1992) Analysis of gene expression in single live neurons. Proc Natl Acad Sci USA, 89, 3010–3014.

35 Kacharmina JE, Crino PB, Eberwine J (1999) Preparation of cDNA from single cells and subcellular regions. Methods Enzymol, 303, 3–18.

36 Bowtell D, Sambrook J (2003) DNA microarrays : a molecular cloning manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

37 Puskas LG, Zvara A, Hackler L, Jr., Van Hummelen P (2002) RNA amplification results in reproducible microarray data with slight ratio bias. Biotechniques, 32, 1330–1334, 1336, 1338, 1340.

38 Zhao H, Hastie T, Whitfield ML, Borresen-Dale AL, Jeffrey SS (2002) Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis. BMC Genomics, 3, 31.

39 Attia MA, Welsh JP, Laing K, Butcher PD, Gibson FM, Rutherford TR (2003) Fidelity and reproducibility of antisense RNA amplification for the study of gene expression in human CD34+ haemopoietic stem and progenitor cells. Br J Haematol, 122, 498–505.

40 Li Y, Li T, Liu S, Qiu M, Han Z, Jiang Z, Li R, Ying K, Xie Y, Mao Y (2004) Systematic comparison of the fidelity of aRNA, mRNA and T-RNA on gene expression profiling using cDNA microarray. J Biotechnol, 107, 19–28.

41 van Gijlswijk RP, Talman EG, Janssen PJ, Snoeijers SS, Killian J, Tanke HJ, Heetebrij RJ (2001) Universal Linkage System: versatile nucleic acid labeling technique. Expert Rev Mol Diagn, 1, 81–91.

42 DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet, 14, 457–460.

43 Zhu Z, Waggoner AS (1997) Molecular mechanism controlling the incorporation of fluorescent nucleotides into DNA by PCR. Cytometry, 28, 206–211.

44 Randolph JB, Waggoner AS (1997) Stability, specificity and fluorescence brightness of multiply-labeled fluorescent DNA probes. Nucleic Acids Res, 25, 2923–2929.

45 Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA, 96, 2907–2912.

46 Wahl GM, Berger SL, Kimmel AR (1987) Molecular hybridization of immobilized nucleic acids: theoretical concepts and practical considerations. Methods Enzymol, 152, 399–407.

47 Wang Y, Wang X, Guo SW, Ghosh S (2002) Conditions to ensure competitive hybridization in two-color microarray: a theoretical and experimental analysis. Biotechniques, 32, 1342–1346.

48 Grills G, Griffin C, Lilley K, Massimi A, Bao Y, VanEe J (2000) 1999/2000 ABRF Microarray Research Group Study. The State of the Art of Microarray Analysis: A Profile of Microarray Laboratories. The Association of Biomolecular Resource Facilities.

49 Holloway AJ, van Laar RK, Tothill RW, Bowtell DD (2002) Options available – from start to finish – for obtaining data from DNA microarrays II. Nat Genet, 32 Suppl, 481–489.

50 Sharov V, Kwong KY, Frank B, Chen E, Hasseman J, Gaspard R, Yu Y, Yang I, Quackenbush J (2004) The limits of log-ratios. BMC Biotechnol, 4, 3.

51 Leung YF, Cavalieri D (2003) Fundamentals of cDNA microarray data analysis. Trends Genet, 19, 649–659.

52 Garcia de la Nava J, Van Hijum S, Trelles O (2004) Saturation and Quantization Reduction in Microarray Experiments using Two Scans at Dierent Sensitivities. Statistical Applications in Genetics and Molecular Biology, 3, Article 11.

53 Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res, 29, 2549–2557.

54 Yang MC, Ruan QG, Yang JJ, Eckenrode S, Wu S, McIndoe RA, She JX (2001) A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. Physiol Genomics, 7, 45–53.

55 Wang X, Hessner MJ, Wu Y, Pati N, Ghosh S (2003) Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. Bioinformatics, 19, 1341–1347.

56 Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet, 32 Suppl, 496–501.

57 Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res, 30, e15.

58 Smyth GK, Speed T (2003) Normalization of cDNA microarray data. Methods, 31, 265–273.

59 Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc, 74, 829–836.

60 Cui X, Kerr MK, Churchill GA (2003) Transformations of cDNA microarray data. Statistical Applications in Genetics and Molecular Biology, 2, Article 4.

61 Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 95, 14863–14868.

62 Hughes TR, Shoemaker DD (2001) DNA microarrays for expression profiling. Curr Opin Chem Biol, 5, 21–25.

63 Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH (2000) Functional discovery via a compendium of expression profiles. Cell, 102, 109–126.

64 Quackenbush J (2003) Genomics. Microarrays – guilt by association. Science, 302, 240–241.

65 Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science, 302, 249–255.

66 Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12, 111–139.

67 Lonnstedt I, Speed TP (2002) Replicated microarray data. Statistica Sinica, 12, 31–46.

68 Smyth GK (2003) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, 3, Article 3.

69 Dudoit S, Gentleman RC, Quackenbush J (2003) Open source software for the analysis of microarray data. Biotechniques, Suppl, 45–51.

70 Quackenbush J (2001) Computational analysis of microarray data. Nat Rev Genet, 2, 418–427.

71 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25, 25–29.

72 Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA (2003) Global functional profiling of gene expression. Genomics, 81, 98–104.

73 Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet, 29, 365–371.

74 Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, Causton HC, Cavalieri D, Gaasterland T, Hingamp P, Holstege F, Ringwald M, Spellman P, Stoeckert CJ, Jr., Stewart JE, Taylor R, Brazma A, Quackenbush J (2002) The underlying principles of scientific publication. Bioinformatics, 18, 1409.

75 Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA, 93, 10614–10619.

76 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286, 531–537.

77 Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell, 1, 133–143.

78 Hampton GM, Frierson HF (2003) Classifying human cancer by analysis of gene expression. Trends Mol Med, 9, 5–10.

79 Sanchez-Carbayo M, Socci ND, Charytonowicz E, Lu M, Prystowsky M, Childs G, Cordon-Cardo C (2002) Molecular profiling of bladder cancer using cDNA microarrays: defining histogenesis and biological phenotypes. Cancer Res, 62, 6973–6980.

11
Yeast Two-hybrid Technologies

Gregor Jansen, David Y. Thomas, and Stephanie Pollock

11.1
Introduction

More than 100 genomic sequences have now been completed and multitudes of genome sequencing projects are in progress. Major developments in DNA sequencing technology have enabled the rapid acquisition of high-quality genome sequence data. The assignment of function has lagged far behind sequencing and methods have chiefly relied on bioinformatics. Comparison of total genomes and the inference of gene function by identification of orthologs is a well established procedure, but even in the most completely annotated genomes (for example that of *Saccharomyces cerevisiae*) there are approximately 25 % of ORFs that do not yet have a possible function assigned. The magnitude of the task of obtaining a complete understanding of the function of a protein can be appreciated by the enormous number of publications on single proteins, for example 31,793 publications on p53 and 88,496 that mention Ras. Thus, beyond a first assignment of enzymatic or cellular function there is a vast amount to be learnt before a knowledge of the function of a protein is complete. Even initial assignment of the proteins is, currently, a major challenge, however. Bioinformatics approaches are powerful, but are limited by the experimental data available, and high-throughput methods such as genome-wide gene deletions or siRNA gene knock-downs, still need extensive further experimental analysis. Another genome-wide analytical approach uses DNA microarrays. The global analysis of gene expression profiles after a variety of experimental changes has provided many clues about the function of proteins and their associated pathways.

A different approach to establishing protein function is to find their interaction partners. Protein–protein interactions are found for most proteins, and partners that interact are expected to participate in the same cellular process, thus providing clues to the function of unknown interaction partners. Many techniques have been developed to study protein–protein interactions. These include traditional biochemical methods such as co-purification, and coimmunoprecipitation to identify the members of protein complexes. This approach usually requires complex and specific optimization of the experimental setup, however. Recently, new proteomics-based strategies have been

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

used to determine the composition of complexes and to establish interaction networks. For example, a recent large-scale approach used affinity tagging and mass spectroscopy to study protein–protein interactions in yeast [1, 2]. So far this method has been limited, because it requires the target organism to be amenable to genetic manipulation.

The yeast two-hybrid system is an excellent means of high throughput detection of protein–protein interactions for any organism. Yeast two-hybrid systems detect not only members of known complexes, but also weak or transient interactions. They have been shown to be robust and adaptable to automation. In addition, protein expression in yeast can provide the important protein modifications necessary for stability and proper folding, but not available in bacteria or *in vitro*. Interaction screening can be performed with random cDNA libraries from any organism or involve robotics with grids of predefined clones. Over the last decade a variety of two-hybrid technologies have been developed, enabling study of a broad range of proteins, including membrane proteins. To date, the yeast two-hybrid system has been widely used for determination of protein interaction networks within different organisms such as *Helicobacter pylori* (bacterium) [3], *Saccharomyces cerevisiae* (yeast) [4, 5], *Caenorhabditis elegans* (worm) [6], and *Drosophila melanogaster* (fly) [7].

11.2 The Classical Yeast Two-hybrid System

The yeast two-hybrid system was developed by Fields and Song [8] and variations of this are the most widely used two-hybrid system. It makes use of the modular organization found in many transcription factors that have a DNA-binding (DB) domain and activation domains (AD) that can function independently, but controlled transcriptional activity that can be reconstituted when the independent domains are fused to two proteins that interact. The original system was based on the reconstitution of the DB and AD domains of the yeast transcription factor Gal4p. Fusions of these domains, the DB usually termed the bait and AD the prey, are made to proteins of interest and the interaction of bait and prey via the fused proteins brings the separated domains into proximity. This reconstitutes a complex with the ability to bind to specific upstream activating sequences (UAS) in the promoter region of target genes and to specifically activate their transcription (Fig. 11.1A). In a similar system, the *E. coli* DNA-binding protein LexAp and the independent B42p strong activation domain have also been used [9]. The power of the system is that the UAS that is recognized by the DB can be placed to control the transcription of selectable or detectable reporter genes. A commonly used reporter gene contains the specific UAS for either Gal4p or LexAp in the promoter region of *LacZ* and interactions are monitored by well characterized and sensitive colorimetric assays. Other reporter genes such as *HIS3*, *ADE2* and *URA3* are used and report interactions by conferring growth under selective conditions. The yeast two-hybrid system has been used to generate large-scale interaction maps, but also to define the domains and individual amino acid residues involved the interaction. Occasionally results obtained from the yeast two-hybrid system have been corroborated by structural studies [10].

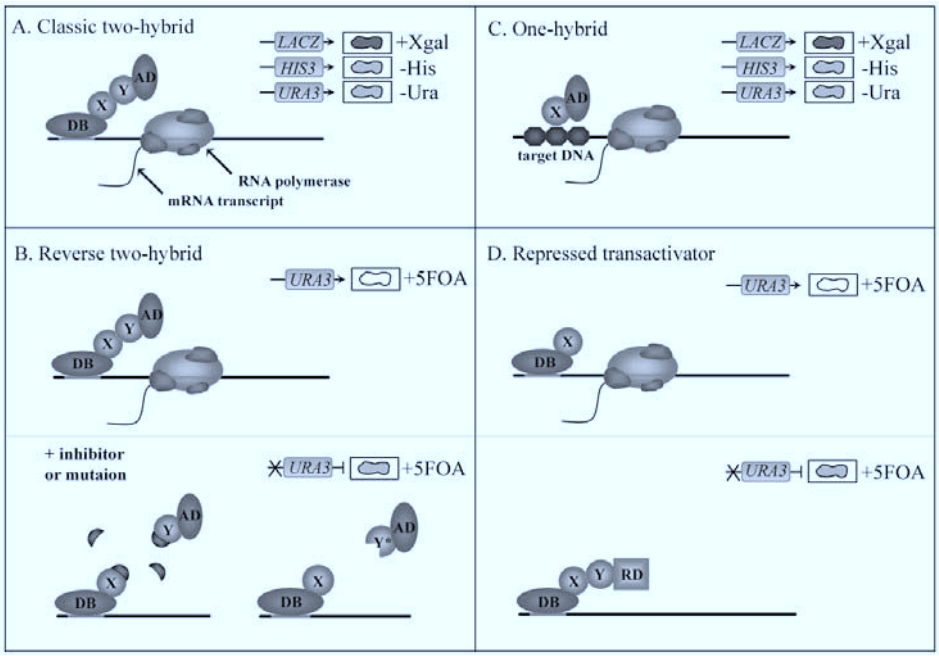


Fig. 11.1 (A) Classic two-hybrid system. The interaction of bait (X) and prey (Y) fusion proteins brings the DNA-binding domain (DB) and activation domain (AD) into close proximity. This enables binding to specific upstream activating sequences in the promoter region of target genes and activates their transcription. Interactions can be measured by growth (*HIS3* and *URA3*) or by colorimetric assays (*LACZ*). Dark gray patches represent growth and change of color, light gray patches represent growth, and white patches represent no growth. (B) Reverse two-hybrid system. On interaction of bait and prey fusion proteins, induced *URA3* expression leads to 5-FOA being converted into the toxic substance 5-fluorouracil by Ura3p, leading to growth inhibition. Mutated or fragmented genes are created, then subjected to analysis, and only loss-of-interaction mutants are able to grow in the presence of 5-FOA. (C) One-hybrid system. In this system the bait is a target DNA fragment fused to a reporter gene. Preys that are able to bind to the DNA-fragment-reporter fusion will lead to activation of the reporter genes (*LACZ*, *HIS3*, and *URA3*). (D) Repressed transactivator system. In this system the interaction of bait-DB fusion proteins and the prey–repressor domain (RD) fusion proteins can be detected by repression of the reporter *URA3*. The interaction of bait and prey enables cells to grow in the presence of 5-FOA, whereas noninteractors are sensitive to 5-FOA, because of Ura3p production.

11.3 Variations of the Two-hybrid System

11.3.1 The Reverse Two-hybrid System

The reverse two-hybrid system was developed to select against protein–protein interactions. It is useful for screening for inhibitors of protein interaction and also for the rapid mapping of residues and domains involved in interactions. It uses the counter selection provided by the UAS-*URA3* growth reporter in the presence of the compound 5-fluoroorotic acid (5-FOA), where Ura3p converts 5-FOA into 5-fluorouracil, which is toxic to cells [11]. In an example of

this technique the genes for proteins in a known interaction are mutated and the cells selected on 5-FOA, where only loss-of-interaction mutants are able to grow (Fig. 11.1B). This approach, together with *in-vivo* recombination techniques simplifies the mapping of domains and interaction surfaces. Other counter-selectable reporters such as cycloheximide resistance have been incorporated into reverse two-hybrid systems [12].

11.3.2 The One-hybrid System

One-hybrid systems are used for identification of DNA-binding proteins that can bind to a DNA fragment of interest, for example a UAS in a promoter region. The "bait" in this system is a DNA sequence placed upstream of a reporter gene. Fusion libraries, in which the preys with potential DNA binding activity are fused to a known activation domain, are used. Preys that are able to bind to the DNA-fragment-reporter fusion lead to activation of the reporter (Fig. 11.1C). The yeast one-hybrid system has been used to identify a yeast origin of replication binding protein [13]. The system has been further refined to identify proteins that interfere with known DNA–protein interactions by introduction of the counter-selectable reporter *URA3* (reverse one-hybrid) [11].

11.3.3 The Repressed Transactivator System

Transcriptional activators are not usually used for two-hybrid studies, because of their ability to auto-activate the reporter genes without having to interact with a bait fusion protein. To overcome this limitation, the repressed transactivator system (RTA) uses the general transcriptional repressor Tup1p. Tup1p participates with Mig1p and other cofactors in the presence of glucose to repress the transcription of genes containing upstream repressing sequences in their promoters. Like transcriptional activators, the transcriptional repressor domain (RD) alone has been shown to be sufficient to confer repression on a $URS_{glucose}$-*lexA* fusion reporter. In the RTA system the interaction of the bait Gal4p-DBfusion proteins and the prey Tup1p-RD fusion proteins can be detected by the repression of the Gal4p-dependent reporter *URA3*. The interaction of bait and prey allows cells to grow in the presence of 5-FOA, whereas noninteractors are sensitive to 5-FOA, because of the increased Ura3p expression caused by the autoactivating bait-transcription-factor fusion protein (Fig. 11.1D) [14].

11.3.4 Three-hybrid Systems

In the three-hybrid system the interaction of bait and prey protein requires the presence of a third interacting molecule to form a complex. The third interacting molecule can be a protein used with a nuclear localization and it acts as a bridge between bait and prey to cause transcriptional activation. The bridging effect might be provided by interacting with both bait and prey at the same time or by forming an intact interaction surface in combination with one partner that enables the second partner to bind (Fig. 11.2). This system has been used to show the ternary complex formation of EGF, GRB2, and SOS [15]. Reporter gene activation only occurs in the presence of all three proteins, but combination of any two proteins is inactive. The expression of a third protein can also be used in a reverse two-hybrid approach by preventing the interaction between bait and prey by competition.

In the RNA three-hybrid system the third interacting component expressed is a bifunctional RNA (Fig. 11.2). This system has

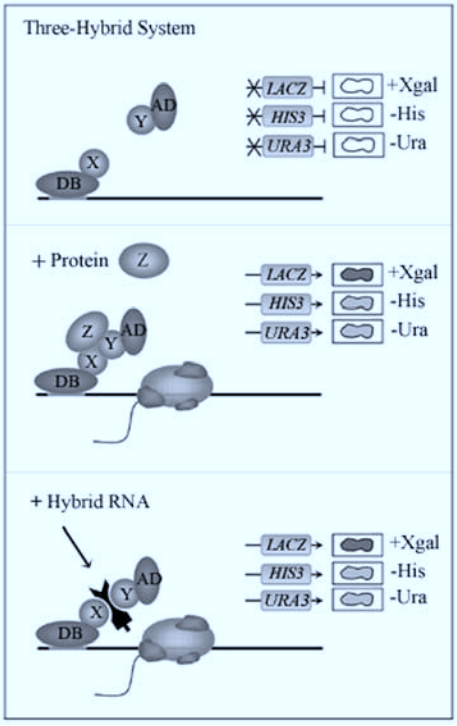


Fig. 11.2 Three-hybrid system. In the three-hybrid system bait and prey fusion proteins require the presence of a third protein or additional RNA molecule to form a complex. The third protein (Z) is expressed with a nuclear localization signal in the presence of bait and prey fusions, and will act as a bridge between them to reconstitute transcriptional activation of reporter genes *LACZ*, *HIS3*, and *URA3*. In the RNA three-hybrid system the third interacting component expressed is a bifunctional RNA. Dark gray patches represent growth and change of color, light gray patches represent growth, and white patches represent no growth.

been used to analyze RNA–protein interactions in which the bifunctional RNA serves as an adaptor between bait and prey, thereby leading to the activation of a reporter gene [16]. For example, the *psi* region of the RNA genome retrovirus has been characterized with this system [17].

The third interacting component can also consist of a small chemical compound as shown for the compound FK506 [18]. This system especially shows promise for identification of the targets of small molecules. One strategy has been to chemically fuse a small molecule with an unknown cellular target to a chemical such as FK506 with a well characterized cellular target. The bifunctional chemical is then introduced into cells where the DB is fused to the known cellular target (FK506 binding protein). An interaction mediated by the small molecule is then determined as for the classical two-hybrid system. This approach has not been as widely used, perhaps because of the difficulties of obtaining hybrid small molecules that retain chemical activity, but a recent publication has shown the power of this approach with protein kinases [19].

Although not strictly a three-hybrid system there are several systems in which introduction of a third gene has been used to modulate the interaction between bait and prey. For example yeast does not have protein tyrosine kinases and the interaction of proteins in the insulin signaling pathway via SH2 domains and specific phosphotyrosine residues has been shown by the expression of a tyrosine kinase [20].

11.4 Membrane Yeast Two-hybrid Systems

At least 30 % of the ORFs in genomes have been estimated to code for membrane proteins and a limitation of classical yeast two-hybrid systems is that it does not detect their interaction. To overcome this problem several membrane based two-hybrid systems have been developed.

These membrane two-hybrid systems do not depend on a direct transcriptional readout for detection of protein–protein interaction. The signaling events that report interaction of bait and prey occur in the cyto-

plasm whereas the interaction itself can occur in a membrane environment or within a compartment such as the ER.

11.4.1 SOS Recruitment System

The SOS recruitment system (SRS) is based on the observation that the human GDP–GTP exchange factor (GEF) hSOS can complement a corresponding yeast temperature-sensitive mutant (*cdc25ts*) when targeted to the plasma membrane. There hSOS stimulates the GDP to GTP exchange on Ras, which in turn leads to restored growth of the *cdc25ts* strain at the restrictive temperature. In the SRS the bait is fused to hSOS and is localized on the plasma membrane by interaction with the prey protein [21]. Because of the lack of specific reporter genes for the Ras pathway induction, an interaction can only be detected by monitoring growth, or lack of it. A variation of this system has been termed the Ras recruitment system (RRS); in this hSOS has been replaced by a human mRas deleted for its membrane attachment modification (CAAX box). The bait is fused to mRas, localized to the plasma membrane by interaction with a specific prey and thereby restores growth to the *cdc25ts* mutant strain (Fig. 11.3A) [22]. The prey protein can either be a plasma membrane protein or it can be artificially localized on the plasma membrane by fusion to the Src myristoylation signal that provides membrane attachment.

11.4.2 Split-ubiquitin System

A variant system for study of membrane protein interactions is the split-ubiquitin system. Ubiquitin acts as a tag for protein degradation, and fusion proteins with ubiquitin are rapidly cleaved by ubiquitin-specific proteases. The split-ubiquitin system was originally developed by Johnsson and Varshavsky using bait and prey separately fused to the N- and C-terminal domain of ubiquitin [23]. Interaction of bait and prey reconstitute an active ubiquitin leading to the specific proteolytic cleavage of a reporter protein from one of the ubiquitin fusions. The proteolytically released reporter in the original system was detected by immunological techniques. A recently developed variation of the system, however, uses the artificial transcription factor lexAp-VP16p as a reporter [24]. The release of the transcription factor on interaction enables its translocation to the nucleus, where it activates classical two-hybrid LexAp coupled reporter genes such as *LacZ* and *HIS3* (Fig. 11.4A). In another version the reporter is a modified RUra3p, which is cleaved on reconstitution of ubiquitin by bait–prey interaction [25]. RUra3p is subsequently degraded, leading to phenotypically ura$^-$ cells that are able to grow in the presence of 5-FOA, reporting the interaction. To reduce the background present in both systems, because of low affinity of the split ubiquitin domains for each other, without the need for interacting fusion proteins, a mutant Nubp (NubGp) has been developed that has lower affinity for Cub [23]. This system has been more widely used [26].

11.4.3 G-Protein Fusion System

This variation of a membrane two-hybrid system uses the heterotrimeric G-protein of the mating signaling pathway in *S. cerevisiae*. The heterotrimeric G-protein consists of the subunits Gα (Gpa1p), Gβ (Ste4p), and Gγ (Ste18p) and is coupled to a pheromone-specific receptor (GPCR). On binding of the pheromone the GPCR triggers the dissociation of the Gα from the Gβγ sub-

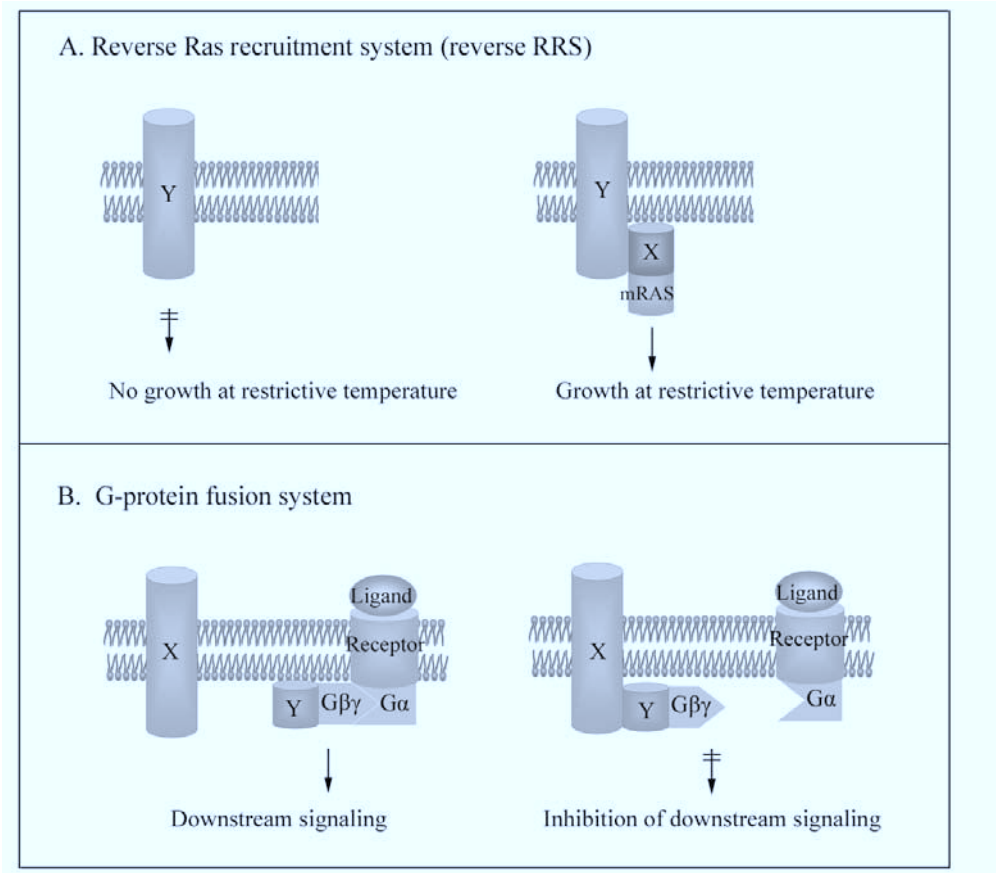


Fig. 11.3 (A) Reverse Ras recruitment system. In this system the bait is fused to mRAS and localized on the plasma membrane by interaction with a specific prey protein, thereby restoring growth in the cdc25ts mutant strain at the restrictive temperature. The prey protein can either be a plasma membrane protein or artificially localized on the plasma membrane. (B) G-protein fusion system. In this system the bait consists of a membrane protein whereas the prey is a cytosolic protein fused to Gγ. Upon interaction with the prey the Gγ fusion protein is dissociated from the Gγ/receptor/ligand complex and is unable to participate in downstream signaling.

units. In wild type yeast, free Gβγ results in the activation of a kinase cascade and the expression of mating-specific genes, which ultimately leads to growth arrest, cell fusion, and mating. In the G-protein fusion system the bait consists of a membrane protein whereas the prey is a cytosolic protein fused to Gγ. On high-affinity interaction with the prey the Gγ fusion protein is sequestered from the effector and is thus unable to participate in pheromone signaling. The yeast tester strain is thereby rendered mating-incompetent, and this can be measured by pheromone sensitivity assays and induction of mating-specific reporters such as *FUS1-lacZ* (Fig. 11.3B). Interactions such as syntaxin2a with neuronal sec1 and fibroblast-derived growth factor receptor 3 with SNT-1 have been demonstrated in this system [27].

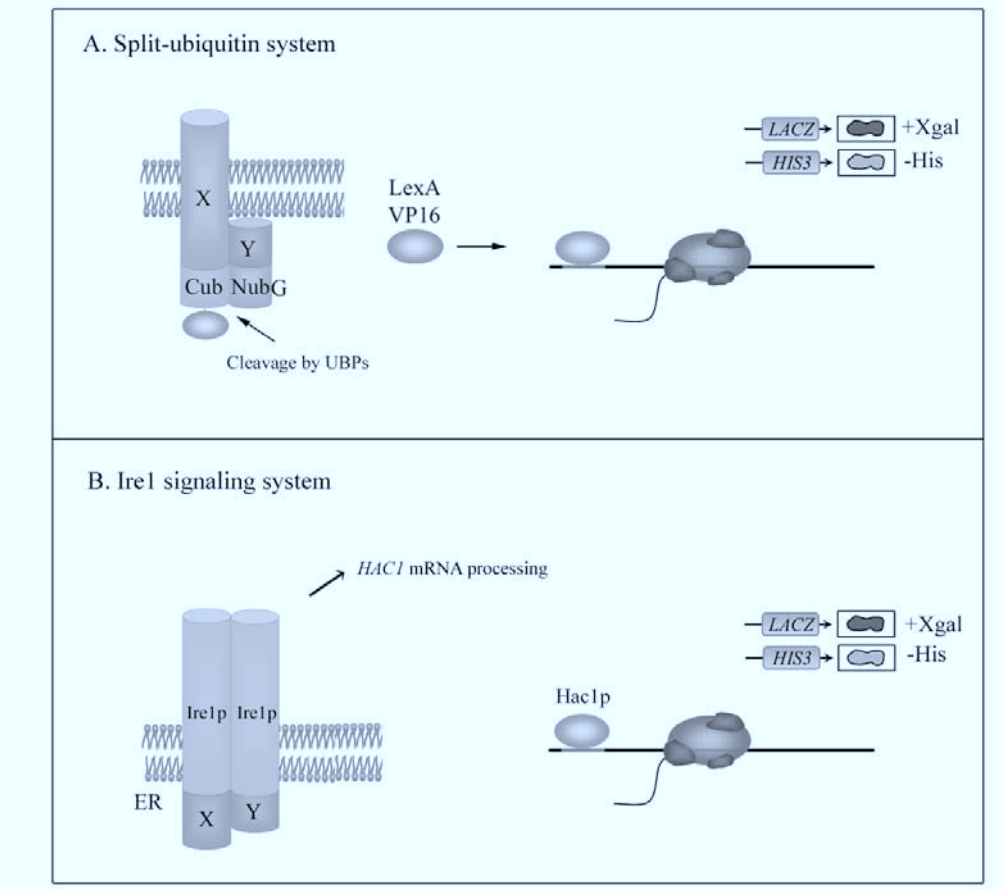


Fig. 11.4 (A) Split-ubiquitin system. In this system interaction of bait and prey reconstitute an active ubiquitin, represented by Cub (C-ubiquitin) and NubG (N-ubiquitin), leading to the specific proteolytic cleavage of the artificial transcription factor lexAp-VP16p by ubiquitin-specific proteases (UBP). Release of the transcription factor enables its translocation to the nucleus, where it activates classical two-hybrid lexAp coupled reporter genes such as *lacZ* and *HIS3*. Dark gray patches represent yeast growth and change of color, and light gray patches represent growth.
(B) Ire1p signaling system. In this system the Ire1p N-terminus is replaced with bait and prey proteins. Upon interaction of N-terminal generated Ire1p-fusions, the unfolded protein response is activated resulting in increased expression of active Hac1p. The interaction-induced expression of Hac1p is monitored by UPRE-coupled *HIS3* and *lacZ* reporters.

11.4.4 The Ire1 Signaling System

Another membrane protein two-hybrid system that can detect interactions between membrane proteins is based on the unfolded protein response system (UPR) in yeast. In wild-type cells this response to unfolded proteins in the endoplasmic reticulum is initiated by the type I membrane protein kinase Ire1p. The N-terminus of Ire1p is located in the lumen of the ER and oligomerizes in response to misfolded proteins within the ER. The Ire1p cytosolic kinase

domains transphosphorylate on dimerization and then activate the endoribonuclease domain present at the C-termini of the Ire1p. This nuclease can then specifically cleave the mRNA of the UPR transcriptional activator Hac1p and this nonspliceosomal mechanism is completed by tRNA ligase. The intron within *HAC1* mRNA has been shown to attenuate the translation of unspliced *HAC1* mRNA, and its removal leads to upregulation of the expression of genes involved in the UPR by binding of Hac1p to unfolded protein response elements (UPRE) in the promoter regions of target genes. On interaction of N-terminal generated Ire1p-fusions (replacing the Ire1p N-terminus with bait and prey) the UPR is activated, resulting in the increased expression of active Hac1p. The interaction-induced expression of Hac1p is monitored by UPRE-coupled growth and *LacZ* reporters (Fig. 11.4B). The system has been used to demonstrate epitope–antibody interactions [28] and the interaction between the ER chaperones Calnexin and ERp57, including the mapping of their specific interaction domains [29].

11.4.5 Non-yeast Hybrid Systems

Two-hybrid strategies have also been applied to organisms other than yeast. A number of mammalian two-hybrid systems have been created by using a protein fragment complementation assay (PCA). The PCA systems are based on protein–protein interactions that are coupled to the refolding of enzymes from fragments, and the reconstitution of enzyme activity is used to detect the interaction. These include the β-galactosidase complementation assay in which β-galactosidase activity is restored by interacting fragments in an otherwise β-galactosidase-defective cell line [30]. In another PCA-based two-hybrid method two separated fragments of the enzyme dihydrofolate reductase (DHFR) are used and interacting bait and prey fused to the fragments will restore enzyme activity [31]. The PCA strategy has also been used to split the enzyme β-lactamase and interaction of bait and prey fused to the fragments leads to restored enzyme activity which can be measured in a colorimetric assay [32]. In general, PCA systems enable quantification and real time monitoring of protein–protein interactions and with rapid advances in imaging technology have great promise for a wide variety of applications.

11.5 Interpretation of Two-hybrid Results

Yeast two-hybrid systems have seen wide application both in the study of specific interactions and in the construction of interaction networks (interactomes). Whereas the authenticity of specific interactions is often supported by other data obtained from genetic and biochemical experiments, evidence for interaction networks relies on indirect methods.

Typically specific interactions that have been found by use of yeast two-hybrid systems are confirmed by a variety of biochemical methods including coimmunoprecipitation. An advantage of the two-hybrid systems are that they enable the rapid delimitation of independently interacting domains and guide construction of peptides that can be used as competitors of the interaction of the proteins *in vitro* [29]. One particularly powerful genetic approach has been to use the two-hybrid system to select mutations in one partner that reduce interaction and obtain compensating mutations in the other partner to define interaction domains. In some notable cases structural

studies have shown that these genetic findings are reflected in interacting domains and residues [10, 29].

Two-hybrid systems can also create artifacts, termed "false positives" and "false negatives". The nature of false positives is likely to be different for some of the variants of the two-hybrid system, but can be basically defined as reporter gene activation without the presence of a specific interaction between bait and prey. For classical two-hybrid systems self-activating baits are frequently observed because of activation of the reporters above threshold level by bait-like transcription factors. In addition mutations in the host cell can lead to unspecific activation of the reporters. Another possibility is that bait and prey fusions might interfere with host metabolism, leading to changed viability under selective conditions. Another source of false positives is that two-hybrid libraries often contain proteins of function homologous to that of proteins used for the selection of interactions. Random cDNA libraries also contain small peptides with high content of charged amino acids that can act as artificial activation domains. Strategies to detect or exclude false positives involve the use of multiple reporters, the switch of bait and prey fusions, and two-hybrid independent methods such as coimmunoprecipitations for verification of observed protein–protein interactions. Nevertheless, it should be kept in mind that even after careful verification it is possible that the observed interaction is nonetheless not biologically relevant and the interaction has been artificially created by the two-hybrid localization of bait and prey.

False negatives are, by definition, interactions that cannot be detected using two-hybrid methods. These are thought to occur as a result of steric hindrance between fusion proteins that prevents the reconstitution of the activity needed to induce reporters. Misfolding, degradation of the fusion protein, absent post-translational modifications, and mislocalization of bait or prey are other possible reasons for the generation of false negatives. For example, in the classical two-hybrid system the interaction has to occur in the nucleus to be able to generate the transcriptional readout. Another source of false negatives is large-scale two-hybrid screens that generally use PCR amplified genes, introduced into bait and prey vectors by *in-vivo* homologous recombination. The resulting recombinants are usually not sequence-verified, because of to the large number of clones, but potentially contain mutations or frame shifts that prevent interactions and therefore appear as false negatives.

11.6 Conclusion

In the 15 years since its introduction the yeast two-hybrid system has seen remarkably widespread adoption for a variety of uses. It has been shown to be remarkably robust and has now been used for construction of interactomes of soluble proteins from a variety of organisms. The development of new two-hybrid systems that can detect interaction of membrane proteins that have previously proved recalcitrant to many established techniques is especially powerful and will enable the completion of genome interactomes.

References

1 Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K. et al. (2002) *Nature*, **415**, 180–183.

2 Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M. et al. (2002) *Nature*, **415**, 141–147.

3 Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V. et al. (2001) *Nature*, **409**, 211–215.

4 Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. et al. (2000) *Nature*, **403**, 623–627.

5 Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) *Proc Natl Acad Sci USA*, **98**, 4569–4574.

6 Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N. and Vidal, M. (2000) *Science*, **287**, 116–122.

7 Uetz, P. and Pankratz, M.J. (2004) *Nat Biotechnol*, **22**, 43–44.

8 Fields, S. and Song, O. (1989) *Nature*, **340**, 245–246.

9 Gyuris, J., Golemis, E., Chertkov, H. and Brent, R. (1993) *Cell*, **75**, 791–803.

10 Clark, K.L., Dignard, D., Thomas, D.Y. and Whiteway, M. (1993) *Mol Cell Biol*, **13**, 1–8.

11 Vidal, M., Brachmann, R.K., Fattaey, A., Harlow, E. and Boeke, J.D. (1996) *Proc Natl Acad Sci USA*, **93**, 10315–10320.

12 Leanna, C.A. and Hannink, M. (1996) *Nucleic Acids Res*, **24**, 3341–3347.

13 Li, J.J. and Herskowitz, I. (1993) *Science*, **262**, 1870–1874.

14 Hirst, M., Ho, C., Sabourin, L., Rudnicki, M., Penn, L. and Sadowski, I. (2001) *Proc Natl Acad Sci USA*, **98**, 8726–8731.

15 Zhang, J. and Lautar, S. (1996) *Anal Biochem*, **242**, 68–72.

16 SenGupta, D.J., Zhang, B., Kraemer, B., Pochart, P., Fields, S. and Wickens, M. (1996) *Proc Natl Acad Sci USA*, **93**, 8496–8501.

17 Evans, M.J., Bacharach, E. and Goff, S.P. (2004) *J Virol*, **78**, 7677–7684.

18 Licitra, E.J. and Liu, J.O. (1996) *Proc Natl Acad Sci USA*, **93**, 12817–12821.

19 Becker, F., Murthi, K., Smith, C., Come, J., Costa-Roldan, N., Kaufmann, C., Hanke, U., Degenhart, C., Baumann, S., Wallner, W. et al. (2004) *Chem Biol*, **11**, 211–223.

20 O'Neill, T.J., Craparo, A. and Gustafson, T.A. (1994) *Mol Cell Biol*, **14**, 6433–6442.

21 Aronheim, A., Zandi, E., Hennemann, H., Elledge, S.J. and Karin, M. (1997) *Mol Cell Biol*, **17**, 3094–3102.

22 Broder, Y.C., Katz, S. and Aronheim, A. (1998) *Curr Biol*, **8**, 1121–1124.

23 Johnsson, N. and Varshavsky, A. (1994) *Proc Natl Acad Sci USA*, **91**, 10340–10344.

24 Stagljar, I., Korostensky, C., Johnsson, N. and te Heesen, S. (1998) *Proc Natl Acad Sci USA*, **95**, 5187–5192.

25 Wittke, S., Lewke, N., Muller, S. and Johnsson, N. (1999) *Mol Biol Cell*, **10**, 2519–2530.

26 Wang, B., Pelletier, J., Massaad, M.J., Herscovics, A. and Shore, G.C. (2004) *Mol Cell Biol*, **24**, 2767–2778.

27 Ehrhard, K.N., Jacoby, J.J., Fu, X.Y., Jahn, R. and Dohlman, H.G. (2000) *Nat Biotechnol*, **18**, 1075–1079.

28 Urech, D.M., Lichtlen, P. and Barberis, A. (2003) *Biochim Biophys Acta*, **1622**, 117–127.

29 Pollock, S., Kozlov, G., Pelletier, M.F., Trempe, J.F., Jansen, G., Sitnikov, D., Bergeron, J.J., Gehring, K., Ekiel, I. and Thomas, D.Y. (2004) *Embo J*, **23**, 1020–1029.
30 Rossi, F., Charlton, C.A. and Blau, H.M. (1997) *Proc Natl Acad Sci USA*, **94**, 8405–8410.
31 Remy, I. and Michnick, S.W. (1999) *Proc Natl Acad Sci USA*, **96**, 5394–5399.
32 Galarneau, A., Primeau, M., Trudeau, L.E. and Michnick, S.W. (2002) *Nat Biotechnol*, **20**, 619–622.

12
Structural Genomics

Aalim M. Weljie, Hans J. Vogel,
and Ernst M. Bergmann

12.1
Introduction

The completion of various genome-sequencing efforts has made it apparent that genomics can only offer a glimpse at the blueprint of life. In order to fully understand the physiology of any biological system, from the simple bacteria with relatively small genomes, to the complex plants and humans with very large genomes, we need to know what specific proteins are expressed in what cells, at what times and how these interact with each other. This has created a need for the burgeoning field of proteomics. In parallel a need arose for better understanding of proteins at a molecular level; the original objective of structural genomics, also known as structural proteomics, was to solve as many protein structures as possible in an attempt to identify the complete protein "fold-space". By having examples of all protein folds in the protein database it would then be possible to derive homology models for any protein. Broad-based structural genomics initiatives have been established to achieve this ambitious goal, all relying on synchrotron protein crystallography and high-resolution NMR (nuclear magnetic resonance) spectroscopy to generate the 3D structures. In this chapter we will discuss the technical advances that were necessary to enable protein crystallography and solution NMR to move toward becoming high-throughput techniques capable of dealing with the expected onslaught of requests for detailed structural information. Throughout this chapter we will also review the current status of the structural genomics field. It should become apparent to the reader that the field is moving away from simply filling in the fold-space, and that attention has been shifting more towards "functional genomics", i.e. defining the role of unknown gene products by identifying the ligands and other partners they interact with. Clearly these developments also have a large impact on rational drug design using proteins as potential targets.

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

12.2 Protein Crystallography and Structural Genomics

12.2.1 High-throughput Protein Crystallography

Protein crystallography remains the predominant means of providing detailed structural information about biological macromolecules [1–3]. As such it plays an important role in all structural and functional genomics initiatives that endeavor to solve the structures of proteins at the atomic level. Most of the large – national and international – structural genomics projects combine protein crystallography projects with complimentary techniques, including high resolution NMR spectroscopy, cryo electron microscopy, and small-angle X-ray scattering, for elucidation of protein structures [4].

Several large-scale structural genomics initiatives aimed at elucidating a complete set of three-dimensional structures for the gene products of entire organisms have evolved as a consequence of successful genome sequencing efforts [5, 6]. The feasibility of these projects did, however, also depend on the timely availability of newly developed methods in protein crystallography. The two most important advances were the development of cryogenic data collection and multiple wavelengths anomalous dispersion (MAD) phasing [7–10]. The first method was developed by Hope and others in the late eighties to prevent radiation damage to protein crystals by high intensity X-rays [7, 8, 11]. It involves capturing and mounting a protein crystal in a drop of liquid inside a slightly larger fiber loop. Cryogenic data collection has become the standard method used by crystallographers. It has greatly facilitated progress in the field and enabled the utilization of powerful new third-generation synchrotron X-ray sources [12–14]. It was also a prerequisite for the widespread use of MAD phasing, which requires very large data sets and long data-collection times [9]. MAD phasing was developed by Hendrickson and others in the nineteen eighties. Since then it has become the preferred and most widely used method for phasing and solving novel protein crystal structures. MAD phasing relies on the wavelength-dependent differences in intensities introduced by a few anomalously scattering atoms [10]. The most commonly used element for MAD phasing is Se, which is introduced in the form of seleno-methionine in the protein via specific expression systems [15]. It has long been accepted that this is a nonperturbing substitution in proteins.

Other technological developments also contributed to a revolution in the field of protein crystallography. Amongst these were new computational methods and new X-ray detectors [9, 12, 16]. Developments outside the field itself led to phenomenal advances in molecular genetics that made many more proteins available through cloning and expression – powerful and affordable computers and third-generation synchrotron X-ray sources with dedicated experimental stations for protein crystallography. As a consequence of all these recent developments protein crystallography has reached a point were determination of the crystal structure of a protein can be accomplished in days or even hours once a good quality crystal has been obtained. This progress in the field was a prerequisite to support the concept of large-scale structural genomics efforts. Since the inception of the worldwide structural genomics initiatives a few years ago these are now, in turn, contributing to developments in protein crystallography methods [17].

Substantial structural genomics efforts are in progress in North America, Europe, and Japan [18–20]. Most notable in North

America are the nine NIH/NIGMS-funded pilot projects and some smaller Canadian initiatives. A major Welcome Trust funded project to elucidate human protein structures is based both in the UK and Canada. Major European projects include the Proteinstrukturfabrik in Berlin and SPINE (centered in Oxford). A large-scale Japanese project is headquartered in Yokohama at the RIKEN Institute. All of these centers have ambitious goals to solve thousands of protein structures within the next few years; to be able to fulfill these goals it will be necessary for almost all centers to increase their current output and efficiency. As of the time of writing this chapter, early 2004, structural genomics efforts worldwide have contributed approximately 1000 or so structures to the database of known protein structures. Obviously the impact of these efforts is just now starting to become significant. Many of the major national and international structural genomics efforts have initially focused on developing methods and procedures for high-throughput structure determination and initially applied these methods to pilot projects.

Commercial structural genomics efforts are designed with the goal of discovering and developing novel drugs [2, 20]. Structural and functional genomics initially help to identify and characterize suitable drug targets. Beyond doubt, the outcome of the structural genomics efforts will help identify thousands of novel drug targets. The methods and technologies developed for structural genomics can then be used to improve lead compounds for drugs. The latter process requires the determination of dozens or hundreds of crystal structures of a drug target, usually a protein, complexed with modified versions of the lead compound. This process requires close cooperation and communication between pharmacologists, medicinal chemists, and structural biologists. The process benefits greatly from the high-throughput methods and processes developed by the structural genomics initiatives [21].

The process of determining the structure of a target protein by X-ray crystallography requires several steps [1, 17]. The gene must be cloned into a suitable vector and expressed. The protein must be available in sufficient quantity and in soluble form from the expression system. Subsequently it has to be purified. The purified protein must then be crystallized and the crystals must be of sufficient size and quality to enable recording of an X-ray diffraction pattern. Diffraction data must be collected at a synchrotron X-ray source. The structure must finally be solved, refined, and analyzed. Some of these steps are more difficult and time-consuming than others and the process can fail for individual proteins at any step along the way. The first results from the worldwide structural genomics projects show that structures of only five to twenty percent of the gene products that have been targeted can be completed in an initial round. Attrition occurs at all the individual steps of the process. Figure 12.1 shows the success statistics from several of the NIH/NIGMS-funded structural genomics efforts as of Spring 2004. The success rate has increased since the onset of these initiatives [19].

The most difficult steps during determination of a protein crystal structure are currently the expression and purification of the protein and the preparation of suitable, well-diffracting crystals [18]. It has become rare that solution of the structure fails after X-ray diffraction data have been collected from a crystal. The phasing, solution, and refinement of a crystal structure used to be a major research effort but has become a routine procedure once a set of diffraction data from a protein crystal containing anomalously scattering atoms has been collected.

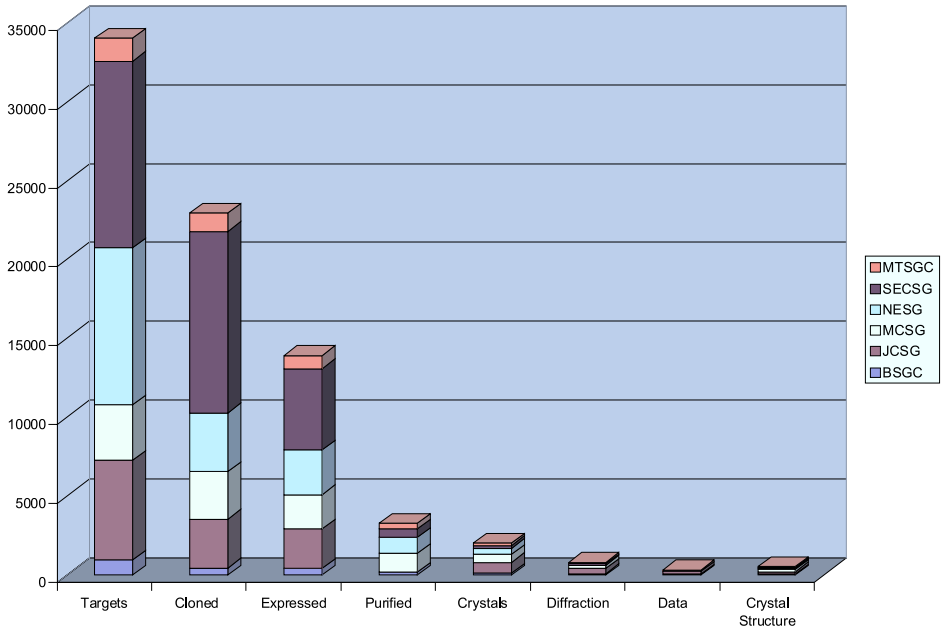


Fig. 12.1 A snapshot of the progress of six of the NIGMS-funded US structural genomics initiatives as of spring 2004. The six consortia (M. Tuberculosis Structural Genomics Consortium, Southeast Collaboratory for Structural Genomics, Northeast Structural Genomics Consortium, Midwest Center for Structural Genomics, Joint Center for Structural Genomics and Berkeley Structural Genomics Center) have cloned 22955 of 34041 targets and completed 521 crystal structures. Some data for purified protein and diffraction data are missing. Data are from the web sites of the consortia (http:www.nigms.nih.gov/psi/centers.html).

This development has shifted the major bottleneck in protein crystal structure determination to the crystallization of the protein; this, in turn, depends on the availability of highly purified, soluble protein. Results from structural genomics efforts have demonstrated generally how improving methods in one of the steps outlined above shifts the rate-limiting step of the process and creates new bottlenecks. At present the expression, purification, and crystallization of the protein has become the most difficult, demanding, and unpredictable part of the overall process.

12.2.2 Protein Production

Protein production for protein crystallography projects within structural genomics efforts consists of the cloning of the gene, expression of the gene in a suitable host system and the purification of the expressed protein. In the context of other ongoing structural biology research, proteins are still sometimes purified from the original source material (e.g. blood, organs, plants, etc.). This approach, however, is no longer commonly used – it is hard to scale-up to large numbers and amounts of proteins and is, therefore, not suitable within the context

of high-throughput projects. The cloning and expression of the gene in question have therefore become prerequisites for most structural biology research on specific gene products.

The cloning of the genes in the context of a large structural genomics effort employs as much as possible standard, modular, and flexible procedures and automation [22]. It is desirable to design a clone so that it can be utilized with several different expression systems (bacteria, insect cells, yeast, *in-vitro* systems) without the need for much additional work [22]. Furthermore, it is desirable to have the flexibility to express the clone with different affinity tags or as different constructs. Unfortunately it is not predictable *a priori* which construct will give the best chance of successful expression and purification for any given protein [23]. Amino- and carboxy-terminal histidine tags, glutathione *S*-transferase tags, maltose binding protein and thioredoxin fusion proteins have all been used successfully. Many structural genomics efforts therefore simultaneously produce clones and expression systems of several different constructs, to maximize the possibility of successful expression and purification. The most popular expression systems are still bacterial, most notably *Escherichia coli*. Insect cells are the preferred choice for proteins that cannot be expressed in bacteria, because of a requirement for post-translational modifications. *In-vitro* expression systems are not yet in widespread use but they are gaining in popularity [24]. One of their main advantages is the possibility of adding specific detergents, chaperones, metal ions, or other ligands, as required, to improve the folding and stability of the expressed protein. *In-vitro* expression systems also provide an opportunity to test and optimize expression conditions, initially on a small test scale, and they are relatively easy to automate.

Purification of the target proteins is best accomplished in a single step by use of affinity chromatography. More complex purification schemes are usually not applicable within the context of the high-throughput nature of structural genomics efforts and are almost impossible to automate. Because the choice of the affinity tag that will be successful and optimum for any given protein cannot be rationalized *a priori*, the cloning of the target gene in a flexible vector and the expression of several constructs with different affinity tags can be critical for success during this stage. Nevertheless, protein production is still a difficult step for many proteins and for some it can turn into the most difficult bottleneck. Many proteins cannot be over-expressed, or expressed at all, without at least optimization of the conditions within the context of the correct expression system. Others can be expressed but appear in an insoluble form or are misfolded. Proteins that are insoluble, often in the form of inclusion bodies, can sometimes be renatured. If renaturing of the insoluble protein is possible it might help with the purification, because the inclusion bodies are simple to separate from the remainder of the cell extract. Solubilization is, however, not possible with all insoluble proteins. Considerable research efforts have been directed toward finding methods to solubilize these proteins. Attempts have been made to change the solubility of a protein by random or systematic mutation or directed evolution [23]. Several of these methods are promising but none has been generally successful with every protein tested. These methods also require considerable additional effort. Purification by affinity chromatography often produces protein that is adequately pure for biophysical or structural studies and, if anything, requires only one more additional purification step, such as size-exclusion chromatography. The affinity

tag can be removed from the purified protein or not. Whether or not it is detrimental to crystallization again depends on the individual protein. Many structural genomics projects attempt protein crystallization initially with the affinity tag still attached; some use both forms of the purified protein for initial crystallization trials, with and without affinity tag. Removal of the affinity tag is usually accomplished via a specific proteolytic cleavage site that was included in the construct for the gene. TEV protease and factor X are often used for this purpose, because they have a relatively narrow specificity, hence reducing the risk of proteolytic cleavage of the desired protein itself.

The purity of the final product of the purification procedure must be analytically confirmed by chromatography, electrophoresis, mass spectroscopy, or a combination of these methods [22, 23]. Dynamic light scattering is also a method frequently used by crystallographers to determine the homogeneity of the protein in solution [25].

12.2.3 Protein Crystallization

Determination of the crystal structures of proteins by X-ray crystallography requires the preparation of high quality single crystals of sufficient size. The advent of new powerful X-ray sources in the form of third-generation synchrotrons has made it possible to obtain X-ray diffraction patterns that are sufficient to solve a structure from crystals that are less than 100 µm, or even less than 50 µm, in size. This greatly reduces the time and effort required to crystallize a protein and solve its crystal structure.

The crystallization of proteins remains, however, a trial-and-error process that is difficult to rationalize [18, 23, 26, 27]. Crystals are grown from small drops, usually a few microliters or less, of an aqueous solution that is buffered and of defined ionic strength and protein concentration. It also contains a precipitant such as a species of polyethylene glycol or a salt. Ammonium sulfate is still one of the most widely and successfully used precipitants in protein crystallography. The crystallization drops are equilibrated against a larger reservoir containing a higher concentration of the precipitant. As the crystallization drop equilibrates, eventually reaching the precipitant concentration of the reservoir, the protein concentration increases until the protein comes out of solution, under the right conditions in the form of crystals. Several different crystallization methods are in use, hanging drop vapor diffusion, sitting-drop vapor diffusion, micro batch under oil, and micro dialysis, among others [18, 25, 27]. For the purpose of structural genomics initiatives the most commonly used method is sitting-drop vapor diffusion [21, 27]. This is probably the method that is the most flexible and easiest to automate.

The initial crystallization conditions for a protein are found by screening a range of conditions, for example sample pH, ionic strength, protein and precipitant concentration, and often small concentrations of additives. Screening procedures currently in use employ between 48 and up to 1536 different conditions. Duplicate experiments are usually performed at several different controlled temperatures (e.g. 4 °C, 14 °C and 20 °C). Typically several hundred crystallization conditions are tested in initial screening for a new protein. The crystallization drops are then observed and monitored for several days or weeks to identify potential crystals. Some proteins yield diffraction quality single crystals even in the initial screening, others produce very small crystals or crystals of poor quality. In the latter circumstance the crystallization conditions must be improved by fine screening of pH, ionic strength, and precipitant concentra-

tion around the conditions of the initial screening. Additional techniques employed at this stage are seeding and additives. Optimization of the initial crystallization conditions can be a difficult and time-consuming process. These steps are difficult to automate and can become a bottleneck in structural genomics initiatives.

Most proteins that can be expressed in soluble form and purified to homogeneity can also be crystallized. Proteins that can be obtained in very pure and soluble form but fail to crystallize completely are rare. A more difficult problem can be to improve the crystallization conditions to the point were the crystals are large enough and of sufficient quality. If no crystals are obtained for a protein from initial screening of several hundred or even thousand conditions, it is probably advisable to analyze carefully the expression and purification of the protein. There might be minor impurities or contamination, the protein might aggregate and form different oligomers in solution, or the protein may be unstable over time. Different expression constructs of the same protein behave differently and should be tested independently.

Crystallization of proteins within structural genomics initiatives is, today, usually fully automated [17, 21, 25]. Liquid handling robots have been adapted to perform the setup of crystals screening and subsequent fine screening. This is not only faster and more cost-efficient but also more reliable than performing these experiments by hand. Liquid-handling robots are much better at producing hundreds of very small crystallization drops than any human. This is particularly true for experiments in the sub-microliter range that are becoming the standard for screening of crystallization conditions [28].

Structural genomics initiatives have contributed a great deal to the automation of protein crystallization. Almost all the steps in the process of producing protein crystals from cloning to crystallization have been automated and these robotic technologies are now widely available [21, 25]. This includes even the screening of the crystallization drops by automated microscopes. Software to automate analysis of the images from these microscope systems is being developed, but currently the reliability of the automatic analysis is such that some human observation is still required [29].

Another contribution to the development of protein crystallography by the world's structural genomics initiatives will be the data accumulating from the crystallization trials, especially the success of the screening methods. Screening for crystallization conditions is currently a trial-and-error process that cannot be rationalized *a priori*. The data currently being accumulated from the large number of crystallization trials of the worldwide structural genomics initiatives, both positive and negative, will be very valuable in providing a better rationale for this process. Initial data from some of the structural genomics initiatives already suggest that some of the conditions commonly used for crystallization screening procedures are redundant, and how others can be improved [30].

12.2.4 Data Collection

Data collection from protein crystals ideally requires access to a beamline in a third-generation synchrotron dedicated to macromolecular crystallography as an X-ray source [9, 12, 14]. These X-ray sources are brilliant and tunable, enable data collection in the shortest possible time, at the best possible resolution, and are capable of collecting several wavelengths for MAD phasing, if required. The development of these beamlines in the

last ten years has greatly helped to facilitate determination of the structure of proteins by crystallography. Many of the individual structures of biological macromolecules that have been published recently could not have been completed without access to modern beamlines.

Before data collection, crystals are harvested from the crystallization drop with a small fiber loop, ideally matched to the size of the crystal, and immediately cooled to near liquid nitrogen temperature by plunging the loop and crystal into liquid nitrogen or mounting it in a stream of cold nitrogen gas [7, 8, 11]. A stream of gaseous nitrogen only a few degrees warmer than liquid nitrogen then cools the crystal during the data collection. The protein crystal is surrounded in the loop by a drop of the mother liquor from which it was harvested. An additive that can act as an antifreeze needs to be present in the mother liquor (20–30 % glycerol or ethylene glycol are commonly used). As a consequence of the rapid cooling and the additive the crystal is maintained in a solid glass-like state. Formation of crystalline ice would result in the destruction of the protein crystal. The process of harvesting and mounting the crystals has not yet been successfully automated. Alternative mounting procedures have been developed but the loop mount is the most commonly used. Research efforts are directed toward procedures that would enable direct mounting of the crystal in the container or capillary where it was grown, because it would be an advantage not to have to manipulate the crystal. Older methods in which the crystal was mounted at room temperature inside a sealed glass capillary have become obsolete for anything but some very specialized experiments.

The loop-mounting of protein crystals enables long-term storage of these crystals under liquid nitrogen, In fact they are usually shipped at liquid nitrogen temperature to synchrotron facilities for data collection. At the synchrotron facilities the crystals are mounted in the X-ray beam by robotic sample-mounting systems [17]. The development of these robotic sample-mounting systems for data collection was one of the most tangible early technologies resulting from structural genomics initiatives. Their development was driven by the need for more efficient use of expensive and oversubscribed beam time at synchrotron facilities for structural biology. Combined with the requisite software development this has led to a situation where data collection from protein crystals is almost completely automated and can be controlled remotely. Two processes necessary for data collection have not been automated successfully at the time of writing. One is capture of the crystals in the mounting loop. This still has to be done by a researcher looking through a microscope. The other is the centering of the loop-mounted crystals in the X-ray beam. The latter step requires identification of the crystal mounted in the loop, which can be difficult. Imaging systems are capable of identifying and centering the loop but a reliable method for picking out the crystal itself has remained elusive. To circumvent this problem it is possible to screen crystals by simply centering the loop and tuning the size of the X-ray beam to the size of the loop. This is not ideal for collection of quality data for which, ideally, the beam is matched to the size of the crystal. Nevertheless, progress in the automation of data collection from protein crystals has been dramatic and led to much more efficient use of synchrotron beam time for protein crystallography [17].

Development of software and powerful but affordable computers has also contributed to the current situation in which it is possible to analyze and process the X-ray

diffraction data from protein crystals very quickly and with little or no user intervention. Very user-friendly and automated software can today provide completely processed X-ray diffraction data within minutes of the completion of the experiment [16, 20, 31, 32]. In favorable cases it is even possible to solve protein crystal structures while data collection is still in progress [32].

12.2.5 Structure Solution and Refinement

Most of the structural targets of structural genomics initiatives are unique structures which do not have any known closely related structural homologues [4, 18]. This necessitates solving the structure by experimental phasing methods. As already mentioned, nowadays by far the most common method of phasing novel protein structures is MAD phasing. MAD phasing requires the presence of heavier atoms with spectroscopic absorption edges in the hard X-ray region of the spectrum, and an X-ray source that can be tuned to the wavelength of these edges. It utilizes the wavelength-dependent differences in intensities introduced by a few anomalously scattering atoms (Se, Br, Zn, Fe, Ga, Cu, Hg, Ir, Au, Pb, Lu, Yb have been used, among others) to derive an initial set of phases for a crystal structure from the position of these atoms. Ideally, it requires collecting entire data sets at each of two or three wavelengths, at least one being at an X-ray absorption edge of the element used. Because data collection for a MAD experiment takes three times longer and decay of the X-ray diffraction, induced by radiation damage to the crystal, is very detrimental to the resulting phases and the experiment, cryogenic data collection is essential. The most commonly used element for MAD phasing is Se in the form of the nonperturbing seleno-methionine, which can be readily introduced into the protein via specific expression systems [15]. MAD phasing has considerable advantages over most other crystallographic phasing methods still in use for protein structures. Most importantly MAD phasing provides the most accurate, unbiased experimental phases from a single crystal. MAD phasing has also enabled further automation of the solution of protein crystal structures. If there are no major problems with the crystal or the X-ray diffraction data and the crystal diffracts to at least reasonable resolution, current software packages can provide nearly complete protein structures with little or no user intervention and starting from a good MAD data set. This will increasingly place the emphasis within a structural project on the preparation of good crystals, which, in turn, requires paying careful attention to expression and purification of the protein.

Almost all crystal structures still require some input from an experienced scientist at the stage of crystallographic refinement [1]. Details of the water structure, multiple or unusual conformations, or the presence of unusual ligands, that are present in almost all protein crystal structures, cannot be modeled accurately by automated procedures. These final details still require several iterations of analysis, if necessary manual fitting and refinement. Because the input to this procedure has, most often, become an almost complete good-quality model, the time required has shortened to weeks or even days from, usually, several months of work by an experienced crystallographer which has enabled the ambitious future goals of the structural genomics initiatives.

Despite all the progress some crystallographic projects will remain serious challenges. Membrane proteins remain a more difficult proposition because the proteins are generally difficult to express, purify, and

crystallize [33]. The procedures employed for membrane proteins are not completely novel and different; the proteins are just more difficult to handle. Expression, purification, and crystallization require the addition of detergents and additives, specialized expression systems, and pose the additional problem of screening for the right detergent conditions.

Other proteins that provide a more serious challenge are cellular proteins from higher organisms that cannot be expressed easily in bacterial systems, because of the requirement for chaperone-mediated folding or covalent modifications. *In-vitro* expression systems might help in some of these cases [24].

12.2.6
Analysis

As increasing numbers of protein crystal structures are being determined and are becoming available, the considerable data base being accumulated by the structural genomics initiatives contains not only structural information but also information about expression, purification, and crystallization of proteins. Detailed analysis of the structures and the data bases will not only enhance our understanding of macromolecular structure but also, eventually, provide a rational strategy for preparation of purified proteins and success in crystallization. In the development of a rational basis for protein purification and crystallization, to replace the screening of suitable conditions, the negative outcome of many experiments forms an equally important part of the data base. The data that are accumulated by the worldwide structural genomics efforts will dramatically improve our understanding of proteins just as the data base of atomic resolution structures enhances our understanding of macromolecular structure.

12.3
NMR and Structural Genomics

12.3.1
High-throughput Structure Determination by NMR

Structural genomics has had an obvious impact on the development of software and instrumentation for high-throughput NMR. Through a variety of structural genomics projects registered with http://targetdb.pbd.org, NMR has proven to be a robust tool for determination of high-resolution structures, with almost 20–25 % of all structures solved to date being determined by NMR spectroscopy. Typical NMR structures and crystal structures for the protein calmodulin are shown in Fig. 12.2. The two experimental techniques can often be complementary with NMR frequently having a better chance at analyzing more flexible protein systems and X-ray techniques being more directly applicable to more rigid proteins that form well-diffracting crystals. It should be noted that although solid-state NMR has much promise in this field, particularly with regard to membrane-associated proteins [34], the scope of this section will be limited to high-resolution solution NMR and structural genomics. Emphasis will also be given on providing the reader with a basic overview of this topic, as most aspects of the role of NMR in structural genomics have been well reviewed recently [35–39].

12.3.1.1
Target Selection

As with X-ray crystallography, target selection is the most important step for successful structure determination by NMR. The two most important factors which affect target selection for NMR spectroscopic analysis are the size of the protein and the ability to label the proteins with NMR-active stable

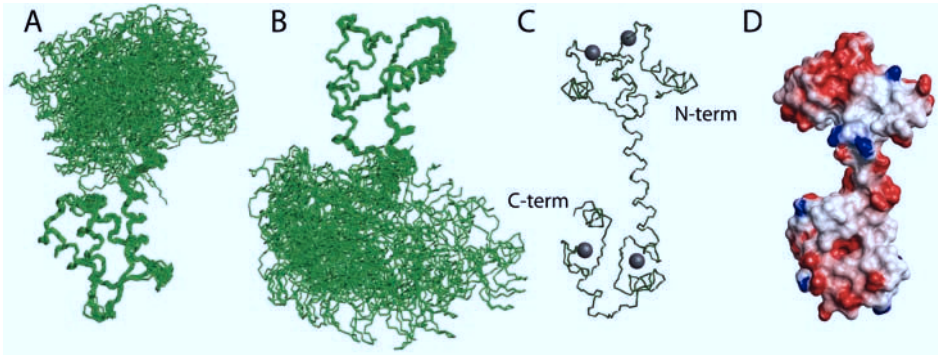


Fig. 12.2 Representative high resolution NMR (A, B) and X-ray crystallography (C) structures of the calcium-binding regulatory protein calmodulin. In (A) and (B) the protein is in its apo state whereas the crystal structure in (C) is for the calcium-bound state. NMR structures are usually presented as an ensemble of structures, whereas crystal structures are typically refined into one unique structure. The ensemble of NMR structures in (A) and (B) is necessary to account for the conformational heterogeneity of calmodulin in the linker region in solution. This flexibility occurs in both the calcium and apo forms, the latter of which does not enable its crystallization. Note that the two independently folded lobes of calmodulin in the NMR structures overlay very well, the flexible point is in the connecting linker between them, enabling the two structurally well-defined lobes to take up widely different positions relative to each other. Addition of calcium produces a more stable form which can be crystallized, and in (C) the crystal structure shows the included Ca^{2+} ions as dark spheres. Note that in NMR structures of calcium-saturated calmodulin the calcium ions would not be seen. In the crystal structure the linker appears as a continuous helical structure, which is stabilized by crystal packing forces and represents one extreme of the conformations that can occur in solution. (D) demonstrates the electrostatic charge surface of calcium-loaded calmodulin, with red areas holding negative charge and blue positive, and provides a partial electrostatic explanation of why the acidic calmodulin binds positively charged calcium ions.

isotopes, usually by expression of the gene for the protein in high quantities. Even though the size barrier has expanded rapidly beyond the 30–40 kDa range in recent years, the desired outcome of an NMR structure is greatly influenced by the size of the protein. Traditionally, NMR structures have been data-intensive, with many hours of data collection required in the best cases, and usually days or weeks for large proteins or difficult samples. Full structural analysis of NMR data can also be intricate and has, traditionally, been a labor-intensive endeavor. If the goal is to characterize the protein fold or fold family, then large proteins (>40 kDa) can be routinely analyzed. Determination of complete high-resolution structures is, however, hampered by the complexity of the spectra for larger proteins. Although technical and software innovations are working to remove these barriers as described below, NMR structures are still generally smaller than this soft limit and also smaller than those typically determined by protein crystallography. Proteins of a size suitable for NMR have been targeted specifically in a Canadian-based pilot study [37] in which 20 % of the 513 NMR-sized proteins targeted were found to be suitable for structural NMR analysis, suggesting a major role for the technique.

An often-stated general goal of structural genomics projects is to map "fold space" [40]. Most proteins are members of families

of sequence homologues from multiple organisms. Analysis of two organisms, *E. coli* and *Thermotoga maritima*, found that between the two organisms, 62 of 68 selected sequence homologues were readily amenable to structural analysis by either X-ray or NMR spectroscopy [41]. Another interesting finding was that the number of proteins amenable to analysis by both X-ray and NMR spectroscopy was surprisingly small [41], hence X-ray and NMR spectroscopy are suggested to be highly complementary techniques.

Another unique approach to NMR targeted sample selection advocated by Iain Campbell's group in Oxford to overcome size limitations is the separation of proteins into their constituent domains before analysis [42]. This modular approach greatly increases the number of potentially solvable NMR targets, while exploiting the advantages of decreasing undesired effects of interdomain flexibility which often hampers both X-ray and NMR determination of the structure of larger intact proteins. It also enables specific soluble parts of proteins to be solved, even if other domains do not lend themselves well for analysis, for example because of transmembrane regions.

Insofar as the labeling of NMR targets is concerned, the general methodology involves expressing the cloned protein gene in a system which can grow on metabolites containing ^{13}C and ^{15}N and/or ^{2}H depending on the experiments required. Often the initial stages involve labeling with ^{15}N only, because this is relatively inexpensive compared with ^{13}C and can provide the basis for ^{1}H,^{15}N correlations which determine if the protein is sufficiently well-behaved to merit further analysis. For complete structure determination it is necessary to fully label proteins >10 kDa with the ^{13}C isotope as well; the structures of smaller proteins can often be solved by using proton NMR spectroscopy alone. Proteins of molecular weight in excess of 25 kDa typically also require deuteration, which results in line narrowing in the spectra. In addition to total isotope labeling strategies it is possible to selectively label only residues of one type (e.g. His) in proteins, The extent of labeling can vary greatly (10–100 %), as can the selectivity of the amino acids labeled. Irrespective of the experiments performed, however, a fundamental requirement is high-level protein expression, and this subject has also been well reviewed recently in the context of structural genomics [43]. By far the most common approach is expression in *E. coli*, although other eukaryotic systems such as yeast are also routinely used. A novel whole-organism-labeling approach adapted for plants has recently been reported in which kilogram quantities of ^{15}N potato material were obtained and several ^{15}N-labelled proteins were purified and proved suitable for NMR analysis [44]. Another innovative protein synthesis strategy which holds a great deal of promise for the production of toxic, unstable, and/or low solubility proteins is the cell-free *in-vitro* expression method [24]. In this approach the isolated cellular components used for protein synthesis are combined to catalyze protein production in a test tube. Importantly, as already discussed, this approach also has great potential for X-ray crystallography, because selenomethionine labeled proteins can also be produced [45] for MAD diffraction.

12.3.1.2

High-throughput Data Acquisition

The core of NMR instrumentation lies in pulsing magnetically susceptible nuclei (e.g.. ^{1}H, ^{13}C, ^{15}N, ^{2}H) with radiofrequency (RF) pulse trains and then measuring the relaxation decay of the observed nucleus back to equilibrium. Innovation in NMR, driven in part by structural genomics, has occurred both in the hardware used for ac-

quiring NMR data and the software used to control the instrumentation.

An inherent issue in conventional NMR is low sensitivity because of the characteristics of the magnetic quantum transitions which underlie the technique. As a result, traditionally highly concentrated samples (>0.5–1 mm) and many milligrams of protein are required for successful data acquisition. The advent of cryogenically cooled probes, higher field strengths, and microvolume sample sizes are working together to reduce this obstacle. In a cryo- or cold-probe, key components such as the detection coil and pre-amplifier used in the transmission and generation of RF pulses are enveloped in a stream of recirculating helium gas cooled to ~25–30 °K. This enables approximately 2 to 4-fold gains in signal-to-noise ratio, depending on sample properties such as solvent and ionic strength. Increasing theoretical sensitivity fourfold enables use of collection times a factor of 16 shorter for similar to conventional sample conditions. Perhaps more importantly it creates the possibility of collecting data from samples a factor of four lower in concentration. NMR cryoprobes enable higher throughput and/or a consequent increase in the scope of proteins amenable to structural analysis [46]. Losses in the sensitivity of the cryoprobe can be offset by appropriate choice of buffer [47].

Another innovation in probe-design is the advent of smaller tubes, requiring less sample. Conventional NMR probes require a 5 mm glass sample tube whereas newer 3 mm probes also hold the promise of use in biomolecular NMR for smaller mass-limited samples. Although cryogenically cooled triple resonance 3 mm probes for protein NMR spectroscopy are not yet available, the combination promises to be potent as structural genomics evolves into the next stages beyond the "low-hanging fruit". Also promising are micro-probes requiring as little as five microliters of sample (~100–300 μg protein) which promise to handle protein triple resonance NMR spectroscopy [48].

Finally the availability of super-high-field instruments (>800 MHz ^{1}H resonance frequency) also merits mention. Although there is certainly an increase in sensitivity with increasing field, the effect is much smaller than that observed with the cryoprobe, whereas the cost increases exponentially. For example, a 500 MHz magnet with a cryoprobe will most probably give better signal-to-noise ratio than a regular 800 or even 900 MHz magnet. The cost of the 800 MHz system is also several times that of a 500 MHz system; the cost of the 900 MHz instrument, currently the most powerful commercially available, is much higher still. Increasing field also leads to other potential issues, for example chemical shift anisotropy, which might effect biomolecular experiments.

The hardware used is driven by the pulse sequences which determine how the magnetic nuclei in the sample interact. Significant interest has been generated recently with research into acquisition or processing methodology which reduces the time needed for the pulse sequences, because the data acquisition time needed for a single sample is substantial. These procedures, the strengths and weaknesses of which have recently been reviewed [49], are still under development but hold much promise for the future. They include unique data-processing algorithms, for example filter diagonalization, which reduces the number of data points needed while maintaining conventional data acquisition procedures. Also promising are "down-sampling" techniques, such as GFT–NMR, which enables high-quality multidimensional spectra to be acquired in a few hours; for example, a 5D spectrum can, in principle, be obtained in less than

3 h. Also included in this category are single-scan experiments in which multidimensional information is encoded by pulse-field gradients which alter the local magnetic field at different points in the sample, and projection reconstruction experiments which reconstruct a 3D spectrum from a small number of 2D projections. Finally, application of Hadamard matrices to previously known chemical shift frequencies can rapidly provide information about selective regions in the protein; this might be useful for high-throughput analysis of active sites and/or ligand binding sites [49].

12.3.1.3

High-throughput Data Analysis

The calculation of a NMR structure is a nontrivial process based on several steps of assignment with various types of data. Each of these has been influenced by structural genomics and the overall methodology is rapidly evolving [36, 38]. The first step is to assign the chemical shifts for all labeled nuclei in the protein, starting with the backbone and then extending into the sidechains. The chemical shift is a type of NMR signature which is unique to a given nucleus, and as discussed above, isotope labeling is required for ^{13}C and ^{15}N of proteins. Although chemical shifts can be degenerate for specific types of atom, it is unlikely that a series of residues in the protein backbone will have completely identical chemical shifts. Hence by performing experiments which probe the connectivity between backbone residues it is possible to specifically identify the chemical shifts for all residues if the NMR spectrum is complete. In the past this peak-assignment process has been done manually, with the user having to examine the spectra which correlate chemical shift information from the various nuclei and assign resonances manually in a stepwise fashion.

In the past few years several software packages have emerged which promise to alleviate the large amount of user input in the assignment process. These include, but are by no means limited to, semi-automated methods such as the SmartNotebook [50], IBIS [51], and a number of fully automated methods such as AutoAssign [52] and PICS [53]. It should be noted that the success of these methods is always highly dependent on the quality of the input spectra (garbage in = garbage out). Partially folded proteins, or proteins undergoing conformational exchange will cause significant difficulties to both manual and automated assignment, although partial assignments for well-behaved regions can often be accomplished.

Subsequent to the assignment of protein resonances, distance and geometry restraint information is incorporated into simulated annealing calculations to determine a family, or ensemble, of structures. Conventionally, the most popular types of information to include are dihedral angle restraints from experimental data or from comparison with a reference data base [54], hydrogen-bond distance restraints, and, most important, ^{1}H–^{1}H inter-nuclear distances from 2D or 3D nuclear Overhauser effect (NOE) spectroscopy (NOESY) experiments. Although the NOESY experiment is the mainstay of structural determination, it has several disadvantages. First, the upper distance limit of the technique is ~6 Å, and the intensity of NOESY peaks decreases with an r^{-6} dependence, where r is the internuclear distance. As a result, many important long-range NOE restraints are of comparatively very low intensity, and in the past much user input has been required to distinguish those peaks from noise. Further complicating the analysis is the relatively small distribution of proton chemical shifts; thus for proteins there are usually many possible ^{1}H,^{1}H pairs

which could give rise to a particular NOE correlation based on chemical shift alone.

The advent of structural genomics has seen a significant progression in algorithms for automated structure determination, of which several representative methods will be mentioned here. All of the programs currently in popular use are iterative, meaning that an initial guess from either a homologous starting structure or chemical shifts alone is used to generate the initial restraint list, along with other experimental distance and geometry information as described in the preceding paragraph. By successive iterations of simulated annealing and energy minimization the criteria used for selection of the "acceptable" NOE distance restraints is slowly narrowed until a reasonable structural family remains which most closely fits the data. The major implementations are in ARIA [55] and CANDID/CYANA [56] which use the global fold from previous iterations to guide the results of subsequent iterations. A new algorithm has recently been published which tracks instead individual NOE peaks and provides greater flexibility in deciding which NOE should be used in the course of the calculation [57]. Another method, not yet widely available at the time of writing, but which promises to move forward the structural calculation process, is RADAR from Professor Kurt Wuthrich's laboratory. This is a combination of the previously published ATNOS algorithm [58] and CANDID [56] programs which promises to automatically perform pick peaking and complete structural calculations.

Another important experimental advance is the introduction of residual dipolar couplings (RDC) as additional restraints in structure calculations [59]. In this class of experiments the sample of interest is placed in some type of alignment medium (e.g. bicelles, phage, polyacrylamide gels), which permits a small amount of ordering of the protein molecules and prevents completely random tumbling. The ordering results in a defined relationship among all the bond vectors, and an alignment tensor can be defined from which long-range relationships can be determined. This type of long-range restraint, combined with sparse NOE data, significantly reduces the overall time required for complete structure determination. In addition, RDC can be used for other types of classification and in the initial assignment phase, as described in the next section. Finally it should be mentioned in this section that sometimes additional restraints can be obtained from paramagnetic metal ions bound to metalloproteins [60]. As with RDC, these long-range restraints substantially improve the structure of the protein calculated on the basis of NOE data alone.

12.3.2 Other Non-structural Applications of NMR

Although structure determination is currently the most obvious application of NMR in structural genomics, there is a whole host of related non-structural applications of NMR. These range from pre-structural characterization of the extent to which the protein is folded, and rapidly determining the protein fold family, to establishing the merits of pursuing structural characterization. Perhaps the most important role NMR will play in the future is in the post-structural characterization of proteins of unknown function, thereby bridging the gap between structural and functional genomics. NMR is a very powerful tool for performing ligand and/or drug screening, and for elucidating the complex chemical dynamic and exchange characteristics of proteins.

12.3.2.1

Suitability Screening for Structure Determination

By their nature, NMR chemical shifts are highly sensitive to the folded state of proteins. The chemical shifts of residues in α-helical, β-sheet, and random coil conformations are reasonably unique, and can be monitored in a relatively straightforward manner to establish if the protein is well behaved for NMR and/or X-ray structural analysis. Even the most basic of NMR experiments, a 1D ^{1}H spectrum generally contains enough information to establish a semi-quantitative estimate of the extent of folding of a protein on the basis of chemical shift dispersion [61]. This is advantageous to structural genomics applications because these experiments are extremely rapid, the data can be acquired from very small amounts of sample (tens of micromolar) within a few minutes, and they do not require any special labeling of the proteins.

The obvious disadvantage of using 1D NMR is the lack of resolution and overlap of protein resonances in the ^{1}H spectrum. As a result, 2D $^{1}H,^{15}N$ correlation spectroscopy is a common alternative [37]. In this case resonances from backbone HN atoms are resolved on the basis of two chemical shifts, those of protons and nitrogen atoms, providing greater dispersion and a much more

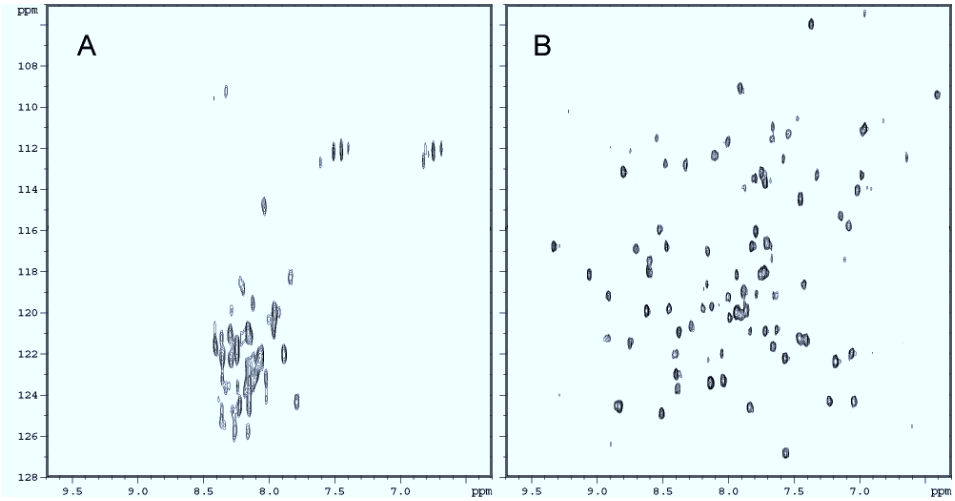


Fig. 12.3 Two-dimensional proton–nitrogen15 correlation spectroscopy of proteins. The panel on the left is for an unfolded protein; note that all the peaks are clustered close together and overlap. The panel on the right is typical of a well-folded protein. In this spectrum the peaks are nicely distributed through the spectrum without overlap. In structural genomics such spectra are being used to decide on the suitability of a protein for structural analysis. The protein giving the spectrum on the left would not be further pursued whereas structure determination would be initiated for the protein giving the spectrum shown in the right panel. These two $^{1}H,^{15}N$ HSQC NMR spectra were obtained for the acyl carrier protein of *Vibrio harveyi*; this protein of 77 residues is normally unfolded (left) but it becomes compactly folded on addition of calcium (right). These two spectra were acquired for unlabeled protein, which could be done because a highly sensitive cryoprobe was used to enable improved sensitivity of detection. Note that the natural abundance of ^{15}N is only 0.4 %. Normally such spectra are collected however for 100 % ^{15}N-labeled proteins on standard NMR probes. This example therefore also illustrates the sensitivity enhancement that can be achieved by the use of cryo-cooled NMR probes.

dramatic visual clue about the extent of folding of the protein (Fig. 12.3). The protein must be expressed in the presence of ^{15}N labeled metabolites, and it is usually purified before screening. If the quantity of protein material is reasonably high, hundreds of micromolar, data acquisition time will still be within tens of minutes for 2D acquisition. In an intriguing structural genomics screening approach, purification has been shown as optional, because the entire cell lysate can be screened provided only the protein of interest is expressed with labeled metabolites [62]. An advantage of screening by either 1D or 2D NMR is that a rapid test is available to search for conditions under which the protein might be folded, for example the presence of metal ions or cofactors (Fig. 12.3). This is important especially as structural genomics moves into the next stages where easily solved proteins are out of the way (the low-hanging fruit), and only more challenging targets remain.

12.3.2.2

Determination of Protein Fold

NMR spectroscopy promises to play a significant role in the rapid classification of fold families and the rapid determination of protein folds. The former is desirable as a method to identify novel folds or determine convergent structural evolution from proteins of different function. One of the stated aims of structural genomics projects is the completion, or near completion of "protein-fold space" [40]. Recently Prestegard's lab has presented a method by which *unassigned* $^{1}H,^{15}N$ RDC can be used to classify proteins into protein families [63]. Such data are relatively simple to acquire, and should it prove successful in large-scale projects, this will be a major step forward toward selecting unique and previously unknown protein folds.

Besides furnishing information on homology with other proteins of known structure, NMR is also a powerful tool for determining protein fold, equivalent to low-resolution structures. These methods have been well reviewed recently in the context of structure-based drug design [64]. The combination of $^{1}H/^{13}C/^{15}N$ chemical shifts of the protein backbone provides a highly reliable measure of secondary structure, and addition of RDC information provides information about the orientation of the secondary structure elements. Several groups have shown that homology modeling is dramatically more successful when this information is incorporated into structural calculations [65, 66]. Particularly striking is an example of the N-terminal domain of apo-calmodulin, a primarily helical protein which is similar in secondary structure to the Ca^{2+}-calmodulin form, although the orientation of the helices is significantly different. A homology model of apo-CaM based on the Ca^{2+} form has a root-mean square deviation (rmsd) of 4.2 Å from the NMR-determined structure, which drops to 1.0 Å when RDC are incorporated [65]. The incorporation of RDC data has been further extended into *ab-initio* protein structure prediction with the program Rosetta [67], which is useful for prediction of protein folds for sequences with no known analogs.

One disadvantage of the approaches listed above for fold determination is the need to assign NMR resonances before incorporating them in structural calculations. Although this is a relatively rapid step, it is still rate-limiting in terms of high throughput structural proteomics. An intriguing, recently suggested solution to this problem is an algorithm designed to use unassigned NMR data [68] combined with the Rosetta program. Although success is not yet sufficiently reliable for completely hands-off

analysis, the concept promises much to be excited about.

12.3.2.3 Rational Drug Target Discovery and Functional Genomics

Drug discovery is the most commercially lucrative medium-term gain evident from structural genomics projects. The same principles used for drug discovery are often applied to "functional genomics", or attempting to understand the function of unknown genes and proteins. The reader is referred to several recent publications [64, 69–72] which describe the relationship between structural genomics and drug discovery; NMR contributions to this field are briefly outlined here. Study of structure–activity relationships (SAR) by NMR was first introduced by Fesik's group who linked two compounds that bound to a ^{15}N-labelled protein [73]. The result was a super-ligand compound which bound with nanomolar affinity, derived from two predecessors which bound with micromolar affinity. This is an example of protein-based screening whereby the protein NMR resonances are monitored and a large number of potential targets can be screened simultaneously to detect binding. Such screening can also be done using a protein with ^{13}C-labelled methyl groups, which can be detected with very high sensitivity, and might be more sensitive than screening with an ^{15}N-labelled protein because sidechains are more often involved in ligand binding than the protein backbone [74]. In an interesting twist, another pharmaceutical group has shown that by selective labeling of amino acids multiple proteins can be simultaneously monitored by NMR, thereby increasing the screening efficiency [75].

Screening by monitoring protein resonances is useful because specific sites and/or residues from the protein can be identified as binding areas for potential ligands. Ideally, however, this requires large quantities of labeled protein with resonances assigned, which can be a non-trivial step. An alternate approach is monitoring potential ligand resonances by NMR while screening with relatively small amounts of unlabelled protein. The methods of this class are predicated on the different physiochemical properties of large and small molecules when observed by NMR, including relaxation rates, diffusion characteristics, and the ability to selectively transfer magnetization [70, 71]. These methods generally work for ligands which bind in the micromolar to millimolar range, although competition experiments will provide information about tighter binding ligands. Besides the obvious advantages of smaller protein quantity, this approach is not restricted to analysis of proteins which provide reasonable NMR spectra, and hence potential problems such as protein size and chemical exchange, which complicate protein spectra, are non-issues when monitoring ligand resonances. It is also worth mentioning that the use of ^{19}F-labelled substrates for competition experiments has shown great promise for biochemical screening, with the obvious advantage of there being no background ^{1}H signals to filter out [76].

12.4 Epilogue

Although the full impact of the public structural genomics efforts is yet to be realized it seems clear that many more distinct protein structures will become available in the near future. In the first instance this will provide much better appreciation of the extent of the protein fold-space. It has, however, also often been suggested that a protein struc-

ture can be directly related to its function. Unfortunately the relationship between protein fold and protein function is rather complex [77, 78]. It is well known that many proteins can have a similar fold, yet have quite different functions, this is well documented for the versatile TIM-barrel motif structures for example. Likewise some proteins carry out closely related functions or enzymatic reactions, yet can have very different folds, as illustrated by the proteinases trypsin and subtilisin. It will therefore be necessary to start looking for specific functional sites on the protein by use of a variety of computational approaches. Programs are now available for identification of such potential sites by mapping the most conserved residues in large groups of homologous proteins. In this fashion it has been possible to identify enzyme active sites or DNA-binding sites, for example [77, 78]. It has also been suggested that that one could attempt to dock all known metabolites on to a novel protein using commonly used computational computer docking approaches [71]. When a group of potential ligands has been identified in this fashion, they have to be confirmed experimentally by NMR-based screening procedures, as described above.

In the near future it will be possible to view simultaneously the 3D structures of all the enzymes involved in one metabolic pathway. An early glimpse of the structures of all the enzymes involved in the glycolytic pathway reveals that unique insights can be obtained in this manner [79]. Also, because many of the early structural genomics trial projects were focused on bacteria, our knowledge of bacterial protein structures has been already increased significantly in a relatively short time. This has enabled researchers to make comparisons between the different classes of protein superfolds for different genomes [80]. The lessons learned from these simpler organisms can, however, also be extended directly to humans, because many protein functions are the same and homologous proteins can be identified and modeled [81]. In this regard it should be noted that methodologies for identifying sequence- and structure-based homologs continue to improve [82]. Nevertheless , it should be realized that homology models of proteins, particularly those obtained with low levels of sequence identity (<30 %), might give an overall fold that is close to correct but is not sufficiently precise to enable drug design or an understanding of enzyme mechanisms [79].

Finally we would like to point out that at this time our understanding of the "ligand-space" is also incomplete. Although many major metabolites have been known and characterized for a long time, new ones that are often present only in small amounts continue to be discovered. This is particularly true for a variety of drug metabolites, that will continue to appear as new drugs enter the market. Fortunately many metabolomics or metabonomics projects are aimed at identifying the complete metabolome in organisms ranging from bacteria to humans and plants [83–85]. These initiatives might have to be completed before we can understand the roles of all the gene-products that are coded for by the genomes.

References

1 Rupp B (2003) High throughput protein crystallography. In: Proteomics (Edwards A, Ed) Marcel Dekker.
2 Goodwill KE, Tennant MG, Stevens RC (2001) High-throughput X-ray crystallography for structure based drug design. Drug Discovery Today 6:S113–S118.
3 Stevens RC (2000) High-throughput protein crystallization. Curr Opin Struct Biol 5:558–563.
4 Stevens RC, Yokoyama S, Wilson IA (2000) Global efforts in structural genomics. Science 294:89–92.
5 Bentley DR (2004) Genomes for medicine. Nature 429:440–445.
6 Venter GC, *et al* (2001) The sequence of the human genome. Science 291:1304–1351.
7 Hope H (1990) Crystallography of biological macromolecules at ultra-low temperatures. Annu Rev Biophys Biophys Chem 19:107–126.
8 Garman EF, Mitchell TR (1997) Macromolecular cryocrystallography. J Appl Crystallogr 30:211–237.
9 Hendrickson W, (2000) Synchrotron Crystallography. Trends Biochem Sci 25:637–643.
10 Ogata CM (1998) MAD phasing grows up. Nature Struct Biol Suppl 5:638–640.
11 Rodgers DW (2001) Cryo-crystallography techniques and devices. Intl Tables Crystallography F:202–208.
12 Sweet R (2000) The technology that enables synchrotron structural biology. Nature Struct Biol Suppl 5:654–656.
13 Garman E, Nave C (2002) Radiation damage to crystalline biological molecules: current views. J Synchrotron Rad 9:327–328.
14 Garman E (1999) Cool data: quantity AND quality. Acta Cryst D55:1641–1653.
15 Doublie S, Carter C (1992) Preparation of selenomethionyl protein crystals. In: Crystallization of nucleic acids and proteins (Ducruix A, Giege R, Eds) Oxford University Press.
16 Lamzin VS, Perrakis A (2000) Current state of automated crystallographic data analysis. Nature Struct Biol 7:979–981.
17 Abola E, Kuhn P, Earnest T, Stevens RC (2000) Automation of X-ray crystallography. Nature Struct Biol Suppl 7:973–977.
18 Hui R, Edwards A (2003) High-throughput protein crystallization. J. Struct. Biol. 142:154–161.
19 Norvell JC, Zapp-Machalek A (2000) Structural genomics programs at the US National Institute of General Medical Sciences. Nature Struct. Biol Suppl. 7:931.
20 Harris TE, (2001) The commercial use of structural genomics. Drug Discovery Today 6:1148.
21 Hosfield D, Palan J, Hilgers M, Scheibe D, McRee DE, Stevens RC (2003) A fully integrated protein crystallization platform for small-molecule drug discovery. J. Struct. Biol. 142:207–217.
22 Goulding CW, Perry LJ (2003) Protein production in E. coli for structural studies by X-ray crystallography. J Struct Biol 142:133–143.
23 Dale GE, Oefner C, D'Arcy A (2003) The protein as a variable in protein crystallization. J Struct Biol 142:88–97.
24 Kigawa T, Yabuki T, Yoshida Y, Tsustui M, Ito Y, Shibata T, Yokoyama S (1999) Cell-free production and stable-isotope labeling of milligram quantities of protein. FEBS Lett. 442:14–19.
25 Wilson, WW (2003) Light-scattering as a diagnostic for protein crystal growth – A practical approach. J. Struct. Biol 142:56–65.

26 Luft JR, Collins RJ, Fehrman NA, Lauricella AM, Veatch CK, DeTitta GT (2003) A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J. Struct. Biol. 142:170–179.

27 DeLucas LJ, Bray TL, Nagy L, McCombs D, Chernov N, Hamrick D, Cosenza L, Belgovskiy A, Stoops B, Chait A (2003) Efficient protein crystallization. J. Struct. Biol. 142:188–206.

28 Cusack S, Belrhali H, Bram A, Burghammer M, Perrakis A, Riekel C (1998) Small is beautiful: protein micro-crystallography. Nature Struct Biol Suppl 5:634–647.

29 Spraggon G, Lesley SA, Kreusch A, Priestle JP (2002) Computational analysis of crystallization trials. Acta Cryst D58:1915–1923.

30 Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002) PHENIX: building new software for automated crystallographic structure determination. Acta Cryst D58:1948–1954.

31 Winn MD, Ashton AW, Briggs PJ, Ballard CC, Patel P (2002) Ongoing developments in CCP4 for high-throughput structure determination. Acta Cryst D58:1929–1936.

32 Holton J, Alber T (2004) Automated protein crystal structure determination using ELVES. Proc Natl Acad Sci USA 101:1537–1542.

33 Loll PJ (2003) Membrane Protein structural biology: the high-throughput challenge. J. Struct. Biol. 142: 144–153.

34 Luca S, Heise H, Baldus M (2003). High-resolution solid-state NMR applied to polypeptides and membrane proteins. *Acc Chem Res* 36:858–865.

35 Montelione GT, Zheng D, Huang YJ, Gunsalus KC, Szyperski T (2000). Protein NMR spectroscopy in structural genomics. *Nat Struct Biol* 7 Suppl:982–985.

36 Kennedy MA, Montelione GT, Arrowsmith CH, Markley JL (2002). Role for NMR in structural genomics. *J Struct Funct Genomics* 2:155–169.

37 Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, Le B, Ramelot T, Lee GM, Bhattacharyya S, Gutierrez P, Denisov A, Lee CH, Cort JR, Kozlov G, Liao J, Finak G, Chen L, Wishart D, Lee W, McIntosh LP, Gehring K, Kennedy MA, Edwards AM, Arrowsmith CH (2002). An NMR approach to structural proteomics. *Proc Natl Acad Sci U S A* 99:1825–1830.

38 Prestegard JH, Valafar H, Glushka J, Tian F (2001). Nuclear magnetic resonance in the era of structural genomics. *Biochemistry* 40:8677–8685.

39 Norin M, Sundstrom M (2002). Structural proteomics: lessons learnt from the early case studies. *Farmaco* 57:947–951.

40 Vitkup D, Melamud E, Moult J, Sander C (2001). Completeness in structural genomics. *Nat Struct Biol* 8:559–566.

41 Savchenko A, Yee A, Khachatryan A, Skarina T, Evdokimova E, Pavlova M, Semesi A, Northey J, Beasley S, Lan N, Das R, Gerstein M, Arrowmith CH, Edwards AM (2003). Strategies for structural proteomics of prokaryotes: Quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. *Proteins* 50:392–399.

42 Staunton D, Owen J, Campbell ID (2003). NMR and structural genomics. *Acc Chem Res* 36:207–214.

43 Yokoyama S (2003). Protein expression systems for structural genomics and proteomics. *Curr Opin Chem Biol* 7:39–43.

44 Ippel JH, Pouvreau L, Kroef T, Gruppen H, Versteeg G, van den PP, Struik PC, van Mierlo CP (2004). In vivo uniform (15)N-isotope labeling of plants: using the greenhouse for structural proteomics. *Proteomics* 4:226–234.

45 Kigawa T, Yamaguchi-Nunokawa E, Kodama K, Matsuda T, Yabuki T, Matsuda N, Ishitani R, Nureki O, Yokoyama S (2002). Selenomethionine incorporation into a protein by cell-free synthesis. *J Struct Funct Genomics* 2:29–35.

46 Monleon D, Colson K, Moseley HN, Anklin C, Oswald R, Szyperski T, Montelione GT (2002). Rapid analysis of protein backbone resonance assignments using cryogenic probes, a distributed Linux-based computing architecture, and an integrated set of spectral analysis tools. *J Struct Funct Genomics* 2:93–101.

47 Kelly AE, Ou HD, Withers R, Dotsch V (2002). Low-conductivity buffers for high-sensitivity NMR measurements. *J Am Chem Soc* 124:12013–12019.

48 Peti W, Norcross J, Eldridge G, O'Neil-Johnson M (2004). Biomolecular NMR using a microcoil NMR probe – new technique for the chemical shift assignment of aromatic side chains in proteins. *J Am Chem Soc* 126:5873–5878.

49 Freeman R, Kupce E (2003). New methods for fast multidimensional NMR. *J Biomol NMR* 27:101–113.

50 Slupsky CM, Boyko RF, Booth VK, Sykes BD (2003). Smartnotebook: a semi-automated approach to protein sequential NMR resonance assignments. *J Biomol NMR* 27:313–321.

51 Hyberts SG, Wagner G (2003). IBIS – a tool for automated sequential assignment of protein spectra from triple resonance experiments. *J Biomol NMR* 26:335–344.

52 Moseley HN, Sahota G, Montelione GT (2004). Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. *J Biomol NMR* 28:341–355.

53 Malmodin D, Papavoine CH, Billeter M (2003). Fully automated sequence-specific resonance assignments of hetero-nuclear protein spectra. *J Biomol NMR* 27:69–79.

54 Cornilescu G, Delaglio F, Bax A (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 13:289–302.

55 Linge JP, Habeck M, Rieping W, Nilges M (2003). ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* 19:315–316.

56 Herrmann T, Guntert P, Wuthrich K (2002). Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* 319:209–227.

57 Kuszewski J, Schwieters CD, Garrett DS, Byrd RA, Tjandra N, Clore GM (2004). Completely Automated, Highly Error-Tolerant Macromolecular Structure Determination from Multidimensional Nuclear Overhauser Enhancement Spectra and Chemical Shift Assignments. *J Am Chem Soc* 126:6258–6273.

58 Herrmann T, Guntert P, Wuthrich K (2002). Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J Biomol NMR* 24:171–189.

59 Lipsitz RS, Tjandra N (2004). Residual dipolar couplings in NMR structure analysis. *Annu Rev Biophys Biomol Struct* 33:387–413.

60 Baig I, Bertini I, Del Bianco C, Gupta YK, Lee YM, Luchinat C, Quattrone A (2004). Paramagnetism-based refinement strategy for the solution structure of human alpha-parvalbumin. *Biochemistry* 43:5562–5573.

61 Rehm T, Huber R, Holak TA (2002). Application of NMR in structural proteomics: screening for proteins amenable to structural analysis. *Structure (Camb)* 10:1613–1618.

62 Galvao-Botton LM, Katsuyama AM, Guzzo CR, Almeida FC, Farah CS, Valente AP (2003). High-throughput screening of structural proteomics targets using NMR. *FEBS Lett* 552:207–213.

63 Valafar H, Prestegard JH (2003). Rapid classification of a protein fold family using a statistical analysis of dipolar couplings. *Bioinformatics* 19:1549–1555.

64 Powers R (2002). Applications of NMR to structure-based drug design in structural genomics. *J Struct Funct Genomics* 2:113–123.

65 Chou JJ, Li S, Bax A (2000). Study of conformational rearrangement and refinement of structural homology models by the use of heteronuclear dipolar couplings. *J Biomol NMR* 18:217–227.

66 Meiler J, Peti W, Griesinger C (2000). DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts. *J Biomol NMR* 17:283–294.

67 Rohl CA, Baker D (2002). De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc* 124:2723–2729.

68 Meiler J, Baker D (2003). Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci U S A* 100:15404–15409.

69 Mueller L, Montelione GT (2002). Structural genomics in pharmaceutical design. *J Struct Funct Genomics* 2:67–70.

70 Stockman BJ, Dalvit C (2002). NMR screening techniques in drug discovery and drug design. *Progress in Nuclear Magnetic Resonance Spectroscopy* 41:187–231.

71 Parsons L, Orban J (2004). Structural genomics and the metabolome: combining computational and NMR methods to identify target ligands. *Curr Opin Drug Discovery Dev* 7:62–68.

72 Buchanan SG (2002). Structural genomics: bridging functional genomics and structure-based drug design. *Curr Opin Drug Discov Devel* 5:367–381.

73 Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996). Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274:1531–1534.

74 Hajduk PJ, Meadows RP, Fesik SW (1999). NMR-based screening in drug discovery. *Q Rev Biophys* 32:211–240.

75 Zartler ER, Hanson J, Jones BE, Kline AD, Martin G, Mo H, Shapiro MJ, Wang R, Wu H, Yan J (2003). RAMPED-UP NMR: multiplexed NMR-based screening for drug discovery. *J Am Chem Soc* 125:10941–10946.

76 Dalvit C, Fagerness PE, Hadden DT, Sarver RW, Stockman BJ (2003). Fluorine-NMR experiments for high-throughput screening: theoretical aspects, practical considerations, and range of applicability. *J Am Chem Soc* 125:7696–7703.

77 Whisstock JC, Lesk AM (2003). Prediction of protein function from protein sequence and structure. *Q Rev Biophys* 36:307–340.

78 Jones S, Thornton JM (2004). Searching for functional sites in protein structures. *Curr Opin Chem Biol* 8:3–7.

79 Erlandsen H, Abola EE, Stevens RC (2000). Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites. *Curr Opin Struct Biol* 10:719–730.

80 Frishman D (2003). What we have learned about prokaryotes from structural genomics. *OMICS* 7:211–224.

81 Matte A, Sivaraman J, Ekiel I, Gehring K, Jia Z, Cygler M (2003). Contribution of structural genomics to understanding the biology of Escherichia coli. *J Bacteriol* 185:3994–4002.

82 Goldsmith-Fischman S, Honig B (2003). Structural genomics: computational methods for structure analysis. *Protein Sci* 12:1813–1821.

83 Pelczer I (2003). Structural biology, ligand binding, metabomics – the changing face of high-field, high-resolution NMR spectroscopy. *Theochem* 666/667:499–505.

84 Nicholson JK, Connelly J, Lindon JC, Holmes E (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev Drug Discov* 1:153–161.

85 Sumner LW, Mendes P, Dixon RA (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* 62:817–836. .

Part III
Bioinformatics

13
Bioinformatics Tools for DNA Technology

Peter Rice

13.1
Introduction

Other chapters in this book concentrate on the underlying technologies. This chapter covers the essentials of using bioinformatics tools to analyze DNA sequences. The examples are drawn from the open-source EMBOSS sequence-analysis package [1] developed over the past few years by the a core team of developers on the Hinxton Genome Campus and many contributors from the bioinformatics community.

The nature of DNA sequence data has changed substantially over the years since DNA sequencing became a common laboratory method. The earliest sequences covered just one gene of interest, either from a bacterial clone or from a cDNA clone. The sequences were finished to a reasonable standard, and at 1500 bases on average were an ideal size for analysis by computer. During the early 1990s, EST sequencing projects led to a vast increase in the number of short, single-read sequences with high error rates both for single base substitutions and for single base insertions or deletions. These proved especially difficult for alignment methods. In recent years we have seen massive growth in the release of large clones and complete genomes, so that the databases also contain many sequences of 100 kb or larger. These also lead to major problems for computer methods. This chapter will illustrate some of the solutions that have been applied to these new sequence-analysis problems.

For DNA sequence analysis, the main methods can be divided into those based on alignment and those based on pattern matching.

13.2
Alignment Methods

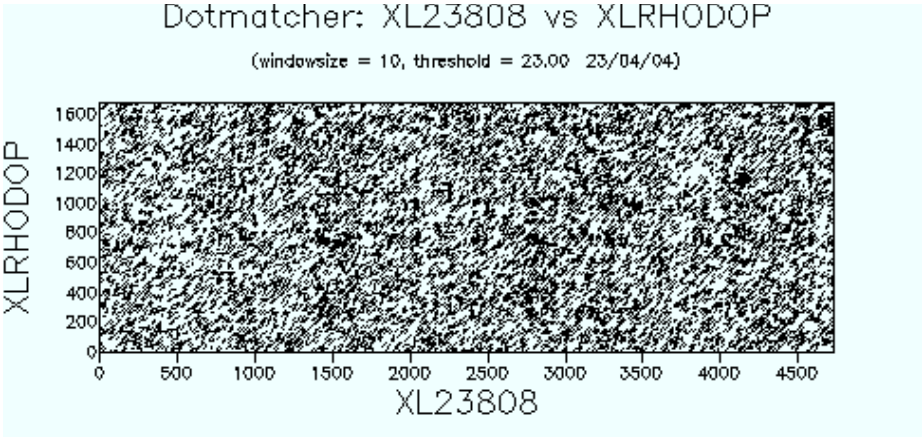
Sequence alignment assumes there are one or more other sequences that can be compared with our starting sequence. The most obvious example is a search of the sequence databases, where we look for any other known sequence that has some similarity to ours, in other words one that can be aligned to it. Having found other sequences that could be similar, we can align them to our starting sequence either one at a time (pairwise alignment) or all together (multiple alignment).

A graphical display of the matches between two sequences is often a good start-

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

Example dotmatcher



ing point, because this is the easiest way to see where there are potential regions of similarity. Most sequence-analysis packages will have one or more programs to produce these plots, known by the generic name of "dot plots". In a dot plot, one sequence is plotted on each axis and a dot is drawn where the two sequences "match". There are, of course, choices to be made about how a match is defined. The simplest is to pit a dot where a base in one sequence matches a base in the other sequence. With only 4 bases to choose from, however, this automatically puts dots in 25 % of the possible places and fills the plot so any real matches become impossible to see.

The general rule for dot plots is to have about one dot for each base in the shorter sequence, so that real matches are obvious as lines on the plot. Some programs do this by counting several bases together, and putting a dot when enough of them match. With suitable scoring methods this can work very well and show similarities between very distantly related sequences.

In the EMBOSS program dotmatcher, short regions (in the example below, 10 bases) are compared, and a dot is plotted when the number of matches exceeds a set score. A reasonable scoring system would require at least half of the residues to be identical. As EMBOSS scores +5 for each match, this would be a score of 25.

A particularly tricky problem, surprisingly, is comparing spliced cDNA sequences with genomic (unspliced) sequences. The reasons will become clearer when we consider pairwise alignment below. A graphical dot plot view, however, makes the problem much simpler. In the example we use a window of 10 bases, enough to avoid "random" matches (unless there are simple repeats in both sequences), and insist on perfect matches with a score of 50 over 10 bases). The five exons in the example sequences are then easy to see.

13.2.1 Pairwise Alignment

To see, base-by-base, how two sequences are related, we need to carry out a rigorous alignment of the two sequences, using string-matching techniques from computer science. The standard alignment method is that of dynamic programming. The name refers to the way in which the result is calculated automatically as the sequences are

compared base by base. Scores for each comparison are stored in a table, like a spreadsheet, inside the program. These individual scores are then used to build an alignment score, stepping through the table from beginning to end. The most favorable alignment is simply calculated by tracing back from the highest value in the table.

There are two very similar methods, one for global alignment [2] and another for local alignment [3]. Of the two, the local alignment is usually preferred in bioinformatics because it will show the most significant alignment between the two sequences. In most practical examples only part of the sequences will be highly conserved. The global alignment forces the entire length of both sequences to be compared, and a long stretch of low or no similarity can prevent even a strong local high-scoring match from being found.

A key part of these alignment methods is the way in which insertions and deletions ("gaps") in one or other sequence are scored. The favored method is to provide three score values to the program. The first is a comparison matrix which gives a single score for every possible match or mismatch between two bases, including, if appropriate, the nucleotide ambiguity codes. The second score is a penalty to be subtracted each time a gap is made in one sequence so that two other matching regions can be better aligned. The third score is a penalty to be subtracted each time a gap is extended by another base. Clearly the sizes of the gap penalty and the gap extension penalty depend on the score values in the comparison matrix. Mainly for this reason, most programs will hide the comparison matrix from the user and prompt only for a pair of gap penalties.

In EMBOSS the Needleman–Wunsch method is implemented in a program called needle. Sequence alignments for DNA are typically very long. The example below, purely for illustration, uses two short sequences.

In EMBOSS, the standard DNA comparison matrix has a simple score of +5 for a match between a base in each sequence, and _4 for a mismatch. The same scoring is used by the BLAST database search program. Typical gap penalties are 10 for opening a gap, so it will be worth opening a gap to enable an extra two matches to be made, and 0.5 for a gap extension so that up to 10 gaps can be added to create one additional match.

The score is calculated as follows: 7 matches (marked by the vertical lines) score +5 each, a total of 35. One mismatch (T to A) scores –4. One gap scores –10, made up from –10 for the gap, and 0 for its length of 1. The resulting score is 21. We know from the dynamic programming method that this is the highest alignment score possible

Example needle

```
########################################
# Program:  needle
# Rundate:  Fri Apr 23 14:01:15 2004
# Align_format: srspair
# Report_file: needle1.out
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: SHORTA
# 2: SHORTB
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 10
# Identity:       7/10 (70.0%)
# Similarity:     7/10 (70.0%)
# Gaps:           2/10 (20.0%)
# Score: 21.0
#
#
#=======================================

SHORTA              1 ATTACCACAT     10
                        .|| |||||
SHORTB              1  ATA-CACAT      8


#---------------------------------------
#---------------------------------------
```

for these two sequences (and this scoring method), although it is possible there could be more than one solution.

Although it is tempting to align the As at the starts of the sequences, this would produce a lower score because of the high penalty for adding another gap. This calculation is left as an exercise for the reader. The score would be 20.00. The gap penalty system is designed to allow only a relatively small number of gaps.

There are instances when we would prefer to allow a large number of gaps, but to keep them short. The most common is in comparison of single sequencing reads in which single base insertions or deletions are quite likely. We can achieve this result relatively easily by changing the gap penalties. For example, scoring –1 for starting a gap and –0.5 for each extra gap position. Now we find that aligning the As at the starts of the sequences is possible, because the gap penalty is low enough to allow two gaps even in such short sequences.

13.2.2 Local Alignment

For most DNA sequence comparisons, we are not really interested in aligning the full length of both sequences. Fortunately, simple adjustment of the dynamic programming method enables us to find the best alignment anywhere within the two sequences. The program looks for the highest score anywhere in the results table, and traces through to the position where the score falls below zero. This method, first proposed by Temple Smith and Michael Waterman, requires that mismatches are given a negative score, something which all alignment programs will include in their scoring methods.

Example water

```
########################################
# Program:  needle
# Rundate:  Fri Apr 23 14:01:15 2004
# Align_format: srspair
# Report_file: needle2.out
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: SHORTA
# 2: SHORTB
# Matrix: EDNAFULL
# Gap_penalty: 1.0
# Extend_penalty: 0.5
#
# Length: 10
# Identity:       8/10 (80.0%)
# Similarity:     8/10 (80.0%)
# Gaps:           2/10 (20.0%)
# Score: 38.0
#
#
#=======================================

SHORTA             1 ATTACCACAT     10
                       | || |||||
SHORTB             1 A-TA-CACAT      8


#---------------------------------------
#---------------------------------------
```

```
########################################
# Program:  water
# Rundate:  Fri Apr 23 14:01:15 2004
# Align_format: srspair
# Report_file: water1.out
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: SHORTA
# 2: SHORTB
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 5
# Identity:       5/5 (100.0%)
# Similarity:     5/5 (100.0%)
# Gaps:           0/5 ( 0.0%)
# Score: 25.0
#
#
#=======================================

SHORTA             6 CACAT     10
                       |||||
SHORTB             4 CACAT      8


#---------------------------------------
#---------------------------------------
```

```
########################################
# Program:  water
# Rundate:  Fri Apr 23 14:01:16 2004
# Align_format: srspair
# Report_file: water2.out
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: SHORTA
# 2: SHORTB
# Matrix: EDNAFULL
# Gap_penalty: 1.0
# Extend_penalty: 0.5
#
# Length: 10
# Identity:       8/10 (80.0%)
# Similarity:     8/10 (80.0%)
# Gaps:           2/10 (20.0%)
# Score: 38.0
#
#
#=======================================

SHORTA             1 ATTACCACAT     10
                       | || |||||
SHORTB             1 A-TA-CACAT      8


#---------------------------------------
#---------------------------------------
```

The EMBOSS program for Smith–Waterman alignment is called water. For our short sequence example, the local alignment gives a higher score than the Needleman–Wunsch global method. The reason is simple – the mismatch at the start of the global alignment makes the score worse, and the local alignment stops when it traces back this far.

13.2.3 Variations on Pairwise Alignment

Many sequence analysis packages offer only these simple global and local alignment methods. EMBOSS has some useful extensions which are easily programmed by making small changes to the standard programs. One, merger by Gary Williams from HGMP, merges two sequences by finding the highest scoring overlap.

Example merger

```
########################################
# Program:  merger
# Rundate:  Fri Apr 23 14:01:16 2004
# Align_format: simple
# Report_file: merger.out
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: LONGA
# 2: LONGB
# Matrix: EDNAFULL
# Gap_penalty: 50.0
# Extend_penalty: 5.0
#
# Length: 41
# Identity:       8/41 (19.5%)
# Similarity:     8/41 (19.5%)
# Gaps:          33/41 (80.5%)
# Score: 40.0
#
#
#=======================================

LONGA              1                      agagacatattactagata     19
                                          ||||||||
LONGB              1 tattgcgcagttgcagatcgcgagagacat                30


#---------------------------------------
#
# LONGA position base         LONGB position base     Using
#
#
#---------------------------------------
```

Setting gap penalties seems complicated at first but is really easy when broken down into simple rules.

- Large gap penalties will make it hard to include more than one matching segment in the highest scoring result, and will tend to produce short, local alignments.
- Small gap penalties will make it easy to include gaps to extend a match, and will tend to produce a global alignment even if only a short local match occurs in practice.
- A low gap extension penalty will tend to produce a few very long gaps. A gap extension penalty of 0 is allowed, and can be useful for very long gaps such as introns when comparing a cDNA with a genomic sequence.
- A low gap open penalty will tend to produce a large number of very short gaps. A gap open penalty of 0 is allowed, and can be useful for allowing many single base gaps when aligning a single read to a finished sequence.

13.2.4 Beyond Simple Alignment

Only one set of gap penalties is allowed at any one time. For more complicated situations a specialized application is needed. For example, to align an EST sequence (many short gaps from poor quality sequence and a few long gaps from intron excision) to a genomic sequence can be done with the EMBOSS application *est2genome*. This program, by Richard Mott at the Sang-

Example est2genome

```
Note Best alignment is between forward est and forward genome,
     and splice sites imply forward gene

Exon    469 100.0 1182 1650 XL23808    2  470 XLRHODOP X.laevis rhodopin
+Intron -20   0.0 1651 1898 XL23808
Exon    169 100.0 1899 2067 XL23808  471  639 XLRHODOP X.laevis rhodopin
+Intron -20   0.0 2068 2668 XL23808
Exon    166 100.0 2669 2834 XL23808  640  805 XLRHODOP X.laevis rhodopin
+Intron -20   0.0 2835 3084 XL23808
Exon    240 100.0 3085 3324 XL23808  806 1045 XLRHODOP X.laevis rhodopin
+Intron -20   0.0 3325 4029 XL23808
Exon    639 100.0 4030 4668 XL23808 1046 1684 XLRHODOP X.laevis rhodopin

XL23808 vs XLRHODOP:

 XL23808   2015 acatggatcatggctttgtcttgtgctgctcctcctctcttcggatggtc   2064
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
XLRHODOP    587 acatggatcatggctttgtcttgtgctgctcctcctctcttcggatggtc    636

 XL23808   2065 caggtaaa.....tgcagatacatcccagagggaatgcaatgctcatgcg   2700
              |||>>>>> 601 >>>>>|||||||||||||||||||||||||||||||||
XLRHODOP    637 cag..............atacatcccagagggaatgcaatgctcatgcg    671

 XL23808   2701 gagtagactactacacactgaagcctgaggtcaacaatgaatcctttgtt   2750
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
XLRHODOP    672 gagtagactactacacactgaagcctgaggtcaacaatgaatcctttgtt    721

 XL23808   2751 atctacatgttcattgtccacttcaccattcccctgattgtcatcttctt   2800
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
XLRHODOP    722 atctacatgttcattgtccacttcaccattcccctgattgtcatcttctt    771

 XL23808   2801 ctgctatggtcgcctgctctgcactgtcaaagaggtgag.....catagg   3085
              ||||||||||||||||||||||||||||||||||>>>>> 250 >>>>>|
XLRHODOP    772 ctgctatggtcgcctgctctgcactgtcaaagag..............g    806

 XL23808   3086 ctgcagcccagcaacaggaatctgctaccacccagaaggctgagaaagag   3135
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
XLRHODOP    807 ctgcagcccagcaacaggaatctgctaccacccagaaggctgagaaagag    856
```

er Center [4], includes splice site consensus sequences and two different gap-scoring systems for introns and exons.

One problem with the dynamic programming method is that it does not scale to large sequences. It uses a comparison table to match each base of each sequence. For a typical bacterial gene or cDNA reading frame this could be 1000 bases in each sequence, or 1,000,000 possible comparisons. As the sequences grow longer this quickly becomes too big a problem to fit in the computer's memory or to be computed in a reasonable time.

For large sequences, for example BAC clones or complete genomes, a short cut is needed. One of the most common is to look only at localized ungapped alignments by comparing the two sequences in short stretches and then looking for ways to extend the match region but without allowing gaps, so the problem is far easier to compute. Where a gap appears in a long region of similarity the expectation is that there will be two ungapped matches reported. Even so, too many gaps will make matches impossible to detect.

A particularly rapid way of finding local ungapped matches is to look for "words" in common between two sequences [5]. One sequence is converted to a list of all possible sequences of, for example, six bases and each of these is then sought in the second sequence. Where there is a match, the neighboring bases can be checked to find whether the match can be extended. It is then very easy to look for clusters of high scores and to report the best matches. Sequences of up to complete bacterial genome size can be compared in a few seconds by this method. For increased speed and flexibility, some programs build a complete index of the words and look up the values instead of computing them each time the program runs. This can take a lot of disk space and a lot of time when the indices are calculated).

Multiple alignment is known to be a difficult problem. Clearly, the dynamic programming method used for aligning two sequences is not practical for more. Just three sequences of 1000 bases each would need 1,000,000,000 comparisons to be made and stored. Instead the problem is usually broken down into a series of pairwise matches which are combined to produce a final alignment [6].

13.2.5 Other Alignment Methods

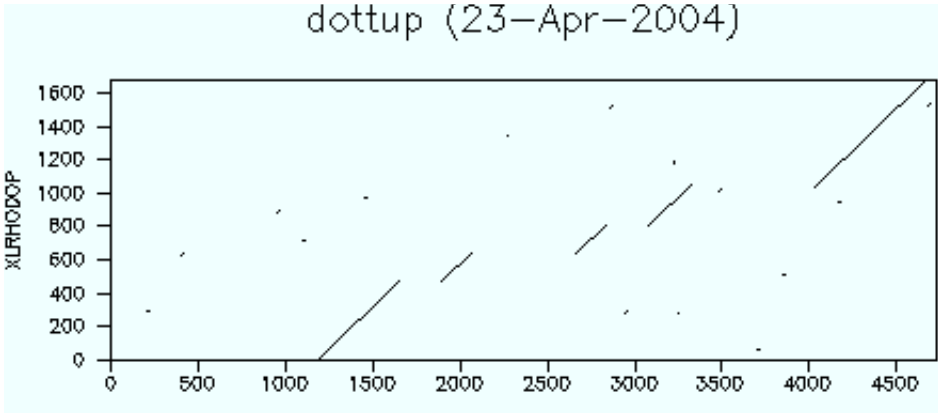
To use all possible alignments, there are other methods. The most popular currently is to create a hidden Markov model (HMM) [7] which "learns" the characteristics of a set of sequences, including the most significant similarities between them. By aligning the original sequences to the HMM we can get a sequence alignment, and by searching a sequence database with the HMM we can search for new sequences to add to the alignment. HMM have been very successfully used for protein domain analysis (Pfam) and have also been applied to gene prediction in bacteria as a pattern-matching method (see below).

13.3 Sequence Comparison Methods

An alternative to full alignment of two sequences is to search for common patterns between them. One of the fastest to calculate is the occurrence of long subsequences ("k-tuples" or "words").

We start again with dot plot methods. The EMBOSS dottup program calculates word matches between two sequences and displays them on a dot plot. The first step is

Example dottup



to build a complete list of all the words in one sequence and make this into an indexed table. Then, for each word in the second sequence, a simple lookup in the table shows every match in the first sequence. With only a few thousand table lookups we have covered several million possible dot positions. This method scales up exceptionally well and can be used up to whole chromosome scale sequences.

Graphical views of the results are all very well, but we also need to have the results as sequence positions we can work with in other programs. Another EMBOSS program, wordmatch, uses the same method as dottup but instead reports the start positions in each sequence and the length. Applied to our earlier cDNA/genomic sequence example we find that the exons are reported almost immediately.

The method does have some slight drawbacks, compared with the rigorous est2genome program. For example, if the next base after the exon is the same in each sequence the match will be extended a little. Also, mismatches or insertions and deletions in the exons will produce two or more shorter matches, but est2genome needs much more time to produce its rigorous result.

Example wordmatch

```
########################################
# Program:  wordmatch
# Rundate:  Fri Apr 23 14:01:20 2004
# Align_format: match
# Report_file: wordmatch.out
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: XL23808
# 2: XLRHODOP
#=======================================

   642 XL23808             4027..4668     XLRHODOP            1043..1684
   471 XL23808             1182..1652     XLRHODOP              2..472
   242 XL23808             3083..3324     XLRHODOP             804..1045
   170 XL23808             1898..2067     XLRHODOP             470..639
   170 XL23808             2666..2835     XLRHODOP             637..806

#---------------------------------------
#---------------------------------------
```

13.3.1
Multiple Pairwise Comparisons

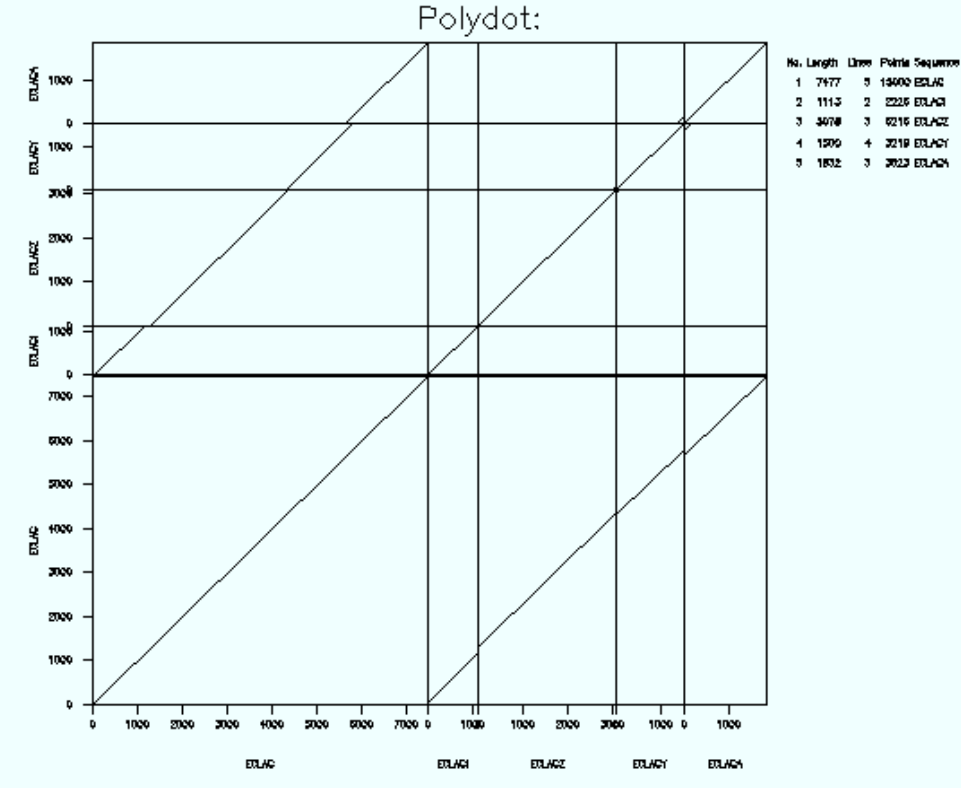
We can extend these rapid word-based methods to achieve very fast comparisons of very long sequences. This is implemented in EMBOSS as the polydot program, which is the dottup method for self comparison of a set of sequences. These could be contigs from a fragment assembly project to check for overlaps (the original use of this program),or complete chromosomes to check for possible duplications.

In the example, we see the *Escherichia coli* lac operon (EMBL entry ECLAC) compared with the sequences of the four individual genes lacI, lacZ, lacY, and lacA.

Just like dottup and wordmatch, polydot has a companion program which produces a report of all the match positions. This program, seqmatchall, reports the alignment length, and the start and end positions in each sequence.

The matches (apart from self matches) are shown below. They include each of the individual gene sequences matched to the complete operon, plus two short matches between lacY and the flanking sequences.

Example polydot



Example seqmatchall

```
########################################
# Program:  seqmatchall
# Rundate:  Fri Apr 23 14:01:21 2004
# Align_format: match
# Report_file: seqmatchall.out
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: ECLAC
# 2: ECLACI
#=======================================

  1113 ECLAC                 49..1161     ECLACI                 1..1113

#=======================================
#
# Aligned_sequences: 2
# 1: ECLAC
# 2: ECLACZ
#=======================================

  3078 ECLAC               1287..4364     ECLACZ                 1..3078

#=======================================
#
# Aligned_sequences: 2
# 1: ECLAC
# 2: ECLACY
#=======================================

  1500 ECLAC               4305..5804     ECLACY                 1..1500

#=======================================
#
# Aligned_sequences: 2
# 1: ECLAC
# 2: ECLACA
#=======================================

  1832 ECLAC               5646..7477     ECLACA                 1..1832

#=======================================
#
# Aligned_sequences: 2
# 1: ECLACZ
# 2: ECLACY
#=======================================

    60 ECLACZ              3019..3078     ECLACY                 1..60

#=======================================
#
# Aligned_sequences: 2
# 1: ECLACY
# 2: ECLACA
#=======================================

   159 ECLACY              1342..1500     ECLACA                 1..159

#---------------------------------------
#---------------------------------------
```

13.4 Consensus Methods

Consensus methods take a sequence alignment and calculate a consensus sequence which represents all the alignment members. These consensus sequences are less useful than the alignment or a weight matrix of hidden Markov model derived from it, but are often found in pattern databases.

Consensus sequences make use of base codes beyond the familiar A, C, G, and T. Each possible combination of bases has its own one-letter code, and with a little practice it is relatively easy to learn them all. The full set of codes is listed in Tab. 13.1, below. For the two-base alternatives some knowledge of base chemistry if needed. Surprisingly, the three-base codes are extremely easy to learn although most biologists will not know them.

Table 13.1 Nucleotide base code table.

Code	*Base(s)*	*Mnemonic*
A	A	Adenine
C	C	Cytidine
G	G	Guanine
T	T or U	Thymine
U	T or U	Uracil (RNA equivalent of T)
R	A or G	Pu-R-ine
Y	C or T/U	p-Y-rimidine
S	C or G	Strong H-bonding
W	A or T/U	Weak H-bonding
K	G or T/U	K-eto group
M	A or C	a-M-ino group
B	C or G or T/U	Not A
D	A or G or T/U	Not C
H	A or C or T/U	Not G
V	A or C or G	Not T or U
N	A or C or G or T/U	a-N-y base
. or –	Gap	

13.5 Simple Sequence Masking

The simplest sequence patterns to detect are often those that cause the most problems. Simple repeats, or runs of single bases, are usually removed before attempting to run a sequence database search, because otherwise the highest scoring hits will be for other sequences with the same simple repeat and any functionally significant matches will be lost. The most common filter for DNA sequences is the "dust" program from NCBI.

13.6 Unusual Sequence Composition

Using the word-based methods familiar from the sequence comparison section above, we can easily identify the most common sub-sequences of any given length. Surprisingly, many genomes do show a strong bias in their composition of sequences ranging in size from two bases to 11 or more.

Example wordcount

```
ccagc 310
tggcg 295
gccag 281
cagcg 281
cagca 273
ctggc 268
ttttt 265
cgcca 263
cgccg 245
gctgg 239
ctagc 6
gctag 6
ctagt 3
actag 3
ctaga 2
cctag 2
ctagg 1
tctag 1
```

The genome of *Escherichia coli* is known to have strong biases for certain 4- and 5-base sequences. The wordcount program reports the frequencies of all short sequences (words) of a given size. This example run on sequence ECUW87 from the EMBL database shows that the rarest five base sequences all contain the four base sequence ctag, and the most common include sequences close to CCAGG or CCTGG [8].

13.7 Repeat Identification

Other programs can find less obvious repeats, or can look for repeats in particular categories. In EMBOSS, the problem of finding tandem (direct) repeats is simplified by breaking it into two steps.

The first program, equicktandem (R. Durbin, personal communication), rapidly scans a sequence and reports possible similarities for a range of possible repeat sizes. In the example below, a run with a range of two to ten bases, it reports a possible repeat size of six bases in a sequence.

The result of equicktandem does not tell us where the repeat was found, and does not guarantee that the repeat will be sufficiently conserved to be identified. The size value can, however, be used by another program, etandem (R. Durbin, personal communication), to identify the precise positions of exact or almost exact repeat runs.

Direct repeats are relatively easy to find, mainly because they appear one after another without gaps. Another program, einverted (R. Durbin, personal communication), looks for almost perfect inverted repeats. This is a trickier problem, because the repeats can be very long and typically have a gap between the two repeated sequences. By limiting the size of the central gap we can significantly increase the speed of the calculation.

Example etandem

```
########################################
# Program: etandem
# Rundate: Fri Apr 23 14:19:04 2004
# Report_format: table
# Report_file: etandem.out
########################################

#=======================================
#
# Sequence: HHTETRA     from: 1   to: 1272
# HitCount: 5
#
# Threshold: 20
# Minrepeat: 6
# Maxrepeat: 6
# Mismatch: No
# Uniform: No
#
#=======================================

  Start     End   Score   Size  Count Identity Consensus
    793     936     120      6     24     93.8 acccta
    283     420      90      6     23     84.8 taaccc
    432     485      38      6      9     90.7 ccctaa
    494     529      26      6      6     94.4 ccctaa
    568     597      24      6      5    100.0 aaccct

#---------------------------------------
#---------------------------------------
```

Example equicktandem

```
########################################
# Program: equicktandem
# Rundate: Fri Apr 23 14:18:08 2004
# Report_format: table
# Report_file: equicktandem.out
########################################

#=======================================
#
# Sequence: HHTETRA     from: 1   to: 1272
# HitCount: 1
#
# Threshold: 20
# Maxrepeat: 600
#
#=======================================

  Start     End   Score   Size  Count
    191     935     339      6    124

#---------------------------------------
#---------------------------------------
```

Example inverted

```
Score 46: 18/20 ( 90%) matches, 0 gaps
    5232 aaaaaacaaaaaaaaggatg 5251
         |||||| |||||||| ||||
    5770 ttttttttttttttttctac 5751

Score 47: 21/25 ( 84%) matches, 0 gaps
    9956 aaatgttagcaacttcaggaagttt 9980
         |||||||| || ||||| | |||||
   12482 tttacaattgtagaagtacatcaaa 12458

Score 80: 44/51 ( 86%) matches, 2 gaps
   12246 ctcctgcctcag-cctccaagtagctgggattaca-gcatgtgccaccatgcc 12296
         |||||| ||||| | |||||   |||||||||||| ||||| |||||||| ||
   13938 gaggacagagtcagaaggtttcacgaccctaatgtccgtactcggtggtatgg 13886
```

Example palindrome

```
Palindromes of:  HSTS1
Sequence length is: 18596
Start at position: 1
End at position: 18596
Minimum length of Palindromes is: 40
Maximum length of Palindromes is: 100
Maximum gap between elements is: 100
Number of mismatches allowed in Palindrome: 10


Palindromes:
60    gaggttgcagtgagccgagatcgcgccactgcactccagcctgggtgagagagcgagactct   121
      |||||||||||||||| |||||   ||| || |||||||||||||| || ||||| ||||| |
281   ctccaacgtcactcggttctaggctggtaacatgaggtcggacccgctgtctcgttctgaca   220

115   agactctgtctcaaaaaaaaaaaaaaaaagaccgccagggct   155
      | |         ||||||||||||||||||||||||||||||
216   ttttttttttttttttttttttttttttctggcggtcccga   176

12672 gatcacgccattgcactccatcctgggcaacagagtgagaccctgaat 12719
      |||||  ||| ||||||||| |||||||||||| |||||||  ||  |
12867 ctagtttggtgacgtgaggtcggacccgttgtcccactctgacacaga 12820
```

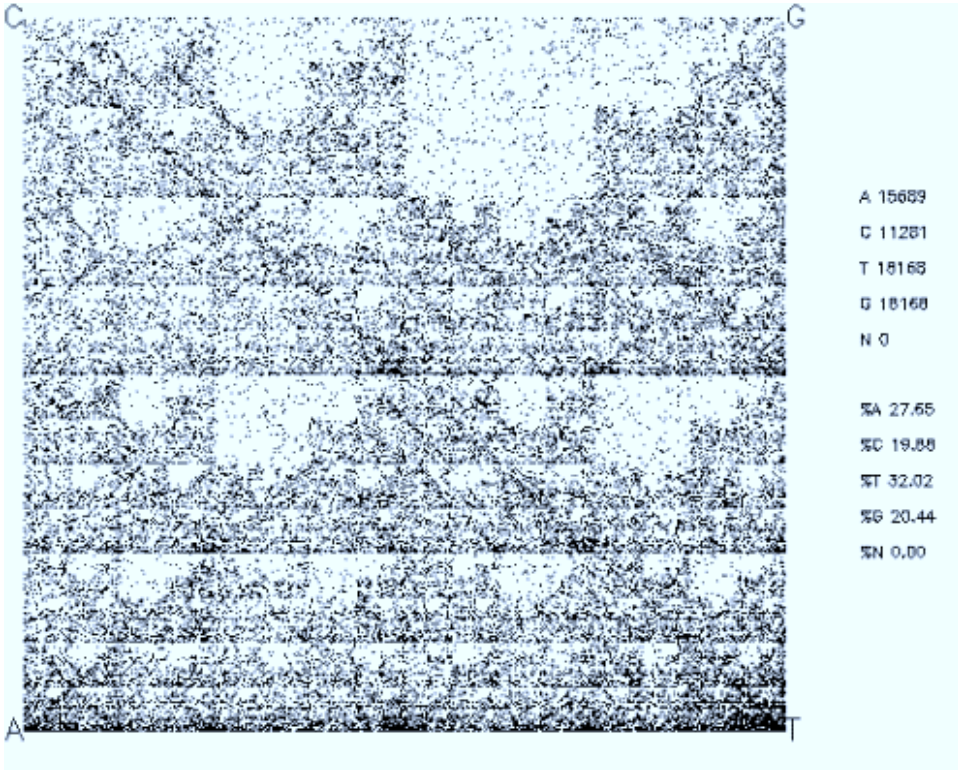
If we are less forgiving of mismatches and gaps, we can make a more exhaustive search for inverted repeats. The program palindrome offers such an alternative approach.

13.8 Detection of Patterns in Sequences

Some applications look for specific patterns in a sequence and can be very useful when first analyzing a new sequence. One of the simplest to program, and yet also one of the most effective for detecting unusual composition in large sequences, is the "chaos game representation" plot [9]. Although the plot looks complicated and difficult to understand, it is really very simple. The example plot shows a typical result of a chaos game plot for human genomic DNA. Clearly, there is a pattern in the sequence although at first sight it is not obvious what the pattern could be.

The chaos game method makes everything clear. Imagine that the program is running on the sequence ATCCG. Now start in the center of the plot. The first base is "A", so go half way to the "A" corner and draw a dot. The next base is "C" so go half

Example chaos



way from this position to the "T" corner and draw a dot. Now go half way to the "C" corner and draw a third dot. Then half way again to the "C" corner and draw a fourth dot. The next move, half way to the "G" corner, will land in the blank area on the plot.

The reason is simple. Moving "half way to the G corner" will always land in the top right quarter of the plot, no matter where you start. But if the previous base was a "C" you start in the top left quarter of the plot, and always move to the blank region. In fact, a sequence that ends with "CG" will always produce a dot in this 1/16 area. The pattern is simply a result of the under-representation of the dinucleotide "CG" in the human genome.

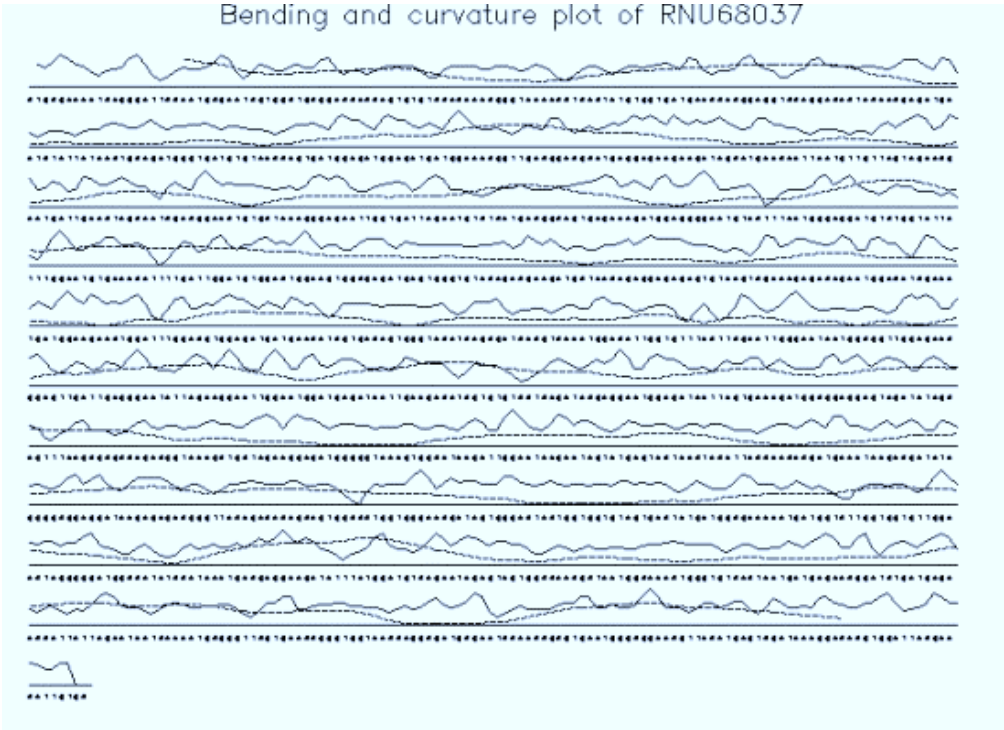
The plot works on longer regions of unusual sequence composition. A plot of the *Escherichia coli* genomic sequence reveals small light and dark boxes, corresponding to the four and five base sequence biases resulting from "very short patch" repair mechanisms [8].

13.8.1 Physical Characteristics

Although accurate calculation of physical properties is not realistic, because factors beyond the sequence are involved, it is possible to estimate the potential for bending in a DNA double helix. Various calculations have been proposed. One method is implemented in the EMBOSS program banana [10].

The program name, incidentally, is more than a simple joke on curvature. The con-

Example banana



sensus sequence for maximum bending is approximately AAAAA surrounded by five bases that are not "A". From the DNA ambiguity codes in Tab. 13.1, you can see that "N" is "any base" and "B" is "not-A". This means that the DNA sequence code "BANANA" matches the bent DNA consensus!

13.8.2 Detecting CpG Islands

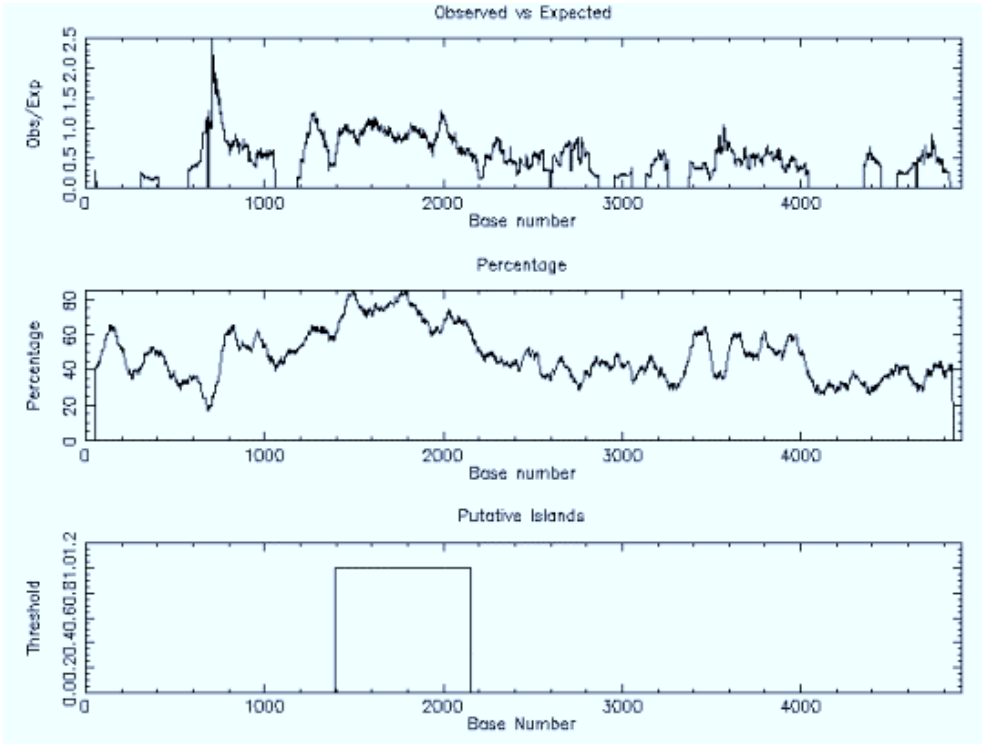
The chaos game plot clearly shows underrepresentation of the dinucleotide sequence "CG" in the human genome. However, in some parts of the genome, especially in the 5′ ends of "housekeeping" genes, the sequence "CG" occurs far more frequently than in the rest of the genome.

Rules have been derived for assigning such sequences as "CpG islands" (the "p" is simply the phosphate between the bases). The EMBOSS program cpgplot identifies CpG islands for the CPGISLE database project [11]. The three plots show the ratio of observed to expected "CG" sequences over a sliding window, the C + G content over the same window (even a sequence that is 100 % C and G will usually have rather fewer than the expected 25 "CG" dinucleotides). The third plot shows the region that meets the "CpG island" criteria.

A second program, newcpgreport, generates the CPGISLE database entries, in an EMBL-like format, from sequences identified by cpgplot. The comments section ("CC" lines) of the entry also shows the CpG island criteria used.

There are alternative approaches to CpG island identification. One used by another CpG island project is provided by the EM-

Example cpgplot



BOSS program newcpgseek (G. Micklem, personal communication), which calculates scores for potential CpG islands. The highest scoring match is the same, although slightly longer, as the region reported by cpgplot and newcpgreport.

Example newcpgreport

```
ID   HSHPRT8A  5000 BP.
XX
DE   CpG Island report.
XX
CC   Obs/Exp ratio > 0.60.
CC   % C + % G > 50.00.
CC   Length > 200.
XX
FH   Key             Location/Qualifiers
FT   CpG island      1395..2152
FT                   /size=758
FT                   /Sum C+G=547
FT                   /Percent CG=72.16
FT                   /ObsExp=0.85
FT   numislands      1
//
```

Example newcpgseek

```
NEWCPGSEEK of HSHPRT8A from 1 to 5000
with score > 17

 Begin    End   Score         CpG  %CG  CG/GC
   737    871      64          11  63.0  0.00
  1003   1009      48           3 100.0  1.00
  1249   2163     814          96  71.0  0.00
  2262   2287      29           3  50.0  3.00
  2657   2661      32           2 100.0  2.00
  3420   3435      39           3  75.0  3.00
  3575   3589      22           2  66.7  1.00
  3638   3649      25           2  66.7  2.00
  3826   3843      37           3  77.8  1.00
  3920   3930      26           2  72.7  2.00
-----------------------------------------------
```

13.8.3 Known Sequence Patterns

Nucleotide sequence patterns are used extensively when there are no similar sequences for comparison. Public databases are available for restriction enzyme target

sites (REBASE) and transcription factor binding sites (TRANSFAC).

13.8.4 Data Mining with Sequence Patterns

Many sequence patterns remain to be discovered. Patterns can be specified in many ways, some of which are particularly difficult to identify in a computer program. Special difficulties are caused by allowing a wide range of possible gaps or a large number of mismatches.

The EMBOSS program fuzznuc enables the user to search for "fuzzy" patterns in DNA sequences. The patterns are specified in a similar way to patterns in the protein motif database PROSITE. Fuzznuc first examines the pattern and then chooses the most appropriate string-matching method from its extensive library.

Example fuzznuc

```
########################################
# Program: fuzznuc
# Rundate: Fri Apr 23 15:00:15 2004
# Report_format: seqtable
# Report_file: fuzznuc.out
########################################

#=======================================
#
# Sequence: ECUW87     from: 1   to: 96484
# HitCount: 12
#
# Pattern: CTAG
# Mismatch: 0
# Complement: No
#
#=======================================

  Start      End Mismatch Sequence
    827      830        . ctag
   1811     1814        . ctag
  32570    32573        . ctag
  40394    40397        . ctag
  49454    49457        . ctag
  57582    57585        . ctag
  60838    60841        . ctag
  68074    68077        . ctag
  69341    69344        . ctag
  75724    75727        . ctag
  88686    88689        . ctag
  95978    95981        . ctag

#---------------------------------------
#---------------------------------------
```

The example run shows results from a simple search of *Escherichia coli* genomic DNA for the under-represented tetranucleotide "CTAG".

13.9 Restriction Sites and Promoter Consensus Sequences

13.9.1 Restriction Mapping

A natural application of pattern searching in DNA sequences is to use the target sites of restriction enzymes to generate a fragment size map from a sequence and to compare this with the expected experimental fragment sizes.

Because there are many and varied uses for such a program, these methods typically offer an unusually large number of options for the user, including restricting the search to rare cutters (those with long specific target sites such as NotI) or looking only for enzymes that will cut the sequence once or twice.

When a program must be especially versatile, it is particularly useful to have access to the program's source code so that new functions can be added and existing functions can be modified. Such changes can easily save several days of tedious analysis.

13.9.2 Codon Usage Analysis

Although more complicated methods for gene identification are covered elsewhere, it is worth reviewing some of the methods available in EMBOSS, especially for prokaryotic (i.e. unspliced) sequences.

Many methods depend on the strong bias in the use of alternative codons in true coding sequences [12]. This generally reflects

Example restrict

```
########################################
# Program: restrict
# Rundate: Thu Apr 29 16:41:08 2004
# Report_format: table
# Report_file: restrict.full
########################################

#=======================================
#
# Sequence: PAAMIR     from: 1   to: 2167
# HitCount: 44
#
# Minimum cuts per enzyme: 1
# Maximum cuts per enzyme: 2
# Minimum length of recognition site: 6
# Blunt ends allowed
# Sticky ends allowed
# DNA is linear
# No ambiguities allowed
#
#=======================================

  Start     End Enzyme_name Rest_site 5prime 3prime
      1       6 KpnI        GGTACC         5      1
      1       6 Acc65I      GGTACC         1      5
     57      62 BsgI        GTGCAG        37     35
     80      85 BssSI       CACGAG        80     84
     97     102 SunI        CGTACG        97    101
    152     159 NotI        GCGGCCGC     153    157
    198     203 EcoRV       GATATC       200    200
    205     210 Eco47III    AGCGCT       207    207
    214     219 MluI        ACGCGT       214    218
    231     236 BsrDI       GCAATG       225    223
    277     282 PvuI        CGATCG       280    278
    374     379 BssSI       CACGAG       369    373
    382     387 BspLU11I    ACATGT       382    386
    500     505 NarI        GGCGCC       501    503
    500     505 KasI        GGCGCC       500    504
    500     505 EgeI        GGCGCC       502    502
    500     505 BbeI        GGCGCC       504    500
    591     596 SacII       CCGCGG       594    592
    648     653 XhoI        CTCGAG       648    652
    671     676 BciVI       GTATCC       682    681
    809     816 NotI        GCGGCCGC     810    814
    817     822 PvuI        CGATCG       820    818
    886     891 BspMI       ACCTGC       872    876
    887     893 AarI        CACCTGC      872    876
    912     917 ClaI        ATCGAT       913    915
    943     948 StuI        AGGCCT       945    945
    950     955 NcoI        CCATGG       950    954
   1065    1070 BsgI        GTGCAG      1086   1084
   1167    1172 ClaI        ATCGAT      1168   1170
   1332    1337 BspMI       ACCTGC      1318   1322
   1333    1339 AarI        CACCTGC     1318   1322
   1379    1384 PvuII       CAGCTG      1381   1381
   1477    1482 MscI        TGGCCA      1479   1479
   1560    1565 SacII       CCGCGG      1563   1561
   1602    1607 BclI        TGATCA      1602   1606
   1606    1611 TaqII       CACCCA      1622   1620
   1693    1698 TaqII       GACCGA      1709   1707
   1718    1723 NaeI        GCCGGC      1720   1720
   1718    1723 NgoAIV      GCCGGC      1718   1722
   1765    1770 NcoI        CCATGG      1765   1769
   2016    2021 BclI        TGATCA      2016   2020
   2107    2112 Eco47III    AGCGCT      2109   2109
   2146    2151 TaqII       CACCCA      2131   2129
   2162    2167 XhoI        CTCGAG      2162   2166

#---------------------------------------
#---------------------------------------
```

the abundance of the respective tRNA species, so the codons with the most abundant tRNA will be used in preference, at least for genes that must be highly expressed under some conditions.

To apply such methods, we need to know the typical codon usage for a given species. Each has its own peculiar bias. In yeast *(Saccharomyces cerevisiae)* tables have been used successfully to distinguish low and highly expressed genes by their different codon usage patterns.

In the example below, which is sorted by amino acid, there is a clear bias in the codon choices for aspartate (D) where 85 % of the codons will be "GAC" and only 15 % will be the alternative "GAT". For leucine (L) there is a strong preference for "CTG" or "CTC" and in this sample of five genes only a single occurrence of "TTA"

13.9.3 Plotting Open Reading Frames

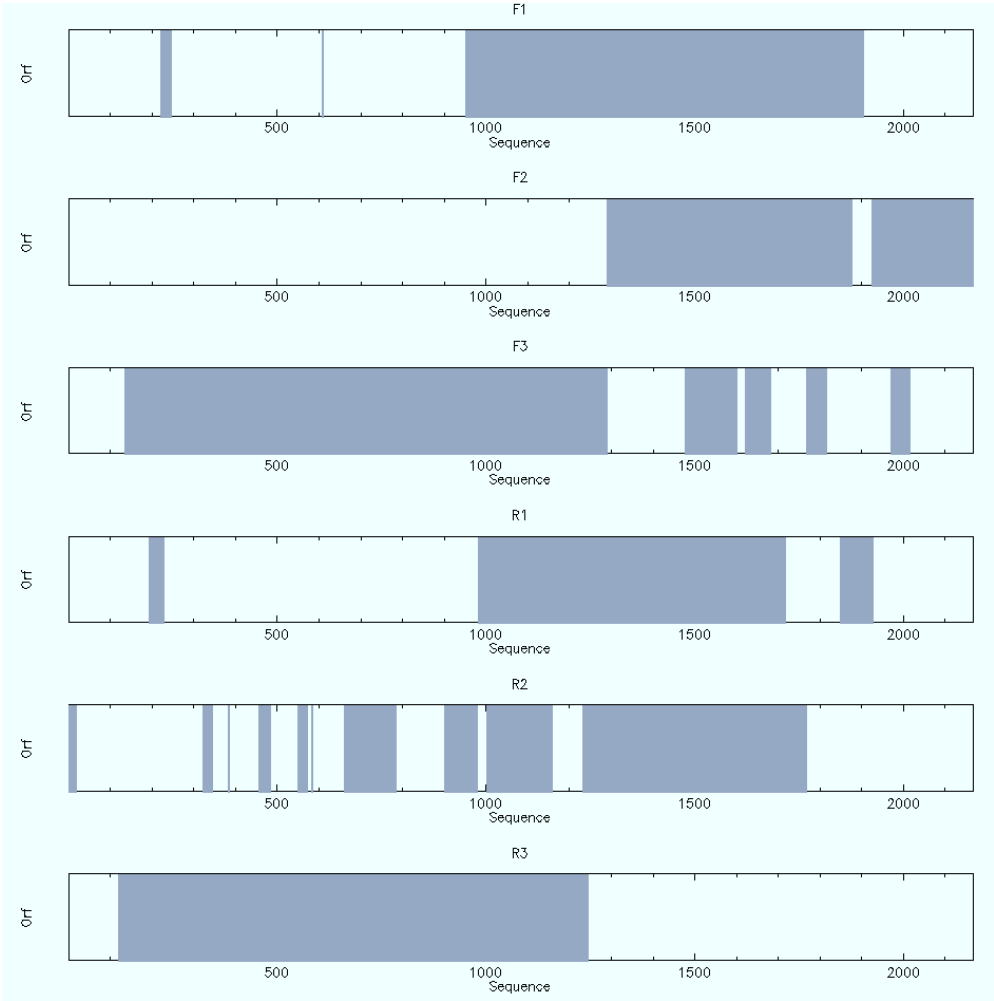
One of the simplest methods, plotorf, displays the open reading frames (ORF) in a sequence, defined as the longest sequences starting with a "start codon"(usually "ATG") and ending with a stop codon. The longest ORF are likely to be true genes, although care is needed in annotation. Although the stop codon is clearly fixed, there is no guarantee that the most distant start codon is that actually used when the gene is expressed.

Example cusp

```
# CUSP codon usage file
# Codon     Amino acid  Fract   /1000     Number
GCA   A           0.134 16.256      29
GCC   A           0.375 45.404      81
GCG   A           0.398 48.206      86
GCT   A           0.093 11.211      20
TGC   C           0.552 20.740      37
TGT   C           0.448 16.816      30
GAC   D           0.851 22.422      40
GAT   D           0.149 3.924 7
GAA   E           0.239 8.969 16
GAG   E           0.761 28.587      51
TTC   F           0.896 24.103      43
TTT   F           0.104 2.803 5
GGA   G           0.100 7.287 13
GGC   G           0.592 43.161      77
GGG   G           0.192 14.013      25
GGT   G           0.115 8.408 15
CAC   H           0.633 10.650      19
CAT   H           0.367 6.166 11
ATA   I           0.111 3.363 6
ATC   I           0.815 24.664      44
ATT   I           0.074 2.242 4
AAA   K           0.250 6.166 11
AAG   K           0.750 18.498      33
CTA   L           0.023 1.682 3
CTC   L           0.192 14.013      25
CTG   L           0.569 41.480      74
CTT   L           0.077 5.605 10
TTA   L           0.008 0.561 1
TTG   L           0.131 9.529 17
ATG   M           1.000 16.816      30

TAA   *           0.000 0.000 0
TAG   *           0.167 1.682 3
TGA   *           0.833 8.408 15
```

Example plotorf



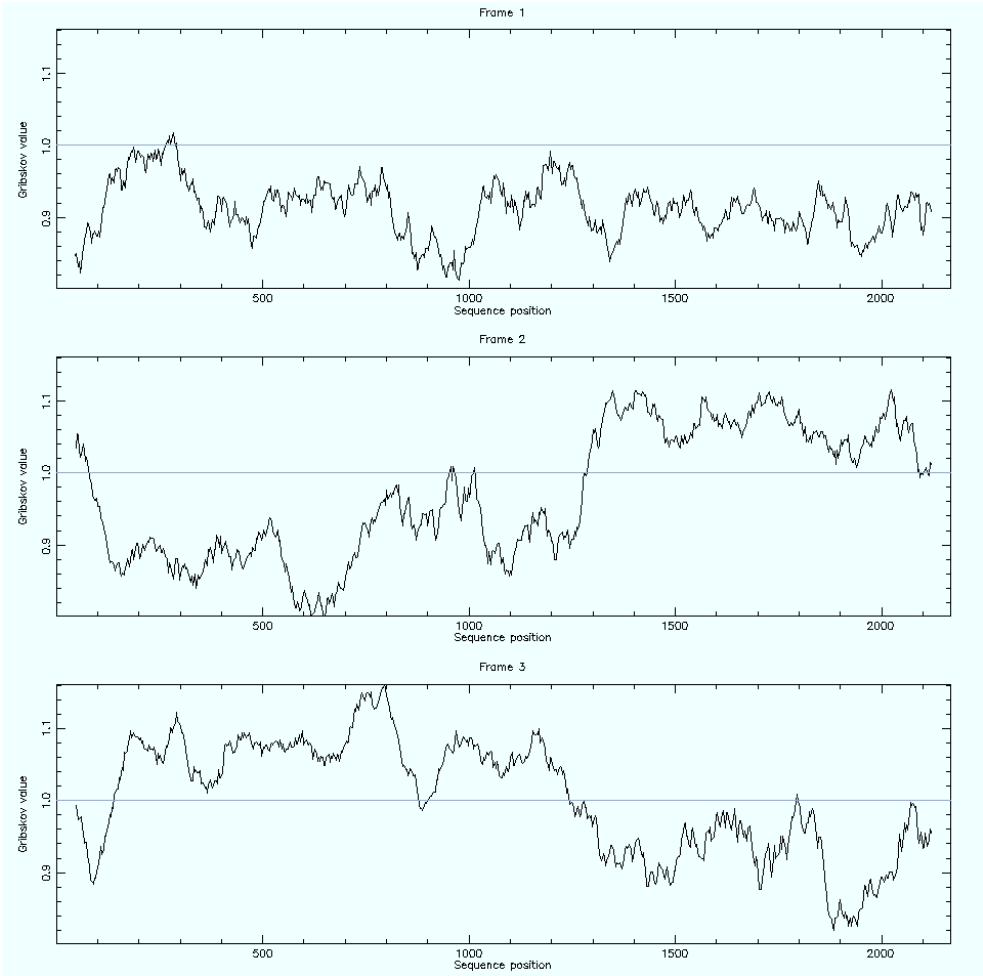
A further complication arises because the nature of amino acid composition and codon bias tends to reduce the number of possible stop codons on the opposite strand, and so it is common to see two open reading frames in opposite directions overlapping each other.

The plot below shows an example for a sequence from *Pseudomonas aeruginosa* [13], in which the high G + C content in this organism further reduces the frequency of stop codons which are, of course, relatively AT-rich. The three true open reading frames can be shown by other methods to be the longest in the second and third panels of the plot.

13.9.4 Codon Preference Statistics

To identify such genes, statistical methods are available which use either the expected

Example syco

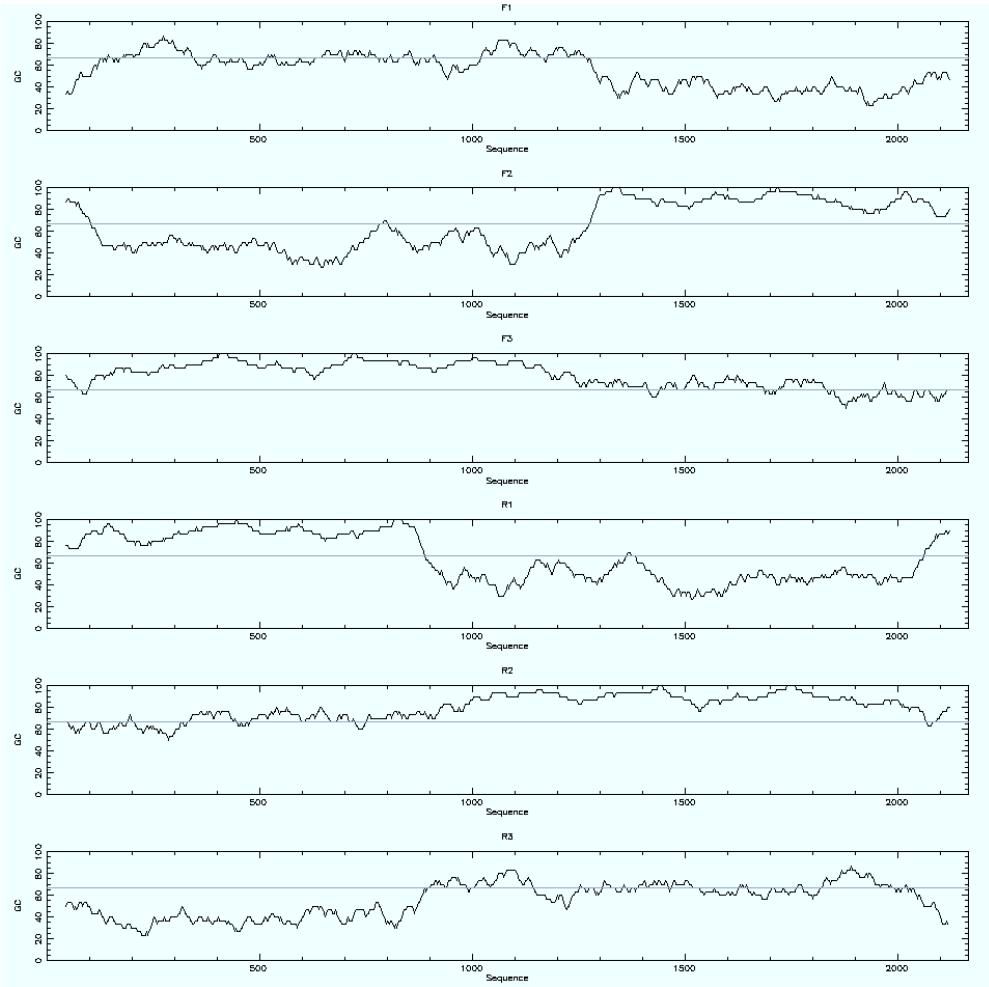


codon usage table or, more crudely, the biased base composition of the third (wobble) base position in each codon.

For codon usage, the EMBOSS program syco (synonymous codons) plots the Gribskov statistic [12] over a sequence in each of the three possible reading frames. The example sequence is the same as that in the plotorf example above.

If a codon usage table is not available, or for some species in which the codon bias is weak, statistical analysis of the third base position offers a practical alternative method. The example below is for the same *Pseudomonas aeruginosa* sequence, in which the high G + C content leads to a specially strong third position bias reflected in the general rule "if any base can be C or G then it will be."

Example wobble



13.9.5

Reading Frame Statistics

When a potential gene has been identified, a variety of statistical methods are available to calculate the codon bias for that particular gene and to relate these to the possible expression level of the gene on the grounds that more highly expressed genes are generally found to have a more biased codon usage.

An example in EMBOSS is Frank Wright's "effective number of codons" statistic [14], calculated by the program chips ("codon heuristics in protein-coding sequences")

In addition to using codon usage table to analyze gene expression in one organism, these tables can be used to compare species. Similar codon frequencies will usually be observed for closely related species, although there are exceptions caused by re-

Example chips

```
# CHIPS codon usage statistics

Nc = 32.951
```

cent changes in GC content, and by recent horizontal transfer of DNA. These methods can be used, for example, to identify DNA or viral origin in bacterial genomes.

The program codcmp compares two codon usage tables, and calculates statistics to indicate their degree of difference.

Example codcmp

```
# CODCMP codon usage table comparison
# Eeco.cut vs Ehum.cut

Sum Squared Difference = 1.409
Mean Squared Difference = 0.022
Root Mean Squared Difference = 0.148
Sum Difference          = 7.350
Mean Difference          = 0.115
Codons not appearing    = 0
```

13.10 The Future for EMBOSS

The EMBOSS project is an open source effort, which means that all the source code is made freely available for everyone to share and develop further. The home of the project is now shared by the European Bioinformatics Institute (http://www.ebi.ac.uk/), and the Human Genome Mapping Project (HGMP) Resource Center http://www.hgmp.mrc.ac.uk/ in Hinxton, next door to the Sanger Center where the project started.

The programs were designed to be run as "commands" by simply typing the program name and answering a series of questions. It was always clear that for many users this would not be enough, and also that few users would, in practice, agree on the best approach to take.

The EMBOSS collaborators made plans to make all the programs available under as many different user interfaces as possible. Already in the first phase of the project (up to the end of the year 2000) we had volunteers who added EMBOSS to Web interfaces such as Pise from the Institut Pasteur, and W2H from the German Cancer Research Center (DKFZ – the German EMBnet node http://www.de.embnet.org/). Web interfaces make the EMBOSS programs available to any user with a web browser, through access to a central service site that can install the databases and maintain the programs – something that EMBnet sites have many years of experience in providing to most of the major countries of the world.

Meanwhile, the EMBOSS core developers produced a graphical user interface called Jemboss, and also collaborated with commercial software companies to provide EMBOSS under proprietary interfaces, for example SRS.

An exciting new area in bioinformatics is the emergence of "web services" using the SOAP protocol. EMBOSS has been adopted by several such projects, including myGrid [15] SoapLab [16] and Taverna [17] which builds EMBOSS-based web services into workflows.

The programs will also continue to develop. In the area of sequence analysis we are adding more input and output data types, and rewriting the program output to give a set of user-selectable report formats; these will make the processing of EMBOSS output and the interconnection of the programs much easier in the next release.

EMBOSS is also extending beyond pure sequence analysis, with groups interested in the fields of protein structure and phylogenetics (to give just two examples). With help from the bioinformatics community the future continues to look very interesting indeed.

References

1 Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite http://www.emboss.org/ and http://emboss.sourceforge.net/ Trends Genet. 16:276–277.

2 Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

3 Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. J. Mol. Biol. 147:195–197.

4 Mott, R.F. (1997) Est_genome: a program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. 13: 477–478.

5 Wilbur, W.J. and Lipman, D.J. (1983) Improved tools for biological sequence comparison. Proc. Nat. Acad. Sci. USA 80:726–730.

6 Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

7 Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) Biological Sequence Analysis, Cambridge University Press, UK.

8 Merkl, R., Kroeger, M., Rice, P., and Fritz, H.-J. (1992) Statistical evaluation and biological interpretation of non-random abundance in the *E. coli* K-12 genome of tetra- and pentanucleotide sequences related to VSP DNA mismatch repair. Nucleic Acids Res. 20:1657–1662.

9 Jeffrey, H.J. (1990) Chaos game representation of gene structure. Nucleic Acids Res. 18:2163–2170.

10 Goodsell, D.S. and Dickerson, R.E. (1994) Bending and Curvature Calculations in B-DNA. Nucleic Acids Res. 22:5497–5503.

11 Larsen, F., Gundersen, G., Lopez, R., Prydz, H. (1992) CpG islands as gene markers in the human genome. Genomics 13:105–107.

12 Gribskov, M., Devereux, J., and Burgess, R.R. (1984) The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res. 12:539–549.

13 Lowe, N., Rice, P.M., and Drew, R.E.(1989) Nucleotide sequence of the aliphatic amidase regulator gene of *Pseudomonas aeruginosa*. FEBS Lett. 246:39–43.

14 Wright, F. (1990) The 'effective number of codons' used in a gene. Gene 87:23–29.

15 Stevens, R., Robinson, A., and Goble, C.A. (2003) myGrid: Personalised Bioinformatics on the Information Grid http://www.mygrid.org.uk/ Bioinformatics 19 (Suppl. 1):i302–i304.

16 Senger, M., Rice, P., and Oinn, T. (2003) Soaplab – a unified Sesame door to analysis tools. In: Cox, S.J. (Ed.) Proc. UK e-Science, All Hands Meeting 2003, pp. 509–513, ISBN 1-904425-11-9.

17 Oinn, T., Addis, M., Ferris, J., Marvin, D., Greenwood, M., Carver, T., Wipat, A., and Li, P. (2004) Taverna: A tool for the composition and enactment of bioinformatics workflows. Bioinformatics J., in press.

14
Software Tools for Proteomics Technologies

David S. Wishart

14.1
Introduction

At the time of this writing there are more than 250 fully sequenced organisms, including more than 200 different species of bacteria and more than two dozen different eukaryotes such as yeast, rice, nematodes, fruit flies, pufferfish, rodents (mice and rats), and humans [1]. It is likely that within two more years complete sets of gene and protein sequences will be known for another 200–300 organisms including most key agricultural animals and plants and all remaining laboratory animals [2, 3]. It is clear that the challenge over the coming decades will be to connect all those protein sequences with their respective actions and to translate that understanding into new approaches for management or treatment of disease, diagnosis of medical conditions, monitoring of drug interactions, improvement of crop yields, or enhancement of the quality of our environment.

Key to translating this raw biological data to practical knowledge will be our ability to recognize or detect patterns occurring within this data [4]. This is where bioinformatics is important. Bioinformatics plays a vital role in all areas of proteomics (the study of proteins and their interactions) by providing the software tools that help sort, store, analyze, visualize, and extract important patterns from raw proteomic data. Computational tools such as correlational analysis, multiparametric fitting, dynamic programming, artificial intelligence, neural networks and hidden Markov models are critical for revealing many of the hidden patterns and relationships in sequence, 2D gel, or mass spectrometric (MS) data. Complementing these tools is a growing array of queryable databases containing protein sequences, pre-calculated mass fragment data, 2D gel images, 3D structures, protein interaction partners, biochemical pathways and functional sites that provide the critical "prior knowledge" necessary for extracting additional information from unprocessed experimental data.

In previous chapters on protein and DNA technology we have seen how raw protein sequence data, isoelectric points, and peptide mass fingerprints can be acquired. In this chapter we will see how these raw data can be transformed into useful biochemical knowledge. In particular, we will show how bioinformatics tools can be used to facilitate protein identification and characterization

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

using 2D gel, MS, and unprocessed protein sequence data. This chapter will be divided into two sections. The first is concerned with describing the software tools and algorithms that can facilitate protein identification from 2D gels, mass spectrometric, or protein sequence data. The second will describe the bioinformatics tools and databases that can be used to predict the functions, locations, and properties of the proteins identified. Particular emphasis will be placed on describing freely available Web tools or software packages that have been published in the scientific literature.

14.2 Protein Identification

Unfortunately for us, proteins do not come with name tags. They also occur in extremely complex matrixes, usually with other proteins that look and behave almost identically. Indeed, the only way to uniquely identify a protein is to carefully separate it and painstakingly determine its sequence or precisely measure its mass. Consequently protein identification is an inherently difficult process that requires close interplay between experimental and computational techniques [5, 6]. The experimental techniques provide the raw data and the computational techniques convert those raw data into a usable protein name or a data bank accession number. These computational tools all rely on a common theme – they identify proteins by looking for close matches (in mass, in sequence, or in 2D gel position) with previously identified proteins. In this way protein identification is facilitated by making use of vast stores of previously accumulated knowledge about the two million proteins already studied. In this section we will review three methods of protein identification and their associated software tools:

1. identification by 2D gel spot position;
2. identification by mass spectrometry; and
3. identification by sequence data.

14.2.1 Protein Identification from 2D Gels

Despite predictions of its imminent demise, 1D and 2D gel electrophoresis is, and will probably continue to be, an essential cornerstone of proteomics research. No other low-cost technology has comparable resolution or sensitivity for protein separation and display. Indeed, as we have seen from Chapt. 7, 2D gel electrophoresis enables precise and reproducible separation of up to 10,000 different proteins. With the introduction of IEF "ultra-zoom" gels [7] to improve resolution, the commercialization of pre-poured immobilized pH gradient (IPG) to improve consistency, and the development of multicolored, multiplexed DIGE [8] to enable rapid and robust in-gel comparisons, it is probable that 2D gels will be with us for a very long time. Continuing technical advances and the rapidly growing use of 2D gels in proteomics has led to the development of several excellent software tools and an increasing number of valuable 2D gel databases to facilitate protein identification, quantification, and annotation. Although most 2D gel analysis software is very image-oriented, the fact that these packages can be used to measure physical properties (p*I* and MW), to quantify protein copy numbers, and to identify proteins makes them a key part of the standard bioinformatics tool chest.

Several popular, commercial software packages are available for facilitating digital gel analysis, including Melanie 4 (GeneBio), Phoretix 2D (Phoretix), ImageMaster 2D (Amersham), PDQuest (BioRad), Delta2D (DeCodon), and Gellab II+ (Scanalytics) (Tab. 14.1). All six packages have an im-

Table 14.1 Protein identification tools – web links.

Tool/Database	*Web Address*
GelScape	http://www.gelscape.org
Flicker (2D gels)	http://www.lecb.ncifcrf.gov/flicker/
Open Source Flicker	http://open2dprot.sourceforge.net/Flicker/
Phoretix 2D	http://www.nonlinear.com/products/2d/2d.asp
PDQuest	http://www.proteomeworks.bio-rad.com/html/pdquest.html
Melanie IV	http://ca.expasy.org/melanie/
Gellab II+	http://www.scanalytics.com/product/gellab/index.shtml
ImageMaster 2D	http://www1.amershambiosciences.com/
Delta2D	http://www.decodon.com/Solutions/Delta2D.html
2DWG (2D databases)	http://www.lecb.ncifcrf.gov/2dwgDB/
SWISS-2DPAGE	http://www.expasy.ch/
2D-HUNT	http://ca.expasy.org/ch2d/2DHunt/
World 2D PAGE	http://ca.expasy.org/ch2d/2d-index.html
Yeast 2D database	http://www.ibgc.u-bordeaux2.fr/YPM/
PeptIdent (MS Fingerprint)	http://us.expasy.org/tools/peptident.html
Profound (MS Fingerprint)	http://prowl.rockefeller.edu/profound_bin/WebProFound.exe
Mowse (MS Fingerprint)	http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse
Mascot (All searches)	http://www.matrixscience.com/search_form_select.html
ProteinProspector (All)	http://prospector.ucsf.edu/
PepSea (MS Fingerprint)	http://pepsea.protana.com/PA_PepSeaForm.html
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
PSI-BLAST	http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi
Swiss-Prot Database	http://ca.expasy.org/sprot/
Owl Database	http://bioinf.man.ac.uk/dbbrowser/OWL/
PIR Database	http://pir.georgetown.edu/home.shtml
GenBank Database	http://www.ncbi.nlm.nih.gov/
UniProt Database	http://www.pir.uniprot.org/
Protein Data Bank	http://www.rcsb.org

pressive array of image-manipulation facilities integrated into sophisticated graphical user interfaces. Most are specific to Windows platforms although PDQuest runs on both Windows and MacOS. Some commercial packages, for example Melanie (Medical ELectrophoresis Analysis Interactive Expert system) and Gellab II+ began as academic projects and have been under development for many years [9, 10]. Most of these packages make use of machine learning, heuristic clustering, artificial intelligence, and high-level image manipulation techniques to support some very complex 2D gel analyses. Several commercial packages are typically sold as part of larger equipment purchases (2D gel systems with robotic gel cutters) and are closely tied to the major proteomics equipment suppliers or 2D gel vendors.

Essentially, all commercial packages offer an array of automated or manual spot manipulation, including spot detection, spot editing, spot normalization, spot filtering, spot (re)coloring, spot quantitation, and spot annotation. This allows users to compare, quantify and archive 2D gel spots quickly and accurately. These kinds of comparison are particularly important for monitoring changes in protein expression from

gel to gel or from experiment to experiment. In addition to individual spot manipulation, whole gel manipulations such as rotating, overlaying, referencing, "synthesizing" and averaging are typically supported in most commercial packages. This is done to facilitate inter-gel comparison and to calibrate gels to p*I* and molecular weight standards. Calibration is particularly important for 2D gels if one wishes to extract accurate molecular weight or p*I* information for protein identification.

No matter how careful one is in casting or running a 2D gel, there is usually some inter-gel variability. The ability to stretch or shrink certain gel regions (or even entire gels) is, therefore, often necessary to enable direct comparisons. Techniques called spot matching, "landmarking", and image "warping" are available in most programs to enable this kind of forced matching. When this kind of image transformation has been completed most commercial packages enable additional gels to be overlaid, subtracted, alternately flashed (flickered), or color-contrasted to identify significant changes or significant spots.

In addition to the commercial 2D gel packages, there is a small but growing collection of high quality freeware packages for 1D and 2D gel analysis including GelScape [11] and Flicker [12, 13]. Both packages are written in Java and both are freely available either as stand-alone applications or as platform-independent servlets/applets that can be readily accessed over the web (Tab. 14.1). The use of web-server technology enables users to access, share, or distribute gel images and gel data in an interactive, platform-independent manner. GelScape (Fig. 14.1) supports many of the features found in commercial, stand-alone gel-analysis packages including interactive spot marking and annotation, spot integration, gel warping, image resizing, HTML image mapping, gel (pH/MW) gridding, automated spot picking and integration, transparent image overlaying, rubber-band zooming, image recoloring, image reformatting, and gel image and gel annotation data storage in compliance with Federated Gel Database [14] requirements. Although not quite as feature-rich as GelScape, the Flicker applet is, nevertheless, quite useful for transforming (warping, rotating, etc.) and visualizing pairs of 2D gels so that the gel of interest can be easily compared with a pre-existing gel obtained from the Web. Flicker's name comes from the fact that the program enables gel images to be switched on and off ("flickered") to facilitate visual comparison. Flicker was recently rewritten as a stand-alone, open source java application (Tab. 14.1). This has led to significant performance and feature improvements over the original applet version.

Protein identification by use of 2D gels can be achieved in any number of ways, whether by p*I*/MW measurements, Western blotting (if an antibody is known), or ^{32}P detection (if the protein of interest is known to be phosphorylated). Often, however, the best method for protein identification by visual comparison with previously annotated gels from databases [12, 15–17]. Over the past 25 years, thousands of 2D gels have been run on cell extracts of many different organisms and human tissues. Many of these gels have been analyzed and their protein spots identified by microsequencing or mass spectrometry. These carefully annotated gels have been deposited in more than 30 different "federated" 2D gel databases (for example SWISS-2D PAGE or WebGel) with the intention that others who might be studying similar systems can use these standardized, annotated gel images to overlay with their own gels and rapidly identify proteins of interest.

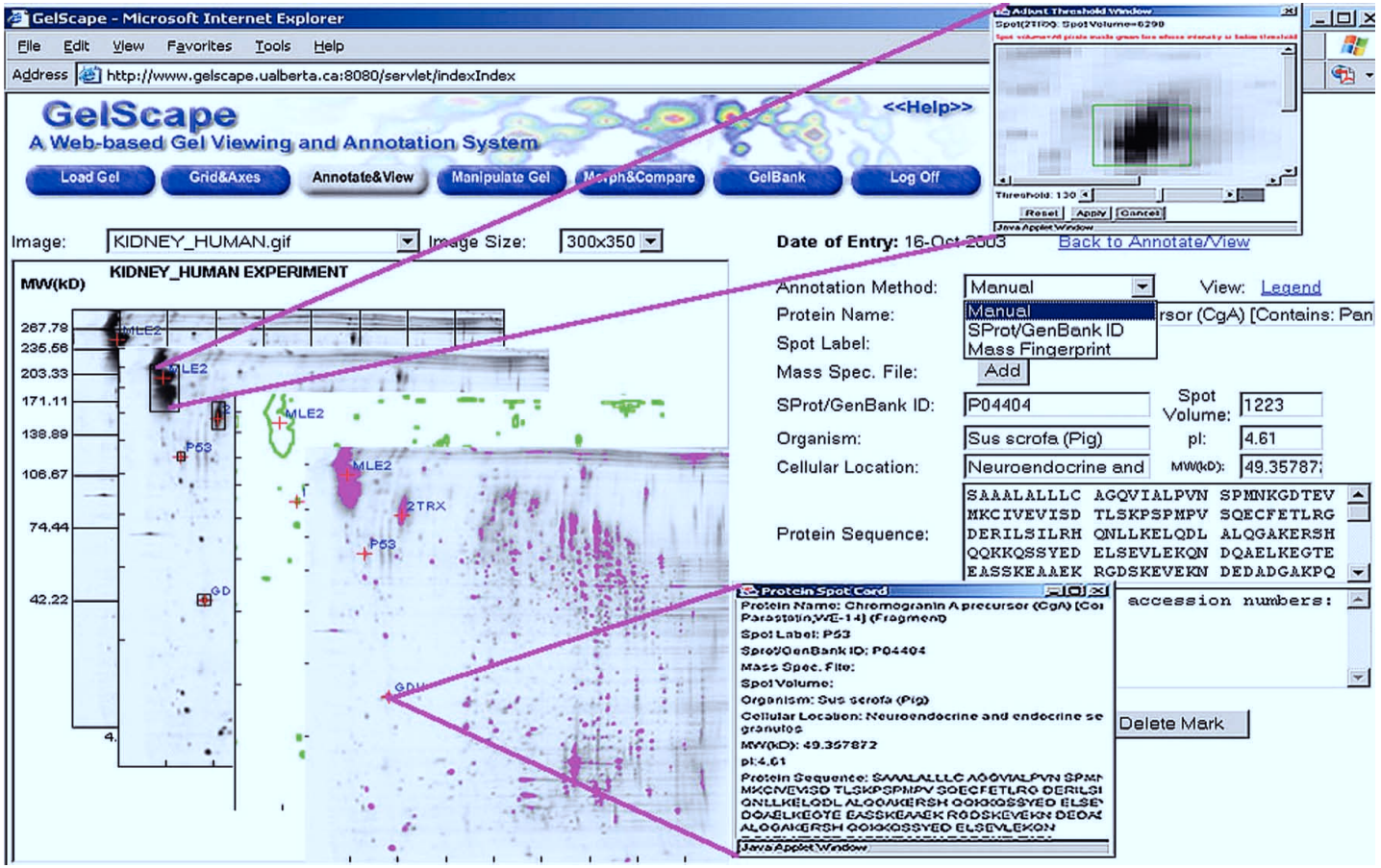


Fig. 14.1 A screen-shot montage illustrating several of the gel viewing, gel annotation, and gel manipulation features found in GelScape.

Many of these gels and gel databases can be found on the internet via WORLD-2-DPAGE or 2D-HUNT (Tab. 14.1) simply by typing an organism or protein name. For instance, suppose you decided to study *S. cerevisiae* under anaerobic conditions. By running a 2D gel of the proteins expressed under anaerobic conditions, you can save literally months of effort by comparing the gel with the fully annotated *S. cerevisiae* 2D gel (grown under aerobic conditions) found at http://www.ibgc.u-bordeaux2.fr/YPM/. By using a freeware package like GelScape, or more sophisticated commercial packages, it should be possible to visually transform the two gels, overlay them and identify nearly 400 yeast proteins or protein fragments in less than an hour. Quantification of the differences in expression might take only a few more hours. Indeed, the objective of setting up these federated 2D gel databases is to avoid costly or repetitive efforts that only lead to re-identification of previously mapped or previously known proteins. The utility of 2D gel databases is bound to grow as more gels are collected and as more spots are progressively identified in laboratories around the world. Indeed, one might optimistically predict that some time in the not-too-distant future, mass spectrometry and micro-sequencing will no longer be needed to routinely identify protein gel spots, because all detectable spots will have been annotated and archived in a set of Web-accessible 2D gel databases.

Much remains to be done before this vision becomes reality, however. 2D gel spot patterns are highly dependent on the methods used to isolate and prepare the initial protein mixture [15]. Consequently, individuals wishing to perform gel database comparisons must take into account such variables as the protein fraction isolated, how the sample was prepared, and how the gel was run. Even if sample preparation issues are eventually clarified, continuing problems concerning 2D gel database maintenance and updates will persist. Indeed, most publicly available annotated 2D gels represent incomplete "best efforts" of a single graduate student rather than collective, sustained efforts arising from multiple labs. If the concept of 2D gel databases and data sharing is going to succeed, it will need a well-funded central repository (for example the NCBI or EBI) and open-minded funding agencies to support sustained gel annotation contributions from the whole scientific community. The recent introduction of XML- (extensible markup language) based procedures to help standardize 2D gel data sharing is an encouraging step in this direction [18].

In the future it is probable that other separation and display techniques, for example 2D HPLC, tandem capillary electrophoresis (Chapt. 8), and protein chips [5] will gain greater prominence in functional proteomics. The resolution and separation reproducibility of these techniques suggest that similar database comparison methods (e.g. elution profile analysis) could eventually enable proteins to be identified without the need for MS or microsequencing analysis.

14.2.2 Protein Identification from Mass Spectrometry

Recent advances in mass spectrometry have led to a paradigm shift in the way peptides and proteins are identified [19, 20]. In particular, the introduction of "soft" ionization techniques (electrospray and MALDI), coupled with substantial improvements in mass accuracy (~2 ppm), resolution (MS–MS) and sensitivity (femtomoles) have made the rapid, high throughput identification of peptides and proteins almost routine [5, 21].

Key to making this shift possible has been the development of bioinformatics software that enables direct correlation of biomolecular MS data with protein sequence databases. Two kinds of MS bioinformatics software exist:

1. software for identifying proteins from peptide mass fingerprints; and
2. software for identifying peptides or proteins directly from uninterpreted tandem (MS–MS) mass spectra.

Peptide mass fingerprinting (PMF) was developed in the early 1990s as a means of unambiguously identifying proteins from proteolytic fragments [22–24]. Specifically, if a pure protein is digested with a protease (for example trypsin) that cuts at predictable locations the result will be a peptide mixture containing a unique collection of between 10 and 50 different peptides, each with a different or characteristic mass. Running this mixture on a modern ESI or MALDI instrument will lead to a mass spectrum with dozens of peaks corresponding to the masses of each of these peptides. Because no two proteins are likely to share the same set of constituent peptides, this mixture is called a peptide mass fingerprint or PMF. By comparing the observed masses of the mixture with predicted peptide masses derived from all known protein sequences it is theoretically possible to identify the protein of interest (providing the protein has been previously sequenced). Specifically, in the course of performing a mass fingerprint search, database sequences are theoretically "cleaved" using known protease-cutting rules, the resulting hypothetical peptide masses are calculated and the whole protein is ranked according to the number of exact (or near exact) cleavage fragment matches made to the observed set of peptide masses. The sequence with the highest number and quality of matches is usually selected as the most likely candidate (an outline is given in Fig. 14.2).

There are nearly a dozen different types of mass fingerprinting software available for protein identification. Although some are sold as commercial products, most are freely available over the Web (a partial listing is given in Tab. 14.1). Nearly all of the packages enable the user to select a protein database (OWL, Swiss-Prot or NCBI-nr), a source organism (to limit the search), a cleavage enzyme (trypsin is the most common), a cleavage tolerance (one missed cleavage per peptide is usual), a mass tolerance (0.1 amu is typical), and a mass type (average or monoisotopic). Most of these values are pre-selected as defaults in the submission form and do not normally need to be changed. All packages expect users to enter a list of masses (with at least two decimal point accuracy) read from the MS spectrum before launching the search. On most days a Web search result can be returned within 10–20 s.

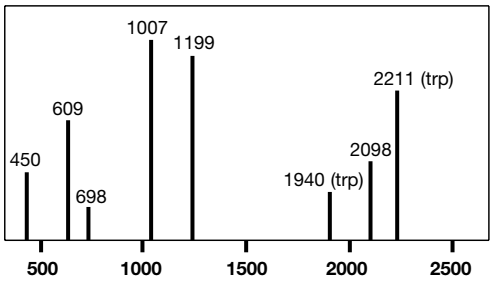
Key to performing any successful peptide mass fingerprint search is to start with the most accurate masses possible. Internally calibrated mono-isotopic standards are essential. If one is very confident in the mass accuracy, restricting the mass tolerance setting to less than 0.1 amu will usually improve the specificity of the search. Restricting the size or scope of the database to search is also wise. The organism being studied is usually known, so it is best to select only the portion of the protein database with protein sequences from the presumptive source organism (or very closely related organisms). It is not (yet) a good idea to search through translated EST databases, because they have too many sequencing errors and contain only partial protein sequence information.

As a general rule one should try to use as many mass values as possible when per-

A

Sequence DB	Calc. Tryptic Frags	Mass List
>P12345	acedfhsak	450.2017 (P21234)
acedfhsakdfqea	dfgeasdfpk	609.2667 (P12345)
sdfpkivtmeeewe	ivtmeeewendadnfek	664.3300 (P89212)
ndadnfekqwfe	gwfe	1007.4251 (P12345)
		1114.4416 (P89212)
>P21234	acek	1183.5266 (P12345)
acekdfhsadfqea	dfhsadfgeasdfpk	1300.5116 (P21234)
sdfpkivtmeeewe	ivtmeeewenk	1407.6462 (P21234)
nkdadnfeqwfe	dadnfeqwfe	1526.6211 (P89212)
		1593.7101 (P89212)
>P89212	acedfhsadfgek	1740.7501 (P21234)
acedfhsadfqeka	asdfpk	2098.8909 (P12345)
sdfpkivtmeeewe	ivtmeeewendak	
ndakdnfeqwfe	dnfegwfe	

B



C

Query Masses	Database Mass List	Results
450.2201 →	450.2017 (P21234)	2 Unknown masses
609.3667 →	609.2667 (P12345)	1 hit on P21234
698.3100	664.3300 (P89212)	3 hits on P12345
1007.5391 →	1007.4251 (P12345)	
1199.4916	1114.4416 (P89212)	Conclude the query
2098.9909 ↘	1183.5266 (P12345)	protein is P12345
	1300.5116 (P21234)	
	1407.6462 (P21234)	
	1526.6211 (P89212)	
	1593.7101 (P89212)	
	1740.7501 (P21234)	
	2098.8909 (P12345)	

Fig. 14.2. An outline of the peptide mass fingerprinting procedure for identifying proteins. In (A) the preparation of the "in silico" trypsin cleavage database from the original sequence database is shown. In (B) the query MALDI mass fingerprint from the unknown protein is shown. In (C) the general process of the PMF search is shown with a sample output.

forming an MS fingerprint analysis. An absolute minimum of five, but more commonly ten or more mass values should be entered for positive identification of a protein. Typically, the number of masses entered should depend on the molecular weight of the protein in kilodaltons (i.e. more mass entries for heavier proteins).

The need for so much mass data is primarily to compensate for the noise inherent in experimental MS data. Indeed, it is not uncommon to have up to half of all predicted peptide peaks absent from any given MS fingerprint spectrum along with any number of additional peaks arising from contaminating proteins. Consequently, one is usually quite content to obtain coverage (the fraction of predicted peptide masses closely matching observed peptide masses) of only 40–50 %. It is very rare to see a perfect match or 100 % coverage.

Because of the experimental noise associated with MS data, analysis of peptide fingerprint searches is not always easy nor is it always reliable. Some of the common complications include:

1. disappearance of key peaks as a result of non-specific ion suppression;
2. appearance of extra peaks from protease (trypsin) autolysis;
3. appearance of peaks from post-translational or artifactual chemical modification;
4. appearance of peaks from non-specific cleavage, or from contaminating proteases; and
5. appearance of peaks from contaminating impurities, contaminating homologs or splice variants.

Because of these complications, the issue of how to score and rank peptide mass matches is actually quite critical to the performance and reliability of peptide mass fingerprint software. Most early fingerprinting programs used simple heuristic scoring schemes and arbitrary cut-offs to select candidate sequences. More recent programs such as Profound [25] use Bayesian statistics and posterior probabilities to rank database candidates. Some of the latest programs take at least some of the complications listed above into account and allow secondary searches with so-called "orphan" masses. Mascot [26] is one program which uses a probabilistic model similar to an expectation or E-value to rank sequences. The use of probabilities enables better estimation of significance (which guards against false positives) and it also enables scores to be compared with those from other types of search algorithm (such as BLAST). Irrespective of the advantages and disadvantages of individual programs, it is usually a good idea to run at least two different peptide mass fingerprinting programs and combine the results. This serves as a form of "signal averaging" and potentially reduces the occurrence of errors arising from algorithmic or database limitations in any single program.

Because peptide mass fingerprinting does not always work, for unambiguous protein identification there has been increasing emphasis on using tandem mass spectrometers equipped with collision-induced dissociation (CID) cells to provide more precise and interpretable peptide data. SEQUEST [27, 28], ProteinProspector [29], and Mascot [26] are three software packages that can be used to analyze tandem mass data of peptide fragments. These programs take uninterpreted tandem mass spectral data (i.e. the actual spectrum or peak lists), perform sequence database searches and identify probable peptides or protein matches. Typically these programs work by first scanning the protein databases for potential matches with the precursor peptide ion then ranking the candidate peptides on the basis of their predicted similarity (ion continuity, intensity, etc.) to the observed fragment ion masses. After this screening step a model MS–MS spectrum for each candidate peptide is generated and then compared, scored, and ranked with the observed MS–MS spectrum using correlational or probabilistic analysis. As with peptide mass

fingerprinting, similar kinds of information (database, source organism, mass tolerance, cleavage specificity, etc.) must be provided before running the programs. The only difference is that instead of typing in a list of masses the user is expected to provide a spectral filename containing the digitized MS–MS spectrum. The reported performance of these programs is quite impressive [26, 28].

It is likely that protein identification via mass spectral analysis will continue to grow in popularity and in importance. The wide availability of easy-to-use, freely available peptide mass fingerprinting software has made the entire protein identification process very accessible. Furthermore, as more protein sequence data is deposited in sequence data banks around the world the utility and reliability of these database-driven techniques is expected to grow accordingly. Sequence databases continue to grow and mass spectrometer technology is also progressing rapidly. With continuing improvements in mass resolution (e.g. Fourier-transform cyclotron mass spectrometers with 1 ppm resolution are now available) it is likely that peptide mass fingerprinting will become less common as only a single tryptic peptide will be sufficient for positive identification of a protein [30, 31].

14.2.3 Protein Identification from Sequence Data

The most precise and accurate way of unambiguously identifying a protein is through its sequence. Historically, proteins were identified by direct sequencing using painstakingly difficult chemical or enzymatic methods (Edman degradation, proteolytic digests). All that changed with the development of DNA-sequencing techniques which proved to be faster, cheaper, and more robust [32]. Now more than 99 % of all protein sequences deposited in databases such as OWL [33], PIR [34], Swiss-Prot + trembl [35], UniProt [36], and GenBank [37] are derived directly from DNA sequence data. While complete sequence data is normally obtained via DNA sequencing, improvements in mass spectrometry and chemical microsequencing now enable routine sequencing of short (10–20 residue) peptides from subpicomole quantities of protein [38, 39]. With the availability of several different rapid sequencing methods (MS–MS, chemical microsequencers, DNA sequencers, ladder sequencing, etc.) and the growing number of protein sequences (>2,000,000) and sequence databases (>200), there is now increasing pressure to develop and use specific bioinformatics tools to facilitate protein identification from partial or homologous sequence data.

Protein identification via sequence analysis can be performed through either exact substring matches or local sequence similarity to a member of a database of known protein sequences. Exact matching of short peptide sequences to known protein sequences is ideal for identifying proteins from partial sequence data (obtained via Edman microsequencing or tandem MS). This type of text matching to sequence data is currently supported by the OWL, Swiss-Prot, and PIR web servers (but not GenBank!). Given the current size of the databases and the number of residues they contain it is usually wise to sequence 7 or 8 residues to prevent the occurrence of false positives. Alternately, if information about the protein mass, predicted p*I* or source organism is available, only four or five residues might need to be determined to guarantee a unique match. Note that exact string matching will only identify a protein if it is already contained in a sequence database.

Although exact string (or subsequence) matching is useful for certain types of pro-

tein identification problem, by far the most common method of protein identification is by "fuzzy matching" via sequence similarity. Unlike exact string matching, sequence similarity determination is a robust technique which enables identification of proteins even if there are sequencing errors in either the query or database sequence. Furthermore, sequence similarity enables one to potentially identify or ascribe a function to a protein even if it is not contained in the database. In particular, the identification of a similar sequence (>25 % sequence identity to the query sequence) with a known function or name is usually sufficient to infer the function or name of an unknown protein [40].

Sequence similarity is normally determined by using database alignment algorithms wherein a query sequence is aligned and compared with all other sequences in a database. In many respects sequence-alignment programs are merely glorified spell checkers. Fundamentally there are two types of sequence alignment algorithm – dynamic programming methods [41, 42] and heuristic "fast" algorithms such as FASTA and BLAST [43–45]. Both methods make use of amino acid substitution matrices such as PAM-250 and Blosum 62 [46, 47] to score and assess pairwise sequence alignments. Dynamic programming methods are very slow N^2 type algorithms that are guaranteed to find the mathematically optimum alignment between any two sequences. On the other hand, heuristic methods such as FASTA and BLAST are much faster N-type algorithms that find short local alignments and attempt to string these local alignments into a longer global alignment. Heuristic algorithms make use of statistical models to rapidly assess the significance of any local alignments, making them particularly useful for biologists trying to understand the significance of their matches. Exact descriptions and detailed assessments of these algorithms are beyond the scope of this chapter; suffice it to say that BLAST and its successors such as FASTA3 [48] and PSI-BLAST [45] have probably become the most commonly used "high-end" tools in biology.

BLAST-type searches are generally available for all major protein databases through a variety of mirror sites and Web servers (Tab. 14.1). Most servers offer a range of databases that can be "BLASTed". The largest and most complete database is GenBank's non-redundant (nr) protein database, which is largely equivalent to the translated EMBL (TREMBL) database. The Swiss-Prot database is the most completely annotated protein database, but does not contain the quantity of sequence data found in OWL or GenBank. The PIR database, which was started in the 1960s, is actually the oldest protein sequence database and contains many protein sequences determined by direct chemical or MS methods (-which are typically not in GenBank records). The PIR has recently joined with TREMBL and SwissProt to produce UniProt [36] which might now be the most complete protein-sequence database available. Most of these protein databases can be freely downloaded by academics, but industrial users must pay a fee.

BLAST and FASTA3 are particularly good at identifying sequence matches sharing between 25 % to 100 % identity with the query sequence. PSI-BLAST (position-specific iterated BLAST), on the other hand, is exceptionally good at identifying matches in the so-called twilight zone of between 15–25 % sequence identity. PSI-BLAST can also identify higher scoring similarities with the same accuracy as BLAST. The trick to using PSI-BLAST is to repeatedly press the "Iterate" button until the program indicates that it has converged. Apparently many first-time users of PSI-BLAST fail to

realize this by running the program only once; they come away with little more than a regular BLAST output. Nevertheless, because of its near universal applicability, PSI-BLAST is probably the best all-round tool for protein identification from sequence analysis.

The stunning success that PSI-BLAST has had in "scraping the bottom of the barrel" in terms of its ability to identify sequence relationships is leading to increased efforts by bioinformaticians aimed at trying to develop methods to identify even more remote sequence similarities from database comparisons. This has led to the development of a number of techniques such as threading, neural network analysis, and hidden Markov modeling – all of which are aimed at extracting additional information hidden in the sequence databases. Many of these techniques are described in more detail in Sect. 14.3.

14.3 Protein Property Prediction

Up to this point we have focused on how to identify a protein from a spot on a gel, an MS fingerprint, or by DNA or protein sequencing. When the identification problem has been solved, one is usually interested in finding out what this protein does or how and where it works. If a BLAST, PSI-BLAST, or PUBMED search turns up little in the way of useful information, it is still possible to employ a variety of bioinformatics tools to learn something directly from the protein's sequence. Indeed, as we shall see in the following pages, protein property prediction methods can often enable one to make a very good guess at the function, location, structure, shape, solubility, and binding partners of a novel protein long before one has even lifted a test-tube.

14.3.1 Predicting Bulk Properties (pI, UV absorptivity, MW)

Although the amino acid sequence of a protein largely defines its structure and function, a protein's amino acid composition can also provide a great deal of information. Specifically, amino acid composition can be used to predict a variety of bulk protein properties, for example isoelectric point, UV absorptivity, molecular weight, radius of gyration, partial specific volume, solubility, and packing volume – all of which can be easily measured on commonly available instruments (gel electrophoresis systems, chromatographic columns, mass spectrometers, UV spectrophotometers, amino acid analyzers, ultra-centrifuges, etc.). Knowledge of these bulk properties can be particularly useful in cloning, expressing, isolating, purifying, or characterizing a newly identified protein.

Many of these bulk properties can be calculated using simple formulas and commonly known properties, some of which are presented in Tabs. 14.2 and 14.3. Typical ranges found for water-soluble globular proteins are also shown in Tab. 14.2. A large number of these calculations can also be performed with more comprehensive protein bioinformatics packages such as SEQSEE [49, 50] and ANTHEPROT [51] and with many commonly available commercial packages (LaserGene, PepTool, VectorNTI).

14.3.2 Predicting Active Sites and Protein Functions

As more protein sequences are being deposited in data banks, it is becoming increasingly obvious that some amino acid residues remain highly conserved even among diverse members of protein fami-

Table 14.2 Formulas for protein property prediction.

Property	*Formula*	*Typical range*
Molecular weight	$MW = \Sigma A_i \times W_i + 18.01056$	N/A
Net charge (p*I*)	$Q = \_A_i/(1 + 10^{\mathrm{pH}} - \mathrm{p}K_i)$	N/A
Molar absorptivity	$\varepsilon = (5690 \times \#W + 1280 \times \#\mathrm{Y})/MW$	N/A
Average hydrophobicity	$AH = \Sigma A_i \times H_i$	($AH = -2.5 \pm 2.5$)
Hydrophobic ratio	$RH = \Sigma H(-)/\Sigma H(+)$	($RH = 1.3 \pm 0.5$)
Linear charge density	$\sigma = (\#K + \#R + \#D + \#E + \#H + 2)/N$	($\sigma = 0.25 \pm 0.5$)
Solubility	$\Pi = RH + AH + \sigma$	($\Pi = 1.6 \pm 0.5$)
Protein radius	$R = 3.875 \times (N^{0.333})$	N/A
Partial specific volume	$PSV = \Sigma PS_i \times W_i$	($PSV = 0.725 \pm 0.025$)
Packing volume	$VP = \Sigma A_i \times V_i$	N/A
Accessible surface area	$ASA = 7.11 \times MW^{0.718}$	N/A
Unfolded ASA	$ASA(U) = \Sigma A_i \times ASA_i$	N/A
Buried ASA	$ASA(\mathrm{B}) = ASA(U) - ASA$	N/A
Interior volume	$V_{\mathrm{int}} = \Sigma A_i \times FB_i \times V_i$	N/A
Exterior volume	$V_{\mathrm{ext}} = VP - V_{\mathrm{int}}$	N/A
Volume ratio	$VR = V_{\mathrm{ext}}/V_{\mathrm{int}}$	N/A
Fisher volume ratio	$FVR = [R^3/(R - 4.0)^3] - 1$	N/A

MW, molecular weight in Daltons; A_i, number of amino acids of type i; W_i, molecular weight of amino acid of type i; Q, charge; pK_i, pK_a of amino acid of type i; ε, molar absorptivity; $\#W$, number of tryptophans; H_i, Kyte Doolittle hydropathy; $H(-)$, hydrophilic residue hydropathy values; $H(+)$, hydrophobic residue hydropathy values; N, total number of residues in the protein; R, radius in Angstroms; PS_i , partial specific volume of amino acid of type i; V_i, volume in cubic Angstroms of amino acid of type i; ASA_i, accessible surface area in square Angstroms of amino acid of type i; FB_i, fraction buried of amino acid of type i. Residue-specific values for many of these terms are given in Tab. 14.3

lies. These highly conserved sequence patterns are often called signature sequences and often define the active site or function of a protein. Because most signature patterns are relatively short (7–10 residues) this kind of sequence information is not easily detected by BLAST or FASTA searches. Consequently it is always a good idea to scan against a signature sequences database (for example PROSITE) in an effort to obtain additional information about a protein's structure, function, or activity.

Active site or signature sequence databases come in two varieties – pattern-based and profile-based. Pattern-based sequence motifs are usually the easiest to work with, because they can be easily entered as simple regular expressions. Typically pattern-based sequence motifs are identified or confirmed by careful manual comparison of multiply aligned proteins – most of which are known to have a specific function, active site, or binding site [57]. Profile-based signatures or "sequence profiles" are usually generated as a combination of amino acid and positional scoring matrices, or hidden Markov models derived from multiple sequence alignments [58, 59]. Although sequence profiles are usually more robust than regular pattern expressions in identifying active sites, the effort required to prepare good sequence profiles has usually precluded their

Table 14.3 Amino acid residue properties (molecular weight [monoisotopic mass], frequency, pK_a, absorbance at 280 nm, hydrophobicity, partial specific volume (mL g^{-1}), packing volume (Å^3), accessible surface area (Å^2), fraction buried).

AA	MW (Da) [52]	v (%)	pK_a	ε_{280}	Hphb [53]	PS [54]	Vol. [55]	ASA [55]	FB [56]
A	71.03712	8.80	–	–	1.8	0.748	88.6	115	0.38
C	103.00919	2.05	10.28	–	2.5	0.631	108.5	135	0.45
D	115.02695	5.91	3.65	–	–3.5	0.579	111.1	150	0.15
E	129.04260	5.89	4.25	–	–3.5	0.643	138.4	190	0.18
F	147.06842	3.76	–	–	2.8	0.774	189.9	210	0.50
G	57.02147	8.30	–	–	–0.4	0.632	60.1	75	0.36
H	137.05891	2.15	6.00	–	–3.2	0.670	153.2	195	0.17
I	113.08407	5.40	–	–	4.5	0.884	166.7	175	0.60
K	128.09497	6.20	10.53	–	–3.9	0.789	168.6	200	0.03
L	113.08407	8.09	–	–	3.8	0.884	166.7	170	0.45
M	131.04049	1.97	–	–	1.9	0.745	162.9	185	0.40
N	114.04293	4.58	–	–	–3.5	0.619	117.7	160	0.12
P	97.05277	4.48	–	–	–1.6	0.774	122.7	145	0.18
Q	128.05858	3.84	–	–	–3.5	0.674	143.9	180	0.07
R	156.10112	4.22	12.43	–	–4.5	0.666	173.4	225	0.01
S	87.03203	6.50	–	–	–0.8	0.613	89.0	115	0.22
T	101.04768	5.91	–	–	–0.7	0.689	116.1	140	0.23
V	99.06842	7.05	–	–	4.2	0.847	140.0	155	0.54
W	186.07932	1.39	–	5690	–0.9	0.734	227.8	255	0.27
Y	163.06333	3.52	–	1280	–1.3	0.712	193.6	230	0.15

widespread adoption by the bioinformatics community.

Perhaps the best known and best documented signature sequence database is PROSITE [60]. It currently contains 1676 signature sequences and sequence profiles and extensive bibliographic and statistical information on all protein families, domains, or functional sites associated with each signature. The PROSITE team has also introduced a facility to search the PDB with a PROSITE entry or a user's pattern and to visualize the matched positions on associated 3D structures. PROSITE can be accessed through a variety of commercial bioinformatics programs and several freely available Web servers (Tab. 14.4). The database itself can also be downloaded and run locally. PROSITE is by no means the only active site or signature sequence site database available. SEQSITE [49] and several smaller databases have also been published or made available over the years [61–63]. More recently a new kind of signature sequence database has started appearing in the literature. These databases are automated compilations of multiply aligned sequence fingerprints, sequence blocks, position-specific scoring matrices, or hidden Markov models [64]. PRINTS [65], BLOCKS [66], InterPro [67] and the Pfam database [59] – all of which have Web access – are examples of a few of these family signature databases. Indeed, InterPro (which combines Pfam, PRINTS, and PROSITE) is now perhaps the most comprehensive protein site and protein domain identification resource available, with well over 10,000 archived motifs and signatures.

Table 14.4 Protein prediction tools – web links.

Tool/database	*Web Address*
Compute p*I*/Mw	http://ca.expasy.org/tools/pi_tool.html
Scan ProSite	http://ca.expasy.org/tools/scanprosite/
PROSITE database	http://ca.expasy.org/prosite/
BLOCKS database	http://blocks.fhcrc.org/
PRINTS database	http://bioinf.man.ac.uk/dbbrowser/PRINTS/
Pfam database	http://pfam.wustl.edu/
NetPhos server	http://www.cbs.dtu.dk/services/NetPhos/
NetOGlyc server	http://www.cbs.dtu.dk/services/NetOGlyc/
PSORT server	http://psort.nibb.ac.jp/
PSORT-B server	http://www.psort.org/psortb/
PA-SubCell	http://www.cs.ualberta.ca/~bioinfo/PA/Sub/
HMMTOP	http://www.enzim.hu/hmmtop/html/submit.html
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/
KEGG database	http://www.genome.ad.jp/kegg/regulation.html
BIND database	http://www.blueprint.org/bind/bind.php
DIP database	http://dip.doe-mbi.ucla.edu/
MINT database	http://mint.bio.uniroma2.it/mint/
IntAct database	http://www.ebi.ac.uk/intact/index.html
PESTfind	http://bioweb.pasteur.fr/seqanal/interfaces/pestfind-simple.html
PredictProtein	http://cubic.bioc.columbia.edu/predictprotein/
ProDom database	http://protein.toulouse.inra.fr/prodom/current/html/form.php
CDD database	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
COILS (coil–coil prediction)	http://paircoil.lcs.mit.edu/cgi-bin/paircoil
SAM-T02 (2° prediction)	http://www.cse.ucsc.edu/research/compbio/HMM-apps/
PSIpred (2° prediction)	http://bioinf.cs.ucl.ac.uk/psipred/
Prospector (Threading)	http://www.bioinformatics.buffalo.edu/
GenThreader Server	http://bioinf.cs.ucl.ac.uk/psipred/
3DPSSM (Threading)	http://www.sbg.bio.ic.ac.uk/~3dpssm/
ANTHEPROT	http://antheprot-pbil.ibcp.fr/
SEQSEE	http://www.pence.ualberta.ca/ftp/seqsee/seqsee.html

A particularly useful feature of the InterPro database is its linkage to the Gene Ontology (GO) database and GO annotations [68]. The GO consortium is a community-wide annotation effort aimed at providing expert-derived, consensus information about three very specific protein features or "qualities" – molecular function, biological processes, and cellular components. The intent is to describe these features/processes using a controlled or carefully structured vocabulary. According to GO nomenclature, *Molecular function* refers to the tasks performed by individual proteins, for example carbohydrate binding and ATPase activity, *Biological process* refers to the broad biological goals, for example mitosis or purine metabolism, accomplished by a collection or sequence of molecular functions, whereas *Cellular component* refers to the subcellular structures, locations, and macromolecular complexes in which a protein may asso-

ciate or reside. The ultimate aim of GO is to annotate all proteins with detailed (experimental or inferred) information about these three kinds of feature. Using the Genome Ontology AmiGO GOst server (Tab. 14.2) it is possible to query the GO database with a protein sequence (via BLAST) and determine both the GO numbers and GO annotation that either it or its homologs might have. In this way protein functions and protein locations can be readily determined (as long as a homolog has been annotated by the GO consortium). Alternatively, via the Pfam or InterPro searches one can also determine appropriate GO annotations.

14.3.3 Predicting Modification Sites

Post-translational modification, for example proteolytic cleavage, phosphorylation, glycosylation, sulfation, and myristylation can greatly affect the function and structure of proteins. Identification of these sites can be quite helpful in learning something about the function, preferred location, probable stability, and possible structure of a protein. Furthermore, knowledge of the location or presence of these modification sites can assist in selecting expression systems and designing purification procedures. Most post-translational modifications occur at specific residues contained within well-defined sequence patterns. Consequently, many of these modification sites are contained in the PROSITE database and can be detected through a simple PROSITE scan. For glycosylation and phosphorylation, however, the sequence patterns are usually less-well defined and the use of simple sequence patterns can lead to many false positives and false negatives. Neural networks trained on databases of experimentally determined phosphorylation and glycosylation sites have been shown to be somewhat more specific (>70 %) for finding these hard-to-define sites [69, 70]. NetPhos and NetOGlyc (Tab. 14.4) are examples of two Web-servers that offer neural network identification of potential phosphorylation and O-glycosylation sites on proteins.

14.3.4 Finding Protein Interaction Partners and Pathways

Sequence information alone can rarely provide sufficient information to indicate how or where a protein interacts with other proteins or where it sits within a given metabolic pathway. Although detailed literature surveys and careful keyword searches through PubMed can be of some help, there are now several freely available databases enabling much easier and much more specific querying, visualization, and identification of protein–protein interactions in the context of metabolic or signaling pathways. Perhaps the oldest and most complete of these databases is KEGG (Kyoto Encyclopedia of Genes and Genomes) maintained by the Kyoto Institute of Chemical Research in Japan [71]. This data-rich, manually curated resource provides information on enzymes, small-molecule ligands, metabolic pathways, signal pathways, reaction diagrams, hyperlinked "clickable" metabolic charts, schematics of protein complexes, EC (enzyme classification) numbers, external database links, and a host of other data that is particularly useful to any protein chemist wanting to learn more about their protein.

Protein interaction information can be quite revealing, particularly if one is trying to gain some "context" about why a particular protein is expressed, how it is being regulated, or where it is found. As a result, there has been an explosion in the number and variety of protein interaction databases designed to store experimentally deter-

mined protein complex or protein interaction information. Among the most comprehensive (and popular) are the Biomolecular Interaction Network Database – BIND [72], the Database of Interacting Proteins – DIP [73], the Interaction Database – IntAct [74], and the Molecular Interaction Database – MINT [75]. Their URLs are listed in Tab. 14.4. Most of these interaction databases provide primitive interactive (applet) visual descriptors, textual information, and hyper-linked pointers to help users understand or explore the role a given query protein plays in a particular pathway, protein complex, or transient protein interaction. Most of these databases also enable users to conduct BLAST searches to identify possible interaction homologs or "interologs". In addition to these archival interaction resources, a more useful and far more visually pleasing compilation of pathways and protein interactions is maintained by BioCarta as part of their Proteomics Pathways Project. Currently more than 400 pathways are maintained in the BioCarta database, with each pathway or process containing detailed textual descriptions, hyperlinked image maps (containing cell components, pathway arrows, hyperlinked protein, topological information, names, cell types, and tissue types) and hyperlinked protein lists.

The field of protein interaction analysis is experiencing a very rapid period of growth, not unlike the rapid growth experienced by the Human Genome Project during the late 1990s. New proteomic technologies are being combined with innovative computational approaches to produce very comprehensive resources and very powerful protein annotation tools. In the not-too-distant future one can expect that many protein interaction databases will become increasingly consolidated and that the quantity and quality of information contained in these resources will eventually make them as important as sequence databases in terms of protein annotation and analysis.

14.3.5 Predicting Sub-cellular Location or Localization

Many proteins have sequence-based characteristics that will localize the protein in certain regions of the cell. For instance, proteins with transmembrane helices will end up in the lipid bilayer, proteins rich in positively charged residues tend to end up in the nucleus, and proteins with specific signal peptides will be exported outside the cell [76]. Being able to identify a signaling or localization sequence can aid understanding of the function, probable location, and biological context of a newly identified protein. This information can also be quite useful in designing cloning, purification, and separation procedures. Several signaling sequences are contained in the PROSITE database and so a simple PROSITE scan can be quite revealing. Not all protein localization sequences are easily defined as simple PROSITE patterns, however. To cover signaling sequences that are not so readily identified, a variety of protein localization tools have been developed by use of several different approaches. One approach is based on amino acid composition, using artificial neural nets (ANN) such as NNPSL [77], or support vector machines (SVM) like SubLoc [78]. A second approach, such as TargetP [79], uses the existence of peptide signals which are short sub-sequences of approximately 3 to 70 amino acids to predict specific cellular localization. A third approach, such as that used in LOCkey [80], is to do a similarity search on the sequence, extract text from any matching homologs, and to use a classifier on the text features to predict the location. A number of other tools, for example PSORT [81] and PSORTB

[82], combine a variety of individual predictors such as hidden Markov models and nearest-neighbor classifiers to predict protein localization. A new approach has recently been described which combines machine learning methods with the text analysis approach pioneered by LOCkey to predict subcellular localization [83]. According to the authors, the PA SubCell server seems to be more accurate and much more comprehensive than any other method described to date, with an average accuracy greater than 90 % (approx. 10 % better than most other tools) and a level of coverage that is 3–5 times greater in terms of the number of sequences or types of organism it can handle. A listing of some of the better subcellular localization Web servers is given in Tab. 14.4.

In addition to these subcellular prediction servers, one might also employ a variety of other programs to identify those proteins that localize to the cell membrane (which account for up to 30 % of all proteins) and which parts of the protein are actually inserted into the membrane. Transmembrane helix prediction was done historically by using hydropathy or hydrophobicity plots [53, 84]. This graphical technique often proved to be inconsistent and unreliable, however. The introduction of neural network-based approaches and hidden Markov models combined with multiple sequence alignments (or evolutionary information) has greatly improved the quality and reliability of transmembrane helix prediction [85]. Although some groups have claimed prediction accuracies in excess of 95 % [86], more recent, independent assessments indicate that membrane helix prediction is approximately 75–80 % accurate [87], with the best tools generally being HMMTOP [85] and TMHMM [88]. Even with this "reduced" level of accuracy, identification of membrane proteins and membrane spanning locations is probably one of the more robust and reliable predictive methods in all of bioinformatics. Interestingly, although transmembrane helices are quite predictable, transmembrane beta-sheets (as found in porins) are not. This problem continues to be one of the unmet challenges in membrane protein sequence analysis. A list of URLs for several membrane spanning prediction Web servers is given in Tab. 14.4.

14.3.6 Predicting Stability, Globularity, and Shape

Before cloning or expressing a protein or protein fragment, obviously one would like to know whether it will be soluble, stable, and globular. Although crude predictors based on average hydrophobicity, localized hydrophobicity, hydrophobic ratios, charge density, and secondary structure are available, prediction of protein solubility and expressibility is still on rather shaky ground [89, 90]. Some general rules or equations are provided in Tab. 14.2, but these, too, are only approximate. Nevertheless, as more data from different structural proteomics efforts around the globe are compiled and analyzed, a clearer idea is being obtained about the key sequence/property features that determine the likelihood of successful, high-yield expression. One example of a protein sequence feature that determines stability is the so-called PEST sequence [91]. Eukaryotic proteins with intracellular half-lives of less than 2 h are often found to contain regions rich in proline, glutamic acid, serine, and threonine (P, E, S and T). These PEST regions are usually flanked by clusters of positively charged amino acids. Identification of PEST sequences in proteins can therefore be a very important consideration in any protein expression project. At least one PEST web server (Tab. 14.4) is available for identifica-

tion of PEST sequences using a standard set of pattern-based rules. Similar information about intracellular protein lifetimes can be extracted using the N-end rule for protein ubiquitination [92].

The propensity of a protein to fold into a particular shape or to fold into a globule can also be predicted. Indeed, as far back as 1964 a very simple but elegant procedure was developed to predict the shape and globularity of proteins on the basis of simple packing rules, hydrophobicity, and amino acid volumes [93]. Specifically, if the calculated volume ratio for a given protein (Tabs. 14.2 and 14.3) is slightly greater than the theoretical Fisher volume ratio, the protein probably forms a soluble monomer. If the calculated volume ratio is much greater than the Fisher volume ratio, the protein probably does not form a compact globular structure (i.e. it is filamentous or unfolded). This procedure, with a few modifications, has been implemented in both the SEQSEE and PepTool software packages [49, 50]. A more sophisticated approach to solubility/globularity prediction has been developed by Burkhard Rost. This technique uses evolutionary information, neural networks, and predicted protein accessibility to determine whether or not a protein will form a globular domain. The PredictProtein Web server [94], listed in Tab. 14.4, provides this domain prediction service with a host of other predictions (secondary structure, accessibility, disulfide bonds, transmembrane helices, coiled-coil regions, and structural switch regions).

14.3.7 Predicting Protein Domains

Larger proteins will tend to fold into structures containing multiple domains. Typically domains are defined as contiguous stretches of 100–150 amino acids with a globular fold or function distinct from other parts of the protein. Many eukaryotic proteins are composed of multiple sub-units or domains, each having a different function or a different fold. If one can identify (by sequence analysis) the location or presence of well-folded, well-defined domains, it is often possible to gain greater understanding of not only the probable function but also the evolutionary history of a protein. The identification of domains can, furthermore, often enable one to "decompose" a protein into smaller parts to facilitate cloning and expression and to increase the likelihood that the protein, or parts of it, can be studied by X-ray or NMR techniques.

As with active-site identification, domain identification is typically performed by using comparisons or alignments to known domains. Several databases are now available, including Pfam [59], the conserved domain database or CDD [95], and Prodom [96, 97] (all have their URLs listed in Tab. 14.4). These represent compilations of protein domains that have been identified both manually and automatically through multiple sequence alignments and hierarchical clustering algorithms. All three databases can achieve quick and reliable identification of most globular protein domains, with CDD results being returned in many BLAST searches performed through the NCBI website. The CDD and Pfam both provide direct links to the domain's 3D structure. Some protein domains seem to defy routine identification, however. For example, specialized software had to be developed to aid identification of coiled-coil domains, with their nondescript sequence character and non-globular nature [98]. Several coiled-coil prediction services are also available over the Web (Tab. 14.4). Another very useful, but unpublished, website (http://www.mshri.on.ca/pawson/domains.html maintained by Dr Tony Pawson's laboratory) pro-

vides detailed descriptions, pictures, and structures of dozens of known protein–protein interaction domains.

14.3.8 Predicting Secondary Structure

The primary structure (the sequence) determines both the secondary structure (helices and beta-strands) and the tertiary structure (the 3D fold) of a protein. In principle, if you know a protein's sequence, you should be able to predict both its structure and function. Prediction of secondary structure is one way of gaining insight into the probable folds, expected domain structure, and possible functions of proteins [99, 100]. Because secondary structure is more conserved than primary structure, prediction of secondary structure might also be used to facilitate remote homolog detection or protein fold recognition [101].

Prediction of secondary structure has been under development for more than 30 years. As a consequence, many different techniques are available with widely varying accuracy and utility. The simplest approaches are statistical [102] wherein intrinsic probabilities of amino acids being in helices and beta-strands are simply averaged over different window sizes. Amino acid segments with the highest local scores for a particular secondary structure are then assigned to that structure. Secondary structure usually depends on more than just averaged conformational preferences of individual amino acids, however. Sequence patterns, positional preferences, long-range effects, and pairwise interactions are also important. To account for these effects more sophisticated predictive approaches have had to be developed. These include information theoretical or Bayesian probabilistic approaches [103]; stereochemical methods [104]; nearest-neighbor or database comparison techniques [105]; and neural network approaches [106]. Typically these methods achieve a three-state (helix, sheet, coil) accuracy of between 55 % (for the simplest statistical methods) to 65 % (for the best neural network or nearest-neighbor approaches) for water-soluble globular proteins.

A significant improvement in the accuracy of prediction of secondary structure occurred in the early 1990s with the introduction of combined approaches that integrated multiple sequence alignments (i.e. evolutionary information) with neural network pattern-recognition methods [107]. This innovation enabled the accuracy of prediction of secondary structure to be improved to better than 72 %. Similar efforts aimed at integrating evolutionary information with nearest-neighbor approaches also led to comparable improvements in prediction accuracy [108]. Most recently, integration of better database-searching methods (PSI-BLAST) has enabled the accuracy of prediction of protein secondary structure to approach 77 % [100, 109]. Many of these "new and improved" methods for prediction of secondary structure are now freely available over the Web (Tab. 14.4).

Given their wide availability and much improved reliability, predictions of secondary structure should now be regarded as integral components of any standard protein sequence analysis. With steady, incremental improvements in prediction accuracy occurring every one or two years, it is likely that secondary structure prediction will soon achieve an accuracy in excess of 80 %. At this level of accuracy it might be possible to use secondary structure predictions as starting points for 3D structure prediction.

14.3.9 Predicting 3D Folds (Threading)

Threading is a protein-fold recognition or structure prediction technique that got its name because it conceptually resembles the method used to thread electrical cables through a conduit. Specifically threading involves placing or threading an amino acid sequence on to a database of different secondary or tertiary structures (pseudo-conduits). As the sequence is fed through each structure its fit or compatibility to that structure is evaluated by using a heuristic potential [110]. This evaluation might be achieved quickly by using an empirical "energy" term or some measure of packing efficiency or secondary structure propensity. In this way it is possible to assess which protein sequences are compatible with the given backbone fold. While one would clearly expect that those sequences homologous to the original template sequence should fit best, it has been found that this approach can occasionally reveal that some seemingly unrelated sequences can also fit into previously known folds. The protein leptin (the hormone responsible for obesity) is one interesting and early example of how successful threading can be in structure and function prediction. Standard threading techniques were able to show that leptin was a helix-like cytokine long before any confirmatory X-ray structure had been determined or biological receptors had been found [111].

Two approaches to threading are used. One, 3D threading, is classified as a distance-based method (DBM). The other, 2D threading, is classified as a prediction-based method (PBM). 3D threading was first described in the early 1980s [112] and then "rediscovered" about 10 years later [113–115] when the concept of heuristic potential functions matured. 3D threading uses distance-based or profile-based [116] energy functions and technically resembles the "pipe" threading description given earlier. In 3D threading, coordinates corresponding to the hypothesized protein fold are actually calculated and the energy functions evaluated on the basis of these 3D coordinates.

Just like 3D threading, 2D threading was first described in the mid-1980s [117] then rediscovered in the mid 1990s [101, 118, 119] when the reliability of secondary structure predictions began to improve. Rather than relying on 3D coordinates to evaluate the quality of a fold, 2D threading actually uses secondary structure (hence the name 2D) as the primary evaluation criterion. Indeed, 2D threading is based on the simple observation that secondary structure is more conserved than primary structure (sequence). Therefore, proteins that have lost detectable similarity at the sequence level, could still be expected to maintain some similarity at the secondary structure level.

Over the past few years 2D threading has matured so that secondary structure, solvent accessibility and sequence information can now be used in the evaluation process. The advantage that 2D threading has over 3D threading is that all this structural information can be encoded into a 1D string of symbols (i.e. a pseudo sequence). This enables the use of standard sequence comparison tools, for example dynamic programming, for rapid comparison of a query sequence and/or secondary structure with members of a database of sequences and secondary structures. Consequently, 2D threading is 10 to 100 times faster than distance-based 3D threading and seems to give comparable (and occasionally even better) results than 3D threading. The fact that the 2D threading algorithm is relatively simple

to understand and to implement has led to the development of a number of freely available 2D threading servers (Tab. 14.4).

Typically the best 2D threading methods score between 30 and 40 % when working with "minimal" databases. If the structural databases are expanded to include more than one related fold representative, the performance can be as high as 70–75 %. As already mentioned, 2D threading performs about as well as 3D threading; it is, however, much faster, and easier to implement. It is generally thought that if 2D threading approaches could improve their secondary structure prediction accuracy and include more information about the "coil" state (such as approximate dihedral angles) then even further performance gains could be realized. Similarly, if initial 2D threading predictions could be verified using 3D threading checks (post-threading analysis) and further tested by looking at other biochemical information (species of origin, known function, ligand contacts, binding partners) additional improvements should be possible.

14.3.10 Comprehensive Commercial Packages

Commercial packages can offer an attractive alternative to many of the specialized or single-task analyses offered over the Web. In particular, these commercial tools integrate sophisticated graphical user interfaces (GUI) with a wide range of well-tested analytical functions, databases, plotting tools, and well-maintained user-support systems to make most aspects of protein sequence analysis simpler, speedier, and "safer" than is possible with Web-based tools. Commercial packages also have their disadvantages, however. Most are quite expensive ($1000–$3000) and most do not offer the range or even the currency of analytical functions available via the Web. Furthermore, most commercial packages are very platform-specific, i.e. they run only on certain types of computer or with a specific operating system. This is not a limitation for Web-based tools, as most are platform independent.

The first commercial packages appeared in the mid 1980s and over the past 20 years they have evolved and improved substantially. Most packages integrate both protein and DNA sequence analysis into a single "suite", although some companies have opted to create separate protein-specific modules. Most commercial packages offer a fairly standard set of protein analysis tools including:

1. bulk property calculations (molecular weight, p*I*);
2. physicochemical property plotting (hydrophobicity, flexibility, hydrophobic moments);
3. antigenicity prediction;
4. sequence motif searching or identification;
5. secondary structure prediction;
6. database searching (internet or local);
7. alignment and comparison tools (dotplots or multiple alignment);
8. plotting or publication tools; and
9. multiformat (GenBank, EMBL, SwissProt, PIR) sequence input/output.

It would be impossible to review all the commercial packages here, but a brief summary of some of the more popular tools is given below.

Accelerys (www.accelerys.com) offers perhaps the widest range of protein analysis tools including the GCG Wisconsin package (UNIX), Discovery Studio Gene (Windows), and MacVector (MacOS). The Wisconsin package, with its new graphical interfaces (SeqWeb, SeqLab, and DS Gene), is one of the most comprehensive and widely distributed bioinformatics packages in

the world. It offers an impressive array of standard tools and some very good motif-detection routines (Meme) and sophisticated sequence profiling algorithms. With literally dozens of protein-analysis tools, the GCG Wisconsin package is much more complete than either Discovery Studio Gene or MacVector. The interface and user-friendliness of GCG are still well behind industry standards, however.

DNAstar (www.dnastar.com) produces a highly acclaimed multi-component suite called LaserGene (MacOS, Windows). The protein sequence-analysis module (Protean) has a well-conceived GUI and an array of sophisticated plotting and visualization tools. In addition to the usual analytical tools the Protean module also offers some very innovative facilities for synthetic peptide design, linear structure display, and SDS PAGE gel simulation. Unfortunately, Protean does not offer the usual sequence comparison or alignment tools typically found in most comprehensive packages. Instead these must be obtained by purchasing a second stand-alone module called Megalign.

Informax (www.informaxinc.com), through its parent company Invitrogen, produces and distributes a very popular package called Vector NTI (Windows, MacOS). Although the emphasis is clearly on DNA analysis, the Vector NTI package also offers a modestly comprehensive set of protein tools presented in an easy-to-use GUI. Vector NTI supports most standard protein analytical functions and, as with many commercial packages, offers integrated internet connectivity to Entrez, PubMed, and BLAST. Vector NTI also supports a good, interactive 3D molecular visualization tool and a number of interesting property evaluation functions to calculate free energies, polarity, refractivity, and sequence complexity.

BioTools (www.biotools.com) produces perhaps the most comprehensive and least expensive (<$200) protein-analysis packages. Its platform-independent (Windows, MacOS, UNIX) protein analysis suite is called PepTool. As with most other commercial packages, PepTool supports all of the major protein analytical functions and offers a particularly broad range of statistical and property prediction tools. In addition to being the only package to offer universal platform compatibility, PepTool also has a particularly logical, easy-to-learn interface with web connectivity to most major protein databases. Most recently PepTool has included a comprehensive 3D structure visualization and modeling module which adds considerably to its utility and ease-of-use.

A question often asked by both novice and experienced users is: Should I choose expensive commercial packages or should I stick with freely available Web tools? There is no straightforward answer. If one is looking for the latest analytical tool or best-performing prediction algorithm, the Web is almost always the best place to find it. If, however, high-quality plotting, graphing, or rendering is important (for papers or presentations) one often has to turn to commercial packages. Helical wheels, colored multiple alignments, overlaid or stacked plots, annotated graphs, graphical secondary structure assignments, etc., are examples of images that cannot yet be rendered (well) on the Web. For many commercial biotech firms, security, uniformity, and reliability are particularly important and so once again, commercial packages – with their resident databases, uniform interfaces, and reliable user-support – are definitely the tools of choice.

Irrespective of one's decision to choose freeware or commercial software, one thing should be clear – without access to the tools

or database described here, we would truly be "lost" in a blizzard of biological data. Biological science and, particularly, protein science has definitely entered the computer age.

Acknowledgments

The author wishes to acknowledge Genome Canada and the Protein Engineering Network of Centers of Excellence (PENCE) for financial support.

References

1 Genome News Network (April, 2004) http://www.genomenewsnetwork.org/sequenced_genomes/genome_guide_p1.shtml.

2 Green, E.D.(2001) Strategies for the systematic sequencing of complex genomes, Nat Rev Genet. 2, 573–583.

3 Broder, S., Venter, J.C. (2000) Sequencing the entire genomes of free-living organisms: the foundation of pharmacology in the new millennium, Annu. Rev. Pharmacol. Toxicol. 40, 97–132.

4 Giometti, C.S. (2003) Proteomics and bioinformatics, Adv. Protein Chem. 65, 353–369.

5 Zhu, H., Bilgin, M., Snyder, M. (2003) Proteomics, Annu. Rev. Biochem. 72, 783–812.

6 Gevaert, K., Vandekerckhove, J. (2000) Protein identification methods in proteomics, Electrophoresis 21, 1145–1154.

7 Hoving, S., Gerrits, B., Voshol, H., Muller, D., Roberts, R.C., van Oostrum, J. (2002) Preparative two-dimensional gel electrophoresis at alkaline pH using narrow range immobilized pH gradients, Proteomics 2, 127–134.

8 Von Eggeling, F., Gawriljuk, A., Fiedler, W., Ernst, G., Claussen, U., Klose, J., Romer, I. (2001) Fluorescent dual colour 2D-protein gel electrophoresis for rapid detection of differences in protein pattern with standard image analysis software, Int. J. Mol. Med. 8, 373–377.

9 Appel, R., Hochstrasser, D., Roch, C., Funk, M., Muller, A.F., Pellegrini, C. (1988) Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning, Electrophoresis 9, 136–142.

10 Appel, R.D., Vargas, J.R., Palagi, P.M., Walther, D., Hochstrasser, D.F. (1997) Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images, Electrophoresis. 18, 2735–2748.

11 Young, N., Chang, Z., Wishart, D.S. (2004) GelScape: a web-based server for interactively annotating, manipulating, comparing and archiving 1D and 2D gel images, Bioinformatics, 20, 976–978.

12 Lemkin, P.F., Thornwall, G. (1999) Flicker image comparison of 2-D gel images for putative identification using the 2DWG meta-database, Mol. Biotechnol. 12, 159–172.

13 Lemkin, P.F., Thornwall, G. (2002) Flicker image comparison of 2-D gel images over the Internet. John Walker (ed) The Protein Protocols Handbook, pp197–214.

14 Appel, R.D., Bairoch, A., Sanchez, J.C., Vargas, J.R., Golaz, O., Pasquali, C., Hochstrasser, .D.F. (1996) Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data, Electrophoresis 17, 540–546.

15 Celis, J.E., Ostergaard, M., Jensen, N.A., Gromova, I., Rasmussen, H.H., Gromov, P. (1998) Human and mouse proteomic databases: novel resources in the protein universe, FEBS Lett. 430, 64–72.

16 Lemkin, P.F., Myrick, J.M., Lakshmanan, Y., Shue, M.J., Patrick, J.L., Hornbeck, P.V., Thornwall, G., Partin, A.W. (1999) Exploratory data analysis groupware for qualitative and quantitative electrophoretic gel analysis over the Internet-WebGel, Electrophoresis 20, 3492–3507.

17 Hoogland, C., Sanchez, J.C., Tonella, L., Binz, P.A., Bairoch, A., Hochstrasser, D.F., Appel, R.D. (2000) The 1999 SWISS-2DPAGE database update, Nucleic Acids Res. 28, 286–288.

18 Stanislaus, R., Jiang, L.H., Swartz, M., Arthur, J., Almeida, J.S. (2004) An XML standard for the dissemination of annotated 2D gel electrophoresis data complemented with mass spectrometry results, BMC Bioinformatics 5, 9.

19 Ashcroft, A.E. (2003) Protein and peptide identification: the role of mass spectrometry in proteomics, Nat. Prod. Rep. 20, 202–215.

20 Yates, J.R. (2000) Mass spectrometry. From genomics to proteomics, Trends Genet. 16, 5–8.

21 Kislinger, T., Emili, A. (2003) Going global: protein expression profiling using shotgun mass spectrometry, Curr. Opin. Mol. Ther. 5, 285–293.

22 Pappin, D.J.C., Hojrup, P., Bleasby, A.J. (1993) Rapid identification of proteins by peptide-mass fingerprinting, Curr. Biology 3, 327–332.

23 Yates, J.R., Speicher, S., Griffin, P.R., Hunkapiller, T. (1993) Peptide mass maps – a highly informative approach to protein identification, Anal. Biochem. 214, 397–408.

24 Mann, M., Hojrup, P., Roepstorff, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases, Biol Mass Spectrom. 22, 338–345.

25 Zhang, W., Chait, B.T. (1995) Proc. 43rd ASMS Conf. Mass Spectrometry and Allied Topics, Atlanta, Georgia.

26 Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis 20, 3551–3567.

27 Eng, J.K., McCormack, A.L., Yates, J.R. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid sequences in a protein database, J. Am. Soc. Mass. Spectrom. 5, 976–989.

28 Yates, J.R., Eng, J.K., Glauser, K.R., Burlingame, A.L. (1996) Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides, J. Am. Soc. Mass Spectrom. 7, 1089–1098.

29 Clauser, K. R., Baker P. R., Burlingame, A. L. (1999) Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Analytical Chemistry, 71, 2871–2875.

30 Bergquist, J. (2003) FTICR mass spectrometry in proteomics, Curr. Opin. Mol. Ther. 5, 310–314.

31 Goodlett, D.R., Bruce, J.E., Anderson, G.A., Rist, B., Pasa-Tolic, L., Fiehn, O., Smith, R.D., Aebersold, R. (2000) Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching, Anal. Chem. 72, 1112–1118.

32 Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., Smith, M. (1977) Nucleotide sequence of bacteriophage phi X174 DNA, Nature 265, 687–695.

33 Bleasby, A.J., Akrigg, D., Attwood, T.K. (1994) OWL – a non-redundant composite protein sequence database, Nucleic Acids Res. 22, 3574–3577.

34 Wu, C.H., Yeh, L.S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E., Vinayaka, C.R., Zhang, J., Barker, W.C. (2003) The Protein Information Resource, Nucleic Acids Res. 31, 345–347.

35 Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. 31, 365–370.

36 Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger E,, Huang. H,, Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S. (2004) UniProt: the Universal Protein knowledgebase, Nucleic Acids Res. 32 Database issue:D115–119.

37 Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L. (2004) GenBank: update, Nucleic Acids Res. 32, Database issue:D23–26.

38 Shevchenko, A., Wilm, M., Mann, M. (1997) Peptide sequencing by mass spectrometry for homology searches and cloning of genes, J. Protein Chem. 16, 481–490.

39 Raska, C.S., Parker, C.E., Sunnarborg, S.W., Pope, R.M., Lee, D.C., Glish, G.L., Borchers, C.H. (2003) Rapid and sensitive identification of epitope-containing peptides by direct matrix-assisted laser desorption/ionization tandem mass spectrometry of peptides affinity-bound to antibody beads, J. Am. Soc. Mass Spectrom. 14, 1076–1085.

40 Doolittle, R.F., Bork, P. (1993) Evolutionarily mobile modules in proteins, Sci. Am. 269, 50–56.

41 Needleman, S.B., Wunsch, C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48, 443–453.

42 Smith, T.F., Waterman, M.S. (1981) Identification of common molecular subsequences, J. Mol. Biol. 47, 195–197.

43 Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA 85, 2444–2448.

44 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool, J. Mol. Biol., 215, 403–410.

45 Altschul, S. F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25, 3389–3402.

46 Dayhoff, M.O., Barker, W.C., Hunt, L.T. (1983) Establishing homologies in protein sequences, Methods Enzymol. 91, 534–545.

47 Henikoff, S., Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. USA 89, 10915–10919.

48 Pearson, W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package, Methods Mol. Biol. 132, 185–219.

49 Wishart, D.S., Boyko, R.F., Willard, L., Richards, F.M., Sykes, B.D. (1994) SEQSEE: a comprehensive program suite for protein sequence analysis, Comput. Appl. Biosci. 10, 121–132.

50 Wishart, D.S., Stothard, P., Van Domselaar, G.H. (2000) PepTool and GeneTool: platform-independent tools for biological sequence analysis, Methods Mol. Biol. 132, 93–113.

51 Deleage, G., Combet, C., Blanchet, C., Geourjon, C. (2001) ANTHEPROT: an integrated protein sequence analysis software with client/server capabilities, Comput. Biol. Med. 31, 259–267.

52 Biemann, K. (1990) Appendix 6. Mass values for amino acid residues in peptides in: Methods in Enzymology (McCloskey, J. A., Ed.) Vol. 193, pp. 888. San Diego: Academic Press.

53 Kyte, J., Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157, 105–132.

54 Zamayatnin, A.A. (1972) Protein volume in solution, Prog. Biophys. Mol. Biol. 24, 107–123.

55 Richards, F.M. (1977) Areas, volumes, packing and protein structure, Annu. Rev. Biophys. Bioeng. 6, 151–175.

56 Chothia, C. (1976) The nature of the accessible and buried surfaces in proteins, J. Mol. Biol. 105, 1–14.

57 Bairoch, A. (1991) PROSITE: A dictionary of sites and patterns in proteins, Nucleic Acids Res. 19, 2241–2245.

58 Gribskov, M., McLachlan, A.D., Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins, Proc. Natl. Acad. Sci. USA 84, 4355–4358.

59 Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., Eddy, S.R. (2004) The Pfam protein families database, Nucleic Acids Res. 32, Database issue:D138–141.

60 Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., Bairoch, A. (2004) Recent improvements to the PROSITE database, Nucleic Acids Res. 32, Database issue:D134–137.

61 Hodgman, T.C. (1989) The elucidation of protein function by sequence motif analysis, Comput. Appl. Biosci. 5, 1–13.

62 Seto, Y., Ikeuchi, Y., Kanehisa, M. (1990) Fragment peptide library for classification and functional prediction for proteins, Proteins: Struct. Funct. Genet. 8, 341–351.

63 Ogiwara, A., Uchiyama, I. Seto, Y., Kanehisa, M. (1992) Construction of a dictionary of sequence motifs that characterize groups of related proteins, Protein Engineering 5, 479–488.

64 Hofmann, K. (2000) Sensitive protein comparisons with profiles and hidden Markov models. Brief. Bioinform. 1, 167–178.

65 Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A., Zygouri, C. (2003) PRINTS and its automatic supplement, preprints, Nucleic Acids Res. 31, 400–402.

66 Henikoff, S., Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching, Nucl. Acids Res. 19, 6565–6572.

67 Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J., Vaughan, R., Zdobnov, E.M. (2003) The InterPro Database, 2003 brings increased coverage and new features, Nucleic Acids Res. 31, 315–318.

68 Harris, M.A., Clark, J., Ireland, A., Lomax .J, Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M,, Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., et al. Gene Ontology Consortium. (2004) The Gene Ontology (GO) database and informatics resource, Nucleic Acids Res. 32, Database issue:D258–261.

69 Blom, N., Gammeltoft, S., Brunak, S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites, J. Mol. Biol. 294, 1351–1362.

70 Hansen, J.E., Lund, O., Tolstrup, N., Gooley, A.A., Williams, K.L., Brunak, S. (1998) NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility, Glycoconj. J. 15, 115–130.

71 Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M. (2004) The KEGG resource for deciphering the genome, Nucleic Acids Res. 32, Database issue:D277–280.

72 Bader, G.D., Betel, D., Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database, Nucleic Acids Res. 31, 248–250.

73 Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 32, Database issue:D449–451.

74 Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R.(2004) IntAct: an open source molecular interaction database. Nucleic Acids Res. 32, Database issue:D452–455.

75 Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G. (2002) MINT: a Molecular INTeraction database, FEBS Lett. 513, 135–140.

76 Nielsen, H., Brunak. S., von Heijne, G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals, Protein Eng. 12, 3–9.

77 Reinhardt, A., Hubbard, T. (1998) Using neural networks for prediction of the subcellular location of proteins, Nucleic Acids Res. 26, 2230–2236.

78 Hua, S., Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction, Bioinformatics 17, 721–728.

79 Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, J. Mol. Biol. 300, 1005–1016.

80 Nair, R., Rost, B. (2002) Inferring subcellular localization through automated lexical analysis, Bioinformatics 18, S78–S86.

81 Horton, P., Nakai, K. (1997) Better prediction of protein cellular localization sites with the k nearest neighbor classifier, Intell. Syst. Mol. Biol. 5, 147–152.

82 Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnády, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K., Brinkman, F.S.L. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res. 31, 3613–3617.

83 Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D.S., Poulin, B., Anvik, J., Macdonell, C., Eisner, R. (2004), Predicting subcellular localization using machine-learned classifiers in proteome analyst, Bioinformatics 20, 547–556.

84 Engelman, D.M., Steitz, T.A., Goldman, A. (1986) Identifying non-polar transbilayer helices in amino acid sequences of membrane proteins, Annu. Rev. Biophys. Biophys. Chem. 15, 321–353.

85 Tusnady, G.E., Simon, I. (2001) The HMMTOP transmembrane topology prediction server, Bioinformatics 17, 849–850.

86 Rost, B., Casadio, R., Fariselli, P., Sander, C. (1995) Transmembrane helices predicted at 95 % accuracy, Protein Sci. 4, 521–33.

87 Chen, C.P., Kernytsky, A., Rost, B. (2002) Transmembrane helix predictions revisited. Protein Sci. 11, 2774–2791.

88 Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305, 567–580.

89 Wilkinson, D.L., Harrison, R.G. (1991) Predicting the solubility of recombinant proteins in Escherichia coli, Biotechnology 9, 443–448.

90 Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T., Gerstein, M. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics, Nucleic Acids Res. 29, 2884–2898.

91 Rechsteiner, M., Rogers, S.W. (1996) PEST sequences and regulation by proteolysis, Trends. Biochem. Sci. 21, 267–271.

92 Bachmair, A., Finley, D., Varshavsky, A. (1986) In vivo half-life of a protein is a function of its amino-terminal residue, Science 234, 179–186.

93 Fisher, H.F. (1964) A limiting law relating the size and shape of protein molecules to their composition, Proc. Natl. Acad. Sci. 51, 1285–1291.

94 Rost, B., Liu, J. (2003) The PredictProtein server, Nucleic Acids Res. 31, 3300–3304.

95 Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., Bryant, S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 30, 281–283.

96 Corpet, F., Servant, F., Gouzy, J., Kahn, D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons, Nucleic Acids Res. 28, 267–269.

97 Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., Kahn, D. (2002) ProDom: automated clustering of homologous domains, Brief Bioinform. 2002 3, 246–251.

98 Lupas, A. (1996) Prediction and analysis of coiled-coil structures, Methods Enzymol. 266, 513–525.

99 Deleage, G., Blanchet, C., Geourjon, C. (1997) Protein structure prediction. Implications for the biologist, Biochimie 79, 681–686.

100 Edwards, Y.J., Cottage, A. (2003) Bioinformatics methods to predict protein structure and function. A practical approach. Mol. Biotechnol. 23, 139–166.

101 Rost, B., Schneider, R., Sander, C. (1997) Protein fold recognition by prediction-based threading, J. Mol. Biol. 270, 471–480.

102 Chou, P.Y., Fasman, G.D. (1974) Prediction of protein conformation. Biochemistry, 13, 222–245.

103 Garnier, J., Ogusthorpe, D.J., Robson, B. (1978) Analysis of the accuracy and implementation of simple methods for predicting the secondary structure of globular proteins, J. Mol. Biol., 120, 97–120.

104 Lim, V.I. (1974) Algorithms for prediction of helices and beta-structural regions in globular proteins, J. Mol. Biol. 88, 873–894.

105 Levin, J.M., Robson, B., Garnier, J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity, FEBS Lett. 205, 303–308.

106 Qian, N., Sejnowski, T.J. (1988) Predicting the secondary structure of globular proteins using neural network models, J. Mol. Biol. 202, 865–884.

107 Rost, B. (1996) PHD: Predicting one-dimensional protein structure by profile-based neural networks, Methods Enzymol. 266, 525–539.

108 Levin, J.M. (1997) Exploring the limits of nearest neighbour secondary structure prediction, Protein Engineering, 10, 771–776.

109 Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292, 195–202.

110 Godzik, A. (2003) Fold recognition methods. Methods Biochem. Anal. 44, 525–546.

111 Madej, T., Boguski, M.S., Bryant, S.H. (1995) Threading analysis suggests that the obese gene product may be a helical cytokine, FEBS Lett. 373, 13–18.

112 Novotny, J., Bruccoleri, R., Karplus, M. (1984) An analysis of incorrectly folded protein models. Implications for structure predictions, J. Mol. Biol. 177, 787–818.
113 Jones, D.T., Taylor, W.R., Thornton, J.M. (1992) A new approach to protein fold recognition, Nature 358, 86–89.
114 Sippl, M.J., Weitckus, S. (1992) Detection of native-like models for amino acid sequences of unknown 3D structure, Proteins 13, 258–271.
115 Bryant, S.H., Lawrence, C.E. (1993) An empirical energy function for threading a protein sequence through a folding motif, Proteins 5, 92–112.
116 Bowie, J.U., Luthy, R., Eisenberg, D. (1991) A method to identify protein sequences that fold into a known 3-dimensional structure, Science 253, 164–170.
117 Sheridan, R.P., Dixon, J.S., Venkataraghavan, R. (1985) Generating plausible protein folds by secondary structure similarity, Int. J. Pept. Protein Res. 25, 132–143.
118 Fischer, D., Eisenberg, D. (1996) Protein fold recognition using sequence-derived predictions, Protein Sci. 5, 947–955.
119 Russel, R.B., Copely, R.R., Barton, G.J. (1996) Protein fold recognition by mapping predicted secondary structures, J. Mol. Biol. 259, 349–365.

15
Applied Bioinformatics for Drug Discovery and Development

Jian Chen, ShuJian Wu, and Daniel B. Davison

15.1
Introduction

Bioinformatics is an evolving research field that incorporates disciplines from biology, mathematics, and computer science. Biology is the central pillar in bioinformatics – the audience and users of bioinformatics are experimental or theoretical biologists. The data resources for bioinformatics are biological experimental data. The output data from bioinformatics are meant to help solve biological problems. Mathematics and statistics are required to develop sophisticated algorithms to model biological data and to deduce hypotheses from biological data. Computer science is an essential tool in bioinformatics. Efficient software programs have been developed to implement the mathematical algorithms. Databases have been established to store and query vast amounts of biological data. High-power supercomputers are used to execute software programs and database queries to generate biological hypothesis. In recent years there has been an expansion of organism genome-sequencing projects in both academic and industrial environments. With progress in genome sequencing, biological studies on those organisms are also expanding enormously. Because of the capacity of bioinformatics to manage a vast amount of data from different biological disciplines and to perform data mining to generate biological hypothesis, it has been increasingly applied in research and development in the pharmaceutical industry.

In this review we focus on biological aspects of bioinformatics. After a descriptive summary of the different biological databases we investigate how bioinformatics can be applied to these different biological databases to facilitate target identification and drug development in the pharmaceutical industry.

15.2
Databases

Completion of the human genome project has resulted in an explosion of sequence-based databases. The public domain Human Genome Project and the private sector Celera have both published their versions of the complete human genome. To support gene findings and functional study of the genome project, expression sequence tagged (EST) sequences have been generated in a high-throughput manner with the new focus of obtaining full-length cDNA sequences of all

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

genes in the genome. Comparison of genomes from different individuals reveals sequence variations and single nucleotide polymorphism. Completion of the human genome project enables systematic study of gene expression in body tissues and disease states. Gene expression and functional study link genes to biological pathways, shifting the focus of the next generation of databases from gene centric to pathway centric. Proteins are the functional units of biological pathways. Sequencing of the human genome and recent advance in protein analysis tools have greatly advanced the field of proteomics. Study of metabolic pathways and their by-products gives rise to metabonomic databases. The ultimate goal of screening and developing small compounds against disease targets in the pharmaceutical industry requires the field of cheminformatics to manage chemical compounds and to explore the structural relationship between proteins and compounds.

The expansion of databases in both number and content provides both excitement and challenge to the scientific community. On the one hand scientists can have access to databases of different disciplines; on the other hand effective use of the different databases requires sophisticated data manipulation. In this section we provide brief summaries of these different databases and show how bioinformatics could be used to perform data mining on these databases. At the end of the chapter we will discuss the use of bioinformatics to integrate different types of data and to study biological systems.

15.2.1 Sequence Databases

15.2.1.1 Genomic Sequence Databases

The completion of human genome project marked a significant milestone in molecular biology [1, 2]. It opens the possibility of identifying genes and the mechanisms responsible for all human diseases. Besides the human genome, those of other species, for example *Drosophila melanogaster* [3], *Rattus norvegicus* [4], *Mus musculus* [5], *Caenorhabditis elegans* [6] have been sequenced. Although completion of the human genome was celebrated with much fanfare, the importance of completing the genome sequencing and analysis of other model organisms, and common pathogens, should not be overlooked. Model organisms like the mouse, rat, and dog are important disease models upon which we rely to develop disease and treatment models. Genome sequencing of pathogens such as *Staphylococcus aureus*, severe acute respiratory syndrome (SARS) Coronavirus [7] enables us to develop drugs and vaccines to treat pathogen-born diseases. Furthermore, comparison of the genomes of different species would facilitate the process of identifying genes and their functions. Some important facts about genome size and estimated gene number are listed in Tab. 15.1 for selected organisms.

Published genome sequences can be found in GenBank in the US [12], in the EMBL Data Library in the UK [13], and in the DNA Data Bank of Japan (DDBJ) [14]. These are regarded as the most comprehensive public databases for nucleotide sequences and supporting bibliographic and biological annotations. All three databases contain publicly available DNA and protein sequences, obtained mostly from individual laboratories and large-scale sequencing facilities. The three databases exchange information daily to ensure their content is up-to-date. Another good resource for genome sequences is the Institute for Genome Research (TIGR) [15], which performs its own genome sequencing and analysis.

There are many online bioinformatic resources enabling genome data to be viewed

Table 15.1. Genome size and estimated gene number for selected organisms.

Organism	*Genome size (bases)*	*Estimated gene number*	*Chromosome number*
Human (*Homo sapiens*)	3310 million	~30,000	46
Mouse (*Mus musculus*)	2695 million	~30,000	46
Fruit fly (*Drosophila melanogaster*)	180 million	~23,600	8
Roundworm (*Caenorhabditis elegans*)	100 million	~19,888	6
Baking yeast (*Saccharomyces cerevisiae*)	12 million	~6419	16
Bacteria (*Escherichia coli*)	4.7 million	~3200	1

Resources for contents of Tab. 15.1: Human [1]; mouse [5]; fruit fly [8]; roundworm [9]; baking yeast [10]; bacteria [11]

and comparative genome analysis to be conducted. EBI's ENSEMBL and UCSC's Genome Browser [16] are two outstanding websites for these purposes. Both provide a graphic view of the assembled genomes to facilitate bioinformatic data analysis and genome comparison. Users can upload their own DNA sequences and have them mapped on to the genomes. Genome synteny study can be conducted with ease because both web sites already map the syntenic regions of selected genomes.

15.2.1.2

EST Sequence Databases

Although the genome represents the entire gene collection in an organism, the existence of an expression sequence tag (EST) provides the first evidence of the functionality of a gene. Before the sequencing of an organism's genome, EST sequences are an important resource for gene discovery and identification. In fact, the human genome project spent its nascent days in EST sequencing and data mining. In the postgenome era, EST continue to play important role in the discovery of novel genes, alternative splicing variants, and single-nucleotide polymorphism. Over the year of 2003 the number of EST in the GenBank sequence collection increased by over 45 % to a total of 18.1 million sequences. Those sequences represent over 580 different organisms, with *H. sapiens* (5.4 million records), *M. musculus* (3.8 million records), *R. norvegicus* (540,000 records) holding the top three positions [12]. There are several key contributors to the public EST sequence collection. The IMAGE (the integrated molecular analysis of genomes and their expression) Consortium shares their high-quality arrayed cDNA libraries and places sequence, map, and expression data on the clones in these arrays in the public domain [17]. The Cancer Genome Anatomy Project (CGAP) generates a large amount of EST sequences in its effort to determine the gene-expression profiles of normal, precancerous, and cancer cells [18]. To facilitate full length cDNA cloning and gene functional study, the Mammalian Gene Collection (MGC) [18, 19] and the RIKEN Genomic Science Center [20] produce and release to the public domain high-quality full-length or near full-length cDNA clones and sequences.

There are several ways of deciphering EST sequence information using bioinformatic tools. GenBank further processes the EST sequences by BLAST search and incorporates their homology information into a companion database called dbEST [21]. Based on the information in dbEST, EST

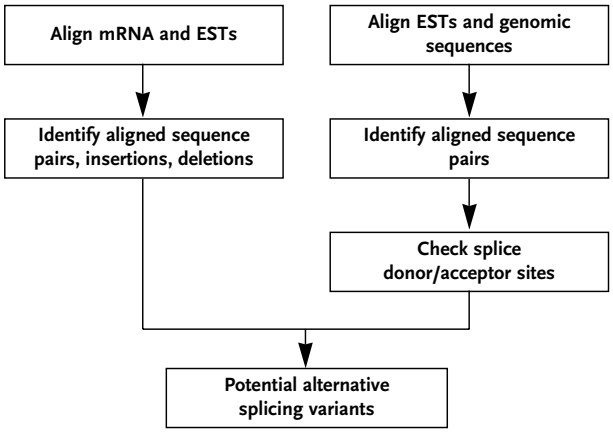


Fig. 15.1 Identifying gene alternative splicing variants.

sequences are grouped into gene-oriented sequence clusters in the UniGene database [22]. Commercial software based on the d^2 algorithm [23] can be used for clustering EST and extracting consensus sequences. At the ENSEMBL and UCSC genome browser web sites, EST sequences can be matched to genomic DNA sequences to facilitate identification of gene exons and to support the novel gene discovery process.

Another important use of EST sequence information is the discovery of alternative splicing variants of genes. Alternative splicing is a widespread phenomenon in mammalian gene expression. It is estimated that at least 59 % or more of human genes have multiple distinct transcripts [1], and the number of variants produced by each gene varies substantially [24]. The alternative splicing patterns are often specific to different stages of development, particular tissues or disease states. Changes in the pattern of alternative splicing of a gene are increasingly regarded as causes of some diseases [25]. A bioinformatic approach to the identification of potential alternative splicing variants is shown in Fig. 15.1. High-scoring EST are aligned to mRNA, taking into account exon–intron boundary information from genome analysis (like ENSEMBL). Each aligned sequence pair is checked for insertion or deletion in the EST sequence, resulting in potential alternative splice variants. Alternative splicing sequence databases are publicly available. EASED (extended alternatively spliced EST database [26]) and ASD (the alternative splicing database [24]) are two good examples. It is important to keep in mind that alternative splicing variants should be validated experimentally by PCR or cDNA cloning methods, followed by protein expression and functional study.

15.2.1.3 Sequence Variations and Polymorphism Databases

Comparative study of genome sequences from different individuals shows there are widespread small genetic changes, especially single-base differences, called single nucleotide polymorphisms (SNP). On average, SNP occur in the human population more than one percent of the time. Some of the SNP occur in the coding region of a gene and potentially affect the protein sequence of the encoding gene [27]. Other SNP occur in the regulation or structural elements of the gene, potentially affecting its transcription regulation [28].

SNP and other DNA polymorphisms can have three major effects on human diseases. Although only a very small number of human diseases are caused by genetic variations within a single gene, the occurrence of DNA polymorphism contributes to a person's genetic predisposition to developing a disease. For example, SNP in the CAPN10 gene locus have been found to be strongly associated with the onset of type II diabetes [29]. The occurrence of SNP and DNA polymorphism can also confer susceptibility or resistance to a disease and determine its severity or progression. SNP and DNA polymorphism can also determine how an individual responds to drug therapy, from drug absorption to drug metabolism and adverse drug effects. For example, polymorphism in cytochrome CYP2D6 protein is known to affect the disposition and anti-tussive effect of dextromethorphan in humans [30]. It is also associated with the efficacy and tolerance of the cholesterol-lowering drug simvastatin [31], and with nicotine metabolism in smokers [32]. Polymorphisms in other target genes, for example β2-adrenoceptor targeted by albuterol and CETP (cholesterol ester transfer protein) targeted by pravastatin have been shown to have a profound effect on an individual's drug response [33, 34].

The dbSNP from the National Center for Biotechnology Information (NCBI) is the largest public SNP database. Records in the dbSNP are cross-referenced to other information resources, for example PubMed, LocusLink, and GenBank. HGVbase is another curated resource describing human DNA variation and phenotype relationships [35]. Besides SNP information, HGVbase also contains current progress in the human heliotype mapping project [36], providing a foundation for combining genotypes and haplotypes into a functional unit for study of the correlation between sequence variations and phenotypes. In recent years SNP have become important tools in genetic association studies to identify genes affecting susceptibility to diseases.

Non-synonymous SNP are implicated in human disease processes because they have the potential to affect a protein's sequence, structure, and function. With genome sequencing and high-throughput prediction of protein 3D structure, structure assignment information for genes within complete genomes are becoming available [37–39]. TopoSNP is a database that combines non-synonymous SNP with 3D structures to facilitate understanding of SNP in disease processes [40]. The assignment of SNP on to 3D protein structures greatly facilitates the study of protein structure–function relationships.

15.2.2 Expression Databases

With completion of the sequencing of the human genome and that of other organisms, and identification of genes inside each genome, the immediate next challenge is to discover where and when these genes are expressed. The expression of genes is precisely controlled spatially and temporary. In addition, as already discussed, different splice variants of a gene could be expressed in different tissues and development stages. Changes in gene-expression profile could be the cause or result of the development of a disease. Gene-expression profile changes could thus be used to predict and monitor disease progress.

15.2.2.1 Microarray and Gene Chip

Several high-throughput methods can be used to define the transcriptome (expression profiles of all genes in an organism). Affymetrix short oligonucleotide-based DNA chip,

cDNA probe-based glass slide DNA microarray, quantitative PCR (Taqman PCR), and serial analysis of gene expression (SAGE) are the most commonly used methods. Advancement of technology in all these methods has enabled them to contribute significantly to the study of gene-expression profiles.

The National Cancer Institute (NCI) sponsors the Cancer Genome Anatomy Project (CGAP), which aims to determine the gene-expression profiles of normal, pre-cancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment of the patient. It initiated the Human Transcriptome Project (HTP) in 2001 to generate a complete collection of transcribed elements of the human genome. Public Expression Profiling Resource (PEPR) is another publicly available resource for human, mouse, and rat Affymetrix GeneChip expression profiles [41]. Through bioinformatic methods, the Gene Expression Database (GXD) [42] integrates with other mouse genome informatics (MGI) databases, placing the gene-expression information in the context of mouse genetic, genomic, and phenotypic information. Integration of database information enables scientists to gain insight into the molecular aspects of tissue and disease development.

15.2.2.2

Others (SAGE, Differential Display)

Serial analysis of gene expression (SAGE) is another well-established technique for gene-expression profiling [43]. It uses short oligonucleotide tags to define expressed mRNA from target genes. Ligation of the tags into long concatemers and their sequencing result in the qualitative and quantitative gene-expression profile of a particular tissue. The serial analysis of gene expression (SAGE) web site enables users to query expression profile information from mouse tissues and cell lines [44].

15.2.2.3

Quantitative PCR

Transcription profiling results from high-throughput profiling analysis must usually be verified by quantitative PCR methods. Taqman quantitative PCR is a reliable method for independent study of the expression of genes. Gene-specific primer design is a key step in obtaining reliable result from quantitative PCR study. PrimerBank [45] is an online database that contains approximately 180,000 PCR primers for approximately 40,000 genes from the human and mouse. The primers are designed with vigorous statistical algorithms and have been used for successful profiling of gene expression.

Bioinformatics provides the tools for performing database integration of expression profiling data from EST databases, DNA chip and microarray databases, and Taqman quantitative PCR experiment data. This integration process could be used to produce a gene-expression body map which provides important insight into the biological function of a gene and its involvement in disease development.

15.2.3

Pathway Databases

Proteins function by participating in one or more biological pathways. Traditionally, elucidation of biological pathways was performed by classic biochemical and genetic studies. Completion of the sequencing of the genomes of numerous organisms makes it possible to construct biological pathways by use of computational methods. Transcription profiling studies also provide powerful assistance to pathway construction, because genes in the same pathway tend to have coordinated expression changes. Pathway databases and their associated software are becoming increasingly critical in under-

standing and simulating the higher-order biological activity of a cell or an organism.

Two of the major pathway databases in the public domain are BioCYC and KEGG (Kyoto Encyclopedia of Genes and Genomes). Both databases provide electronic references to pathways and genomes, enabling scientists to visualize from gene to biochemical reaction to complete biochemical pathways. Two of the major components of BioCYC are MetaCYC and EcoCYC, carefully curated databases with contents derived from biomedical research. EcoCYC [46] contains data from the *E. coli* K12 strain microorganism. MetaCYC [47] is a nonredundant metabolic pathway database containing pathways from many organisms. The KEGG database incorporates genome projects, pathway databases, biochemical compounds, and reactions to enable computational prediction of higher-level complexity of cellular processes and organism behavior [48].

The rise and expansion of pathway databases signal a critical shift in the bioinformatic database paradigm. Proteins are no longer viewed as a simple collection of individual gene products. Instead, they are organized in pathways and networks representing their biological roles. The function of a gene is no longer just a description of the activity a single enzyme. A collection of information is required to describe the function of a the gene – when and where the gene is expressed, the other gene products it interacts with, which pathways it belongs to, how changes in gene sequence and expression affect the multiple pathways its product belongs to, and, ultimately, the physiology of the human body.

15.2.4 Cheminformatics

Chemical compounds have been the holy grail of pharmaceutical companies. Traditionally chemistry and genomics are two different approaches to a problem. With genomics expanding into protein structures, biochemical reactions, and pathways, the link between chemistry and genomics becomes increasing critical for information synergy in pharmaceutical companies. Cheminformatics is about cross-referencing, and about access to biological and chemical data.

A pioneer in cheminformatics is the National Cancer Institute's 60 cell-line anticancer drug screen study [49]. The study screened the ability of ~70,000 compounds to kill or inhibit the growth of a panel of 60 human tumor cell lines. Data mining analysis was performed using correlation analysis algorithms to study the sensitivity to the compounds of these cell lines and thousands of molecular targets [50]. Biological activity data and 3D-structure data of the compounds are interconnected in these analyses to identify potential novel anti-cancer compounds.

A good example of the application of cheminformatics in drug discovery and development is the SMART-IDEA (structure modeling and analysis research tool-integrated data for experimental analysis) project in the pharmaceutical company Bristol– Myers Squibb. SMART-IDEA has an innovative information management capability and enables research scientists to deposit, retrieve, and analyze chemical and biological data about chemical compounds. More importantly, it enables drug-discovery scientists to simultaneously test compounds in multiple experiments and then to use that multidimensional data to refine the characteristics of experimental drug compounds. The success of this cheminformatics platform earned it the 21st Century Achievement Award from ComputerWorld magazine in 2002.

15.2.5 Metabonomics and Proteomics

Metabonomics is an emerging discipline that systematically explores biofluid composition using NMR-pattern-recognition technology to associate target organ toxicity with NMR spectral patterns and identifies novel surrogate markers of toxicity. Metabonomic reports generated using NMR spectroscopy of urine and blood serum are increasingly available in the public domain for marketing drugs [51]. The Consortium for Metabonomic Toxicology (COMET) was recently formed by six pharmaceutical companies and Imperial College of Science, Technology, and Medicine, London, UK. The goal of COMET is to create a comprehensive metabonomic database containing NMR data for a wide range of model toxins and drugs, and to develop multivariate statistical models for prediction of the toxicity of candidate drugs [52]. In the private sector, start-up companies such as Metabometrix (http://www.metabometrix.com/) are developing new ways of using NMR technology to diagnose disease and to predict drug efficacy and toxicity.

Proteomics could be defined as a research field that involves the large-scale identification, characterization, and quantitation of proteins expressed in a cell line, tissue, or organism under given conditions. Proteomics is a powerful approach that integrates recent technological advances in high-throughput protein separation, mass spectrometry (MS), genomic database, and bioinformatics to address important physiological and medical questions. There are several proteomics databases in the public domain. For example, OPD maintained by the University of Texas (http://bioinformatics.icmb.utexas.edu/OPD/), is a public database for storing and disseminating mass spectrometry-based proteomics data. The database currently contains ~400,000 spectra from experiments on four different organisms – *Homo sapiens, Escherichia coli, Saccharomyces cerevisiae*, and *Mycobacterium smegmatis*. Proteomics is playing a key role in novel drug-target discovery, biomarker identification, toxicogenomic study, and many others pharmaceutical processes [53].

15.2.6 Database Integration and Systems Biology

As discussed above, with completion of the human genome project vast amounts of biological research data, on the scale of the genome, are being generated, covering all aspects of biology – sequence, expression, proteins, pathways, and metabolism. Expansion of biological data on this gargantuan scale leads to a proliferation of "omics". For example, the now "classic" genomics includes data from DNA sequencing, cDNA cloning, expression profiling, gene knock outs, and RNA interference. Proteomics consists of protein characterization and biomarker identification. Metabonomics includes the identification and validation of biomarkers using NMR, proteomics and other techniques. Data from these and other "omics" are used to build signal-transduction pathways and metabolism pathways. Systems biology is defined as the process of using the different "omics" to construct and analyze biological pathways to provide a high-level view of the effects of biological molecules and chemical compounds on cells, tissues, organs and whole organisms. Bioinformatics data management has been shifting from gene-centric to pathway-centric. The integration of chemical information provides a new challenge to bioinformatics. To make sense to research scientists different kinds of data from different data resources must be integrated. Many of the database resources discussed

above are beginning to implement data integration and collection to enable users to cross-reference data. In addition, data mining software is being provided to perform in-depth cross-database analysis. One example is the EnsMart system from the European Bioinformatics Institute, which provides a generic data storage for rapid and flexible querying of large sets of biological data and integration with third-party data and tools [54]. Users can group and refine biological data according to many criteria, including cross-species analyses (genome synteny), disease links, sequence variations (SNP and alternative splicing), and expression patterns.

One immediate challenge to data integration and systems biology is to identify a common denominator among different kinds of data. The development of ontologies as annotation standards provides a means of interpreting domains of knowledge and facilitates cross-discipline communication. The Gene Ontology (GO) project is a collaborative effort to address two aspects of information integration – providing consistent descriptors for gene products in different databases and standardizing classifications for sequences and sequence features [55]. Gene ontology has been used for data integration in many major data repositories.

The ultimate goal of applied bioinformatics is to integrate information and data sources from a variety of experiments, both actual and computational. Doing so helps to uncover non-obvious relationships between genes and to cross-validate information from separate sources. In addition, by looking at the union of multiple sources one can cover larger parts of the genome [56]. Expansion of data integration enables scientists to hypothesize and examine the topology of biological networks, leading to a systematic approach to the study of biology [57].

On the other hand, errors in some measurements could be compounded when combined with errors from other technologies. For example, block effects in high-throughput and sensitive measurements, for example expression profiling and metabonomics, are of particular concern. Vigorous statistics in experiment design and data analysis could help reduce some of these errors.

Two approaches are used when performing data integration and systems biology study. The global approach is to measure all genes and all proteins by use of high-throughput facilities with global capabilities. Strong computer infrastructure and global tools have been developed to acquire, integrate, analyze, and model different types of biological data. Large-scale data storage has been developed to perform data mining. Good examples of this approach are the ENSEMBL project in EBI [54] and the processes in the Institute for Systems Biology [58]. This kind of "university-style" approach is essential for advancing the state of the art of data integration and systems biology, but is of little relevance to industrial problems.

The focus approach (industrial-style) to data integration and systems biology is to perform high-throughput studies of a limited set of genes and proteins of therapeutic interest to companies. It is tightly coupled to the goal of identifying new chemical and biological entities for therapeutic purposes. High-throughput facilities with company-wide capabilities are established to generate data from DNA sequencing, expression profiling, proteomics, metabonomics, toxicogenomics, and pharmacogenetics. Powerful computer infrastructure is also established. Instead of internal development of global tools, commercial tools are acquired and used to perform data mining. In this model, massive data storage is not used. In-

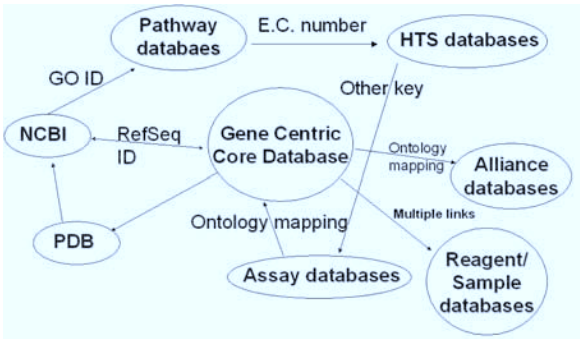


Fig. 15.2 Interconnectivity of databases using standard key identifiers.

stead, an internal core database management system is established to capture gene-centric information. Additional databases capture data generated from various assays and from high-throughput screening. The data are integrated, by use of standard key identifiers, to cross-reference databases. These standard key identifiers include the current NCBI reference sequence ID, locus link ID, gene ontology (GO), and chemical ontology (PROUS). Web interfaces and URL using Perl DBI modules are constructed to enable users to move between genomic, bioinformatic, toxicogenomic, cheminformatic, and other data banks. Figure 15.2 shows an example of how interconnectivity between databases is achieved by using standard key identifiers.

Besides managing and incorporating data from different resources, pharmaceutical bioinformaticians should not forget the task of using this information to facilitate drug discovery and development. In subsequent sections actual examples are discussed to show how these could be implemented.

15.3 Bioinformatics in Drug-target Discovery

Bioinformatics is playing a major role pharmaceutical drug-target discovery. From the pinnacle EST sequencing to the completion of genome projects, bioinformatic scientists have been using both sequence-based and structure-based methods to identify new genes from sequence databases and to computationally assign potential biological functions to these novel gene products. With advances in genome study in almost every discipline of biology, and the possibility of data integration from different resources, drug target discovery is shifting from single-source data mining to target identification as a result of data integration.

15.3.1 Target-class Approach to Drug-target Discovery

Over 80 % of current pharmaceutical drug targets belong to several classes of genes, for example G-protein coupled receptors (GPCR, 45 %), enzymes like kinases and proteases (28 %), nuclear hormone receptors (2 %), and ion channels (5 %) [59]. With completion of the sequencing of the genomes of human and other model organisms, one immediate question for pharmaceutical bioinformatics is: how many genes are there in each of these target classes? Could any of these be targeted to produce novel therapeutic reagents? Systematic querying of the genome for all gene members of a target gene class could be achieved by using bioinformatic approaches.

Sequence-based-homology searching is the classical approach to finding genes of a family. It is based on the observation that homology (evolutionary relationship) can be inferred from sequence similarity. There are several ways of performing sequence-based homology-searching. The Smith–Waterman algorithm [60] is the most sensitive search method, because of its mathematical completeness. Its drawback of being extremely time-consuming can be overcome by adaptation of the algorithm on specialized supercomputer chips supplied by companies like TimeLogic and Compugen. The heuristic search algorithms of BLAST (basic local alignment and search tool) [61] and FASTA [62] provide much faster but less sensitive searches. The position-specific-iterated BLAST (PSI-BLAST) algorithm [63] takes advantage of multiple sequence alignment from its first-pass search results to generate a sequence profile, and continues the similarity search using the profile to generate new homologous hits. This method greatly improves the sensitivity of BLAST searches.

It has been shown that profile-based similarity searches can reveal gene structure or function that would not be seen by simple pair-wise sequence alignment [64]. Profile-based hidden Markov model (PFAM HMM) methods take advantage of biological multiple sequence alignments and the statistical mathematics models of hidden Markov models to reveal statistically significant sequence signatures in biological sequences. The Pfam database [65] contains a collection of over 7000 HMM models covering proteins with both known and unknown biological function. Two popular profile HMM programs, HMMER and SAM, are widely used to build HMM models from multiple sequence alignment and to use the HMM models to search sequence databases for genes that contain these HMM domains [66].

Nucleotides and amino acids of a gene product only represent its linear one-dimensional information. The biological activity of a gene product results from its defined three-dimensional structural form. Both RNA and proteins can only function properly when correctly structurally folded. The functional portions of proteins, like the catalytic sites of proteases and kinases, have very well defined structure conformation. In molecular evolution, protein structures are often much more conserved than nucleotide or amino acid sequences. In addition, convergent evolution often leads to proteins sharing very similar three-dimensional structures despite different nucleotide or amino acid sequences, for example estrogen-interacting proteins [67]. Conversely, one could use the defined three-dimensional structure signature of, for example, the serine protease catalytic site, to perform a homology search against the whole genome structure model database [39] to identify serine proteases from the genome. A computational facility has been implemented in the Oak Ridge National Laboratory [68] for protein structure prediction and genome analysis.

Genes of a target class often contain one or more conserved signatures or domains. A combination of the above sequence, profile, and structure-based search methods tends to produce the most complete set of search results. A non-redundant distillation of the search results would simplify the characterization process. For example, 518 kinase genes in the human genome have been identified by using a combination of eukaryotic protein kinase HMM profiles and PSI-BLAST methods [69]. The application of bioinformatics in target-class gene discovery can thus quickly identify all family members of a target class to facilitate pharmaceutical target identification.

15.3.2 Disease-oriented Target Identification

In comparison with high throughput genome gene-discovery strategy, a new trend in target discovery is being developed to use database integration to perform disease-oriented target identification. Completion of the sequencing of the human genome and use of sophisticated bioinformatics tools in sequence, profile, and structure-based homology search have led to the rapid discovery of many novel genes. Novel gene discovery does not, however, translate linearly into novel drug target discovery. One of the major bottlenecks in drug-target discovery is the process of target validation, which still mostly relies on low-throughput benchtop *in-vitro* and *in-vivo* experiments. The disease-oriented approach to target identification has the advantage of using one set of target validation experiments to validate a smaller set of novel target candidates.

One key to successful disease-oriented target identification is to focus on disease-specific pathways. Another key is to integrate data from different resources. As already discussed, integration of data from different resources has the advantage of reducing noise and cross-validation of data points. Taking the intersection of multiple data resources can greatly reduce the number of genes that must be further validated by functional study. Let us use an example to illustrate how this strategy could be implemented.

Osteoporosis is major health problem that affects over 25 million people in the United States. This disease increases a patient's susceptibility to bone fracture and causes enormous human suffering, loss of productivity, death, and health-care costs in excess of $10 billion [70]. The disease is the result of loss of balanced action of bone-forming osteoblasts and bone-resorbing osteoclasts. Mature osteoclasts develop from hematopoietic stem cells via the monocyte-macrophage lineage whereas mature osteoblasts develop from mesenchymal stem cells. Treatment of osteoporosis thus entails of stimulation of osteoblast differentiation/function and suppression of osteoclast differentiation/function. Because of the very different lineages of osteoclasts and osteoblasts, attempts to discover the genes involved in osteoporosis should examine the two pathways separately. *In-vitro* cell-differentiation experiments could be set up to collect RNA samples from both progenitors and mature osteoclasts. DNA chip [71] and glass slide microarray experiments [72] have been performed to identify genes that are up- or down-regulated during osteoclast differentiation. To obtain osteoclast-specific genes, other cell types that could arise from the same development lineage, for example monocytes and macrophages, are used as controls to subtract non-osteoclast specific genes. A subset of osteoclast-specific genes whose expression levels are changed during osteoclast differentiation could be obtained by means of these experiments.

Human genetic studies have identified many loci in human chromosomes that are linked to bone diseases. For example, a human autosomal recessive osteoporosis gene has been mapped to chromosome 11q13 [73] and another osteoporosis disease, type II Albers–Schonberg disease, to chromosome 16p13.3 [74]. With the complete human genome one could quickly identify all the putative genes beneath these bone disease linked chromosomal regions. Intersection of this pool of genes with the genes identified by means of the transcription profiling experiments described above could further reduce the disease gene candidates. The CLCN7 chloride channel gene and the α subunit of apical vH^+-ATPase were identified by means of this data inte-

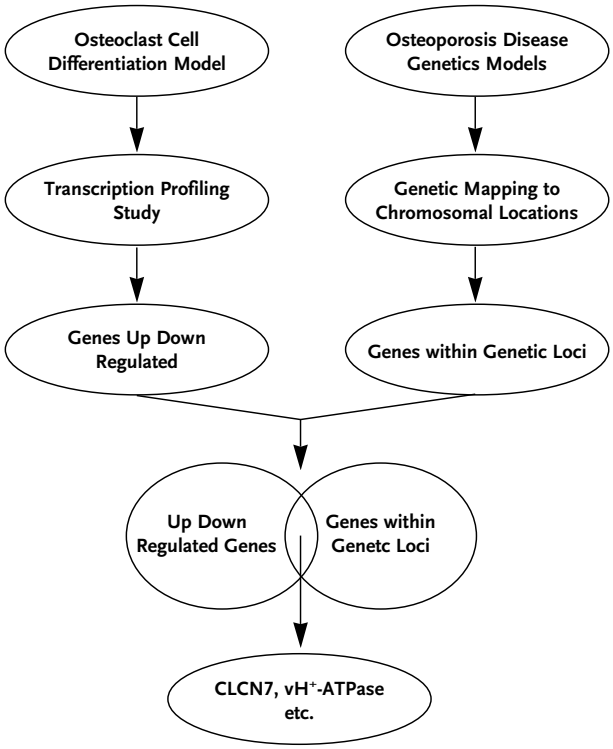


Fig. 15.3 A simple procedure for data integration to identify disease-related genes.

gration process. Figure 15.3 illustrates a simple procedure for this data-integration process.

15.3.3 Genetic Screening and Comparative Genomics in Model Organisms for Target Discovery

The combination of medicinal chemistry and model organism genetics has emerged as a powerful tool for discovery and validation of drug targets. Model systems can be used to perform mechanism of action (MOA) studies to identify target genes of pharmaceutical compounds that have proven *in-vivo* efficacy but unknown targets. Identification of the drug target genes and further study of their biological pathways enable scientists to explore additional intervention points to develop better drugs with potentially fewer adverse drug effects [75]. Genetic screening in model organisms could also identify genes or pathways that interfere with or rescue known disease-related genes or pathways, enabling the discovery of new targets for disease treatment. Reverse genetics is another approach used to identify specific classes of genes that modulate specific pathways key to disease development. Detailed knowledge of model organism genomes enables scientists to generate a collection of model organisms, for example gene knock-out mice or flies, and to screen for desired phenotypes that correspond to human diseases.

The biggest challenge facing the use of model organisms for target identification and validation is the relevance of model-organism phenotypic assays to human phys-

iology and diseases. Emphasizing this relevance is the extent of conservation of genes and biological pathways between the model organism and the human. Bioinformatics plays an important role in helping to determine the conservation of genes and pathways between organisms. Genome synteny study and phylogenetic analysis of multiple organisms helps establish gene orthologs. Gene ontology and pathway databases help evaluate the conservation between biological pathways from model organisms. A good example is the identification and functional study of drosophila p53 gene (Dmp53). In humans the p53 gene plays a critical role in the regulation of cell cycle and apoptosis [76] and has a central role in carcinogenesis [77]. A drosophila p53 gene was identified from the drosophila genome study and found to share most of the functional characterization of its human counterpart [78]. The conservation of p53 gene and pathway between drosophila and the human suggests that genetic screening in drosophila could be used to identify drosophila genes that could modulate the p53 pathway [79]. Human orthologs of these p53-impacting genes could be identified by phylogenetic analysis and evaluated for their function in human p53 pathway (Fig. 15.4).

Both genome-scale target-gene discovery, by use of both target-class and disease-oriented approaches, and genetic screening, generate many hypothetical disease-related target genes, for example the adiponectin receptor for obesity and diabetes [80], the low-density lipoprotein receptor for atherosclerosis [81], and interleukin-4 (IL-4) for allergic diseases [82]. Target genes with hypothetical disease association cannot, however, be used as appropriate intervention points for new drug candidates. It should be borne in mind that a series of validation processes is needed to establish the clear role of the target genes in the phenotype of a disease.

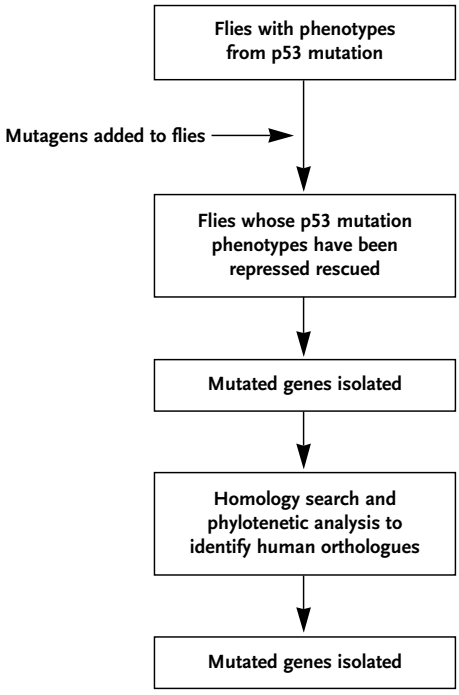


Fig. 15.4 Genetic screening in model organism to identify disease-related genes.

15.4 Support of Compound Screening and Toxicogenomics

In pharmaceutical drug development adverse drug effects arise from two major sources. Some are mechanism-based adverse drug effects, in which perturbation of selected genes or a pathway leads to deleterious effect on the human body. Selection of different intervention points or biological pathways provides a solution to this issue. Other adverse drug effects are non-mechanism based; most of these arise from the cross reactivity of drug compounds with critical biomolecules other than the target gene product. The solution is to develop a series of compound selectivity assays to ensure the specificity of the compound on the target gene.

15.4.1
Improving Compound Selectivity

15.4.1.1
Phylogeny Analysis

Selection of drug compounds with the desired specificity against drug targets is a great challenge to the industry. From a purist's point of view the more specific the selection assays the better. In a single living cell there can be as many as 200–300 kinases regulating almost all biological pathways. Limitations on resources and time do not, however, allow this kind of purist approach. For example, with over 500 kinases in the human genome, it is almost impossible to counterscreen kinase inhibitor compounds against all the kinases in their kinase assays. A bioinformatics approach could be applied here to help guide selection of a representative panel of counterscreen candidates to achieve cost–benefit balance.

As already mentioned, a combination of the profile HMM and PSI-BLAST search methods identifies all 518 kinases in the human genome [69]. Phylogenetic analysis classifies these kinases into 12 major groups, 134 families, and 196 subfamilies, primarily on the basis of sequence comparison of their catalytic domains. Experience has shown that compounds that inhibit kinases by competing with the substrate-binding site are often selective. Thus phylogenetic classification of the kinases provides a starting framework for selecting representative enzymes from each family for counterscreening. For example, a maximum parsimony tree of human "dual" kinases – those that phosphorylate serine, threonine, and tyrosine residues – is shown, in Fig. 15.5. The tree immediate-

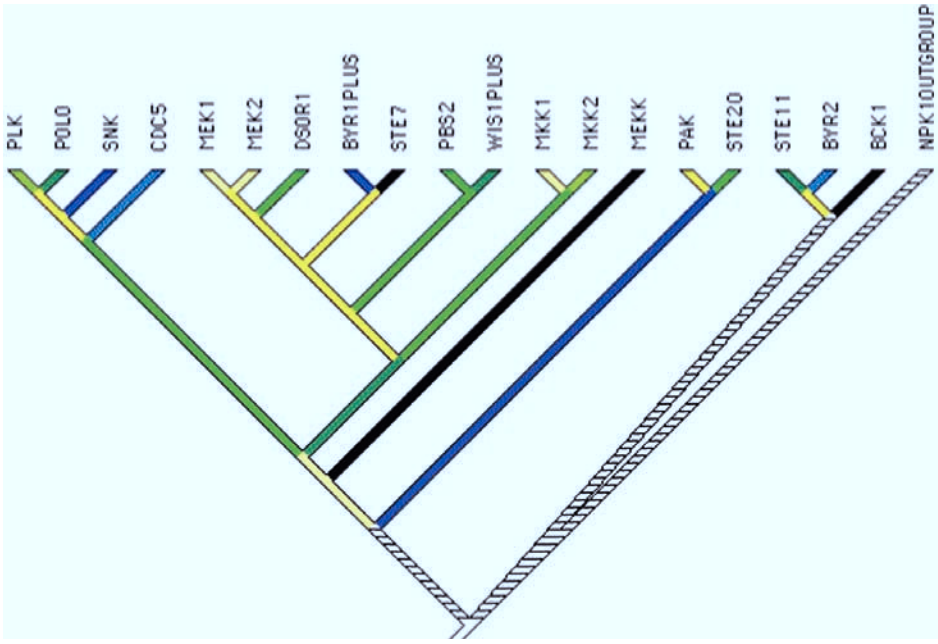


Fig. 15.5 Using phylogenetics to guide counterscreen selection. Screening MEK1 when the intended target is MEK2 might not result in the desired properties. In this tree of human kinases, MEK1 and MEK2 seem to be the result of recent duplication. Other organisms, especially model organisms, should be checked to determine how long ago the duplication actually occurred.

ly suggests that if MEK1 is a drug target, then counterscreening should be done against MEK2 to achieve the highest possible specificity for MEK1. Combined with tissue expression data, potential toxicological problems can be predicted and dealt with in early development, rather than encountering surprises in a clinical trial. Expanding the tree with model organisms can help to define the time of the MEK1–MEK2 duplication. Such analysis can also reveal whether or not the protein (or possibly a fragment of it) is still in the model organism. Such phylogenetic analyses can also be proxies for three-dimensional structure information in helping to establish the biological function of a gene needed for screening or counterscreening.

15.4.1.2

Tissue Expression and Biological Function Implication

To further refine the process of selecting representative genes for counterscreening, tissue expression profiling, and biological function implication data should be integrated with phylogenic data analysis. Drug-development experience tells us that most unmanageable adverse drug effects come from several key biological organs, for example the heart, kidneys, and liver. If a counterscreen gene candidate is expressed highly in key organs and tissues it could be placed at a higher priority on the candidate list. Bioinformatics can provide the tools to perform data integration of expression profiling data from EST databases, DNA chip

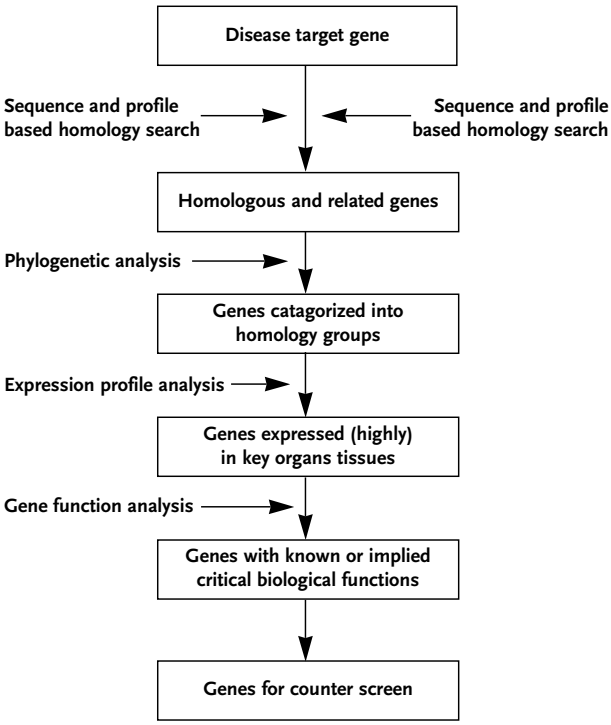


Fig. 15.6 Selection of counterscreening genes using genomic information

and microarray databases, and Taqman quantitative PCR experiment data. This data integration and data-mining process could be used to produce a gene-expression body map. Such a gene-expression body map would provide important insight into a gene's physiological function and facilitate the counterscreen gene-selection process. Figure 15.6 illustrates such a scheme for selecting appropriate counterscreening genes to achieve desired compound specificity.

Another key piece of information that should be integrated into the counterscreen gene selection process is the biological function implication of genes. For novel genes and for less characterized genes, their putative biological function could be implied via a data integration process. Strong homology with genes or domains of known biological function implies the novel gene could have similar biological activity. Co-regulation of a gene's expression profile with genes in a certain biological pathway suggests the gene might belong to or interact with that biological pathway. Phenotypes from genetic study in model organisms in which a gene has been deleted also suggest the biological function of the gene. Bioinformatic integration and interrogation of these data could contribute significantly to the process of counterscreen gene selection.

15.4.2 Prediction of Compound Toxicity

In pharmaceutical industries, a rational strategy for increasing the efficiency and reducing the cost of R&D is to reduce the rate of attrition in the costly late stages such as phases I, II, and III of clinical trials by increasing the rate of attrition in the less costly, early stages, for example the discovery and pre-clinical stages [83]. The main approach toward this involves early detection of the potential toxicity of lead compounds.

15.4.2.1 Toxicogenomics and Toxicity Signature

Traditional compound toxicity evaluation had to use multiple concentrations and time points in several lengthy animal studies, which was time-consuming and labor-intense. Today, development of novel approaches for high-throughput screening for compound toxicity is a major goal in the drug-development process. In the past few years the genomics approach for predictive toxicology is becoming a promising and mature technology [84] and a new discipline, "toxicogenomics", denoting the merging of toxicology with technology that has been developed, together with bioinformatics, to identify and quantify global gene expression changes, is becoming an active research field. It is a new aspect of drug development and risk assessment which promises to generate a wealth of information toward increased understanding of the molecular mechanisms that lead to drug toxicity and efficacy [85].

Gene expression profiling by means of DNA or protein chips has been shown to be an important means of rapidly identifying chemical toxicity issues [86]. More recently, databases containing profiles of compounds for which toxicological and pathological endpoints are well characterized have been developed by the US National Institute of Environmental Health Sciences (NIEHS, http://www.niehs.nih.gov/nct/) [87]. In the private sector several toxicogenomics companies are building data warehouses in which results from thousands of expression array and molecular pharmacology experiments, in which cells and organisms are systematically treated with drugs and drug-related compounds, are deposited and curated (Iconix Pharmaceuticals, http://www.iconixpharm.com/index.php; Gene Logic Inc., http://www.genelogic.com/).

The assumption of toxicogenomics is that cells or tissues exposed to drugs have specific patterns, called signatures, of expressed genes or proteins and that signatures generated by new drugs can be compared with signatures of drugs with well-known toxicity to predict potential toxicological issues with the new drugs [88]. Bioinformatics is playing a key role in data analysis and data mining to identify compound toxicity signature from these high-throughput studies. The use of gene-expression data from chip technologies for classification and predictive purposes has been widely demonstrated in many research fields [89, 90]. Many statistical algorithms tailored to large-volume data analysis have been developed and implemented. For example, linear discriminant analysis (LDA), genetic algorithm (GA), and K-nearest neighbors (KNN) have been used to predict compound safety signatures using DNA microarray data [86]. Principal-component analysis (PCA) has been used to examine the hepatotoxicity of bromobenzene [91].

15.4.2.2

Long QT Syndrome Assessment

Drug-induced long QT syndrome (LQTS), a disorder of cardiac rhythm in which sufferers, who can be otherwise healthy, are subject to a risk of arrhythmia or even sudden cardiac death, has been regarded as a critical side-effect of numerous drugs [92] and in the past decade the single most common cause of withdrawal or restriction of the use of drugs that have already been marketed has been the occurrence of drug-induced LQTS [93]. Today, QT interval prolongation is regarded as a major safety concern by worldwide drug-regulatory bodies, and early detection of new compounds with this undesirable side effect has become an important objective for pharmaceutical companies [94]. Any compounds associated with LQTS should ideally be identified at an early stage of drug discovery, e.g. the pre-clinical phase.

Although many drugs have been associated with LQTS, it appears that only a small subset of individuals is at increased risk of arrhythmia and that a significant component of this risk is genetically determined. Many studies conclude that the human hERG gene responsible for the normal action repolarization in the heart is associated mechanistically with the risk of LQTS for pharmaceutical agents from a wide variety of drug classes [95]. Population genetics and genomics studies found that many mutations in the hERG gene were identified in families suffering from drug-induced LQTS [96]. Moreover, many missense SNP were recently identified in the human hERG gene experimentally or computationally [97, 98]. Some of these missense SNP were found to significantly change the biological function of the channel [99–101]. Several computational algorithms have been used to predict the affinity of compounds for the hERG channel and hence to identify their potential cardiotoxicity. Some computational models have an observed-versus-predicted correlation $r^2 = 0.86$, which is very impressive [102, 103].

Besides numerous mutations and polymorphisms, extensive bioinformatic studies have also found that human hERG has at least three isoforms derived from alternative splice variants that result in different C- or N-termini of the protein. The expression pattern and function of these isoforms are different [104, 105]. Finally, a few reports discuss the expression regulation of human hERG; bioinformatics can play an important role in characterizing regulatory elements within the hERG-promoter region and in understanding how sequence variations with these elements affect the function of the gene and drug-induced LQTS.

LQTS can also be induced by the mutations or polymorphisms of other ion-channel genes, for example the KQT-like voltage-gated potassium channel-1 (KCNQ1), alpha polypeptide of voltage-gated sodium channel type V (SCN5A), and potassium voltage-gated channel subfamily E members 1 (KCNE1) and 2 (KCNE2). It should be noted that several dozen mutations have also been reported throughout the coding region of these genes in different ethnic groups, thus indicating remarkable genetic heterogeneity [106, 107]. Future bioinformatic studies in this field will be essential for more accurate prediction of compound cardiac liability issues in the early stage of drug discovery.

15.4.2.3

Drug Metabolism and Transport

The *in-vivo* response to a chemical compound or drug treatment is a very complex and highly dynamic process that involves many steps. As many as 50 proteins such as metabolizing enzymes and transporters participate in the pharmacodynamic response to drugs. Interestingly, many genes coding for such proteins contain polymorphisms that alter the activity or the level of expression of the encoded proteins [108]. Thus, response to a drug by a given individual reflects the interaction of multiple variable genetic factors that cause important variations in drug metabolism, distribution, and toxicity.

Metabolism in the liver is a major pathway for the elimination of drugs in the bile. Metabolizing enzymes in the liver can be classified into two groups, phase I and phase II enzymes [109]. Phase I metabolism is undertaken mostly by cytochrome P450 isoenzymes (CYP), the most important enzymes involved in the biotransformation of drugs and other xenobiotics. Approximately 80 different forms of P450 have been characterized in humans [110]. Among them, CYP2C9, CYP2D6, and CYP3A4 account for 60–70 % of all phase I metabolic biotransformation of drugs [111]. In the past decade many SNP, mutations, and splice variants in the coding region of these genes have been characterized in different ethnic groups, and some significantly affect the enzymatic activity of the proteins [112]. In collaboration with scientists in the field of protein-structure research, bioinformatics could be used for further investigation of how sequence variations affect the substrate-recognition sites of the enzymes and could, *in silico*, predict how sequence variations impact on enzymatic activity both qualitatively or quantitatively. Neural network computing and machine learning algorithms have been used to predict drug metabolism by CYP enzymes and enzyme isoforms [113].

Often, however, coding variants are insufficient to explain the large inter-individual variation of CYP enzyme activity. These observations have prompted many investigators to seek other causal polymorphisms including intronic mutations, splicing signature deletions/insertions, and exon skipping [114]. In recent years, several research groups have started to investigate regulation of the expression of these CYP genes. It has long been known that the expression level of some CYP genes ranges from 1 to 20-fold in different ethnic groups. This expression variation might be partly derived from the variability of regulatory regions within a gene promoter, and a novel polymorphic enhancer was identified recently in human CYP3A4 [115]. This polymorphism inserts TGT between –11,129 and –11,128 bp of the human CYP3A4 promoter and results in 36 % reduction of CYP3A4 enhancer activity.

With the tremendous progress made by the SNP Consortium (http://snp.cshl.org/)

and the completion of the human, mouse, and rat genomes, SNP data analysis and comparative genomics could be used to study the regulatory regions of these genes. The assumption behind this approach is that critical regulatory elements within a gene promoter might be conserved across species. For example, a novel enhancer of brain expression near the apoE gene cluster was recently identified by comparative genomics [116]. Future bioinformatic studies of sequence variations of these phase I enzymes will uncover more interesting findings and enable understanding of inter-individual variation of drug metabolism in the liver.

Besides huge variability of the phase I enzymes, there are many inter-individual variations of the phase II enzymes. These phase II enzymes, for example uridine diphosphate glucuronosyltransferase 1A1 (UGT1A1), NAT2, and thiopurine *S*-methyltransferase (TPMT), play key roles in conjugating metabolites from phase I to enhance their excretion in urine or bile. Mutations and polymorphisms of UGT1A1, NAT2, and TPMT enzymes have all been reported in different ethic groups [117]. These variations alter the enzymatic activity of these proteins in drug metabolism and even lead to severe toxicity of the drug in some patients.

Finally, many sequence variations of drug transporters such as organic anion-transporting polypeptides (OATP), organic anion transporters (OAT), organic cation transporters (OCT), and multidrug-resistance proteins (MDR) have been reported. These genetic polymorphisms might be one of the key determinants of inter-individual and interethnic variability in drug absorption, disposition, distribution, and excretion [118, 119]. Extensive bioinformatic research in this area is urgent needed in drug discovery.

15.5 Bioinformatics in Drug Development

Drug discovery and development is a risky, expensive and time-consuming process of about 10–12 years [120]. It is estimated that approximately 20 % of the research budget is spent in the drug-discovery phase and approximately 80 % in the drug-development phase. In the past decade, bioinformatics has played a pivotal role in novel drug-target discovery. Although many pharmaceuticals and biotechnology companies have numerous gene targets on which to perform their research, successfully moving research from drug discovery into drug development and, further, into the market is the ultimate goal for all drug companies. In the past few years bioinformatics has being applied to the different stages of drug development, including biomarker discovery, evaluation of drug efficacy, prediction of clinical adverse reactions, and even the life-cycle of drug management.

15.5.1 Biomarker Discovery

According to the FDA's definition, a biomarker is a characteristic, for example blood pressure, CD4 count, or enzymatic activity, that is objectively measured and evaluated as an indicator of normal biological, pathogenic, or pharmacological response to a therapeutic intervention [121]. At the genomic level, changes in the level of mRNA or protein expression are markers that could be used to monitor patients' responses to a drug.

Genomic and proteomic methods offer increased opportunities for biomarker identification and development in the different fields of drug research, especially antitumor research. Drug resistance in cancer therapeutics is a very common observation in clinical practice. Effective treatment for can-

cer requires not only the discovery of new drugs with antitumor activity, but also knowledge of the most accurate way of predicting patents' response to drugs. The recent development of DNA microarrays which enable simultaneous measurement of the expression levels of thousands of genes raises the possibility of statistically unbiased, genome-wide approach on the genetic basis of drug response. A recent study measured the expression levels of 6817 genes in a panel of 60 human cancer-cell lines (the NCI-60) and found that gene expression signatures obtained by a statistical supervised algorithm could be used to determine the chemosensitivity profiles of thousands of chemical compounds. The result suggested that, at least for a subset of compounds, genomic approaches to chemosensitivity prediction are feasible [122]. Bioinformatics is playing a key role in biomarker identification with regard to data integration and collection, data analysis, and mining of genomic and proteomic data [123, 124].

15.5.2 Genetic Variation and Drug Efficacy

Although drugs are designed and prescribed on a population basis, each patient is an individual. In principle, bioinformatics can be used to study drug efficacy. It has been reported that 30 % of patients taking statins, 35 % of patients taking beta-blockers, and up to 50 % of patients taking tricyclic antidepressants do not respond to pharmacological intervention [125]. It is estimated that 85 % of patients' response to drugs depends on genetic variations. "If it were not for the great variability among individuals medicine might as well be a science and not an art." The thoughts of Sir William Osler in 1892 reflect the view of medicine over the past 100 years [126]. Discovering variations in response is extremely important in medical practice. Genetic variants might be because of polymorphisms such as SNP, mutations, insertions, deletions, or alternative splice variants, of a drug target gene or drug metabolizing enzymes and drug transporters (The last of these has been discussed in Sect. 15.3.2). Knowing whether a patient will respond is important so that the best treatment can be given immediately and because of the high cost of some treatments and potential severe side effects of drugs.

Most drugs interact with specific target proteins to exert their pharmacological effects. These target proteins could be GPCR, transcription factors, or ion-channel proteins. Genetic variation of the drug-interactive site on a target protein might directly affect the drug's efficacy. For example, acute promyelocytic leukemia (APL) is characterized by a specific 15;17 chromosomal translocation which generates the PML/RAR chimeric gene [127]. Although all-*trans*-retinoic acid (RA) is a highly effective agent that induces complete remission in a high proportion of APL patients by inducing terminal differentiation of APL cells [128], RA resistance in APL treatment has recently been a serious clinical problem in differentiation-inducing therapy. A group of Japanese scientists found that a novel point mutation in the ligand-binding domain of the RAR portion (Arg611) of the chimeric PML/RAR gene led to RA resistance in APL patients by sequencing the PML/RAR chimeric gene in several APL cell lines established from RA-resistant patients. This mutation significantly reduces the sensitivity of tumor cells to RA [129].

Sequence variations in the regulatory region of a gene can indirectly affect a drug's efficacy. 5-Lipoxygenase (ALOX5) is an enzyme required for the production of both the cysteinyl leukotrienes and leukotriene

B4 [130]. ALOX5 activity therefore partially determines the level of bronchoconstrictor leukotrienes present in the airways. Pharmacological inhibition of ALOX5 activity has been used to control the symptoms of asthma in some patient populations but is not effective for every asthma patient [131]. Previous studies found that DNA sequence variants in the core promoter of ALOX5 are associated with diminished promoter-reporter activity in tissue culture; in particular, an Sp-1 binding motif (–GGGCGG–) located about 100 bp upstream from the ATG start site of ALOX5 is highly polymorphic, with three to six tandem repeats, five being the most common, identified in Caucasians and African Americans [132, 133]. It was found that five repeats of the Sp-1 binding site significantly reduced transcription of ALOX5 mRNA, and hence fewer leukotrienes were produced. Reduced clinical response to the ALOX5 antagonist ABT-761 was observed in patients harboring the mutant promoter [134].

Finally, cyclooxygenase isozyme COX-3, an alternative splice variant of COX-1, is highly expressed in canine cerebral cortex. This variant retains the first intron in its mRNA and hence introduces insertion of 30–34 amino acids; depending on the mammalian species it was found to be inhibited by acetaminophen and other analgesic/antipyretic drugs [135]. It is believed that inhibition of COX-3 might be a primary central mechanism by which these drugs reduce pain and possibly fever.

In the past few years bioinformatics has been widely applied in the field of pharmacogenomics, to enable understanding of the impacts of sequence variations on drug potency and efficacy.

15.5.3 Genetic Variation and Clinical Adverse Reactions

In numerous instances, the optimum dosage of clinically useful drugs is limited by variability in the way individuals respond to these drugs. Genetic variations play key roles in determining the optimum dose of a drug for an individual patient while avoiding adverse drug reactions (ADR) because of inappropriate dosage. For instance, the daily doses required to treat patients vary as much as twentyfold for the antithrombotic drug warfarin, and as much as fortyfold for the antihypertensive drug propanolol [136]. These genetic factors could be sequence polymorphisms of target genes, drug metabolizing enzymes, or transporters. The last two have already been mentioned.

The efficacious effect of warfarin on antithrombosis is substantially limited because of the risk of triggering hemorrhage. Genetic variants associated with the metabolism of warfarin by cytochrome P450 CYP2C9 have specific implications on untoward effects. It is found that for patients with CYP2C9*3 mutations an initial dose of 5 mg warfarin causes an increase in the international normalized ratio and significant risk of bleeding [137]. Recently, the long-sought warfarin target, vitamin K epoxide reductase complex subunit 1 (VKORC1), was identified and cloned by two independent groups [138, 139]. Mutations in this reductase were also found to cause warfarin resistance and multiple coagulation factor deficiency type 2. Carefully examining the genomic structure of VKORC1 on the NCBI human genome build 34 (August 2003, http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?c=locusid&org=9606&l=79001) found the gene maps on chromosome16p11.2. Although it covers only 4.87 kb, the gene con-

tains 13 confirmed introns, 13 of which are alternative. Comparison with the genome sequence shows that at least nine introns follow the consensual [gt–ag] rule. The sequence of this gene is supported by 394 sequences from 352 cDNA clones and produces, by alternative splicing, nine different transcripts altogether encoding seven different protein isoforms. Some of isoforms miss either the enzymatic domain or the transmembrane domains (Fig. 15.7). It would be interesting to see if any of these sequence variations could affect the interaction of the enzyme with warfarin and hence affect patients' sensitivity to warfarin dosage.

Another drug target variation that leads to ADR can be illustrated by the variation of the dopamine D3 receptor. This variation increases the sensitivity of the receptor to drugs and lead to ADR. It is well known that some schizophrenic patients treated with typical neuroleptic agents develop tardive dyskinesia (TD). Researchers found that substitution of serine with glycine in the dopamine D3 receptor leads to greater affinity for dopamine, and supersensitivity to dopamine is believed to cause TD [140].

In short, ADR could be caused by many genetic variations. From a bioinformatic aspect, detailed analysis of sequence variants of genes relevant to a drug's mechanism of action, metabolism, transport, and many other properties might provide a genetic clue in understanding ADR in medical practice.

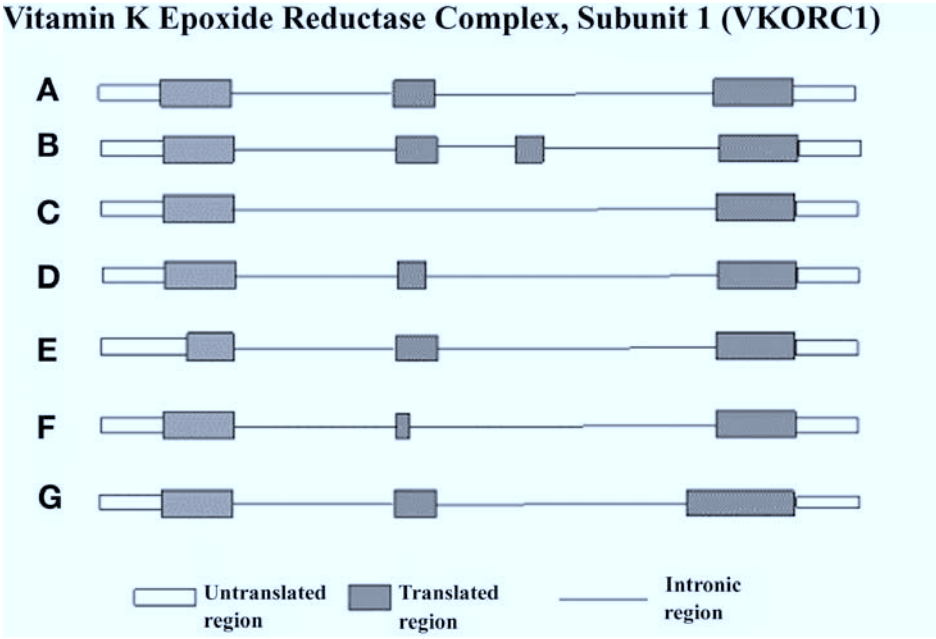


Fig. 15.7 Vitamin K epoxide reductase complex subunit 1 (VKORC1) is warfarin's target gene. According to NCBI human genome build 34 in August 2003, this gene could potentially have nine splice variants and seven of them affect its protein sequence. Variant A was recently identified and characterized functionally [138, 139]. Variants B–G are putative variants based on 394 EST sequences from 352 cDNA clones.

15.5.4 Bioinformatics in Drug Life-cycle Management (Personalized Drug and Drug Competitiveness)

In drug life-cycle management, post-marketing surveillance and pharmacovigilance are becoming very important. The ultimate goal of these activities is to prevent or to minimize the occurrence of ADR. Many ADR appear only after the marketed drug is used by large numbers of heterogeneous patients. Many drugs have been withdrawn in recent decades because of severe ADR. Although these ADR might not be observed during phase II and III clinical trials, when the drugs are marketed and used in a much larger patient population, problems with rare or long-term ADR begin to appear. For example, in 1997, manufacturers had to withdraw the fenfluramines from the market after study results linked their use with valvular heart disease in some obese patients [141]. It is imperative for drug makers to investigate the exact relationship between drug activity/toxicity and genetic polymorphisms so additional caution can be taken when a drug is administrated to "at- risk" patient groups.

Given the completeness of the human genome project and the technical availability of high-throughput sequence variation detection and gene-expression profiling, bioinformatics could use statistical algorithms to identify important genetic variations associated with or responsible for the ADR. For example, two studies have revealed an association between the HLAB*5701 polymorphism and a hypersensitivity reaction (HSR) to the anti-HIV reverse transcriptase inhibitor abacavir [142, 143]. This result can be used in clinical monitoring and management of hypersensitivity reactions among patients receiving abacavir. Development of databases of genetic profiles associated with ADR has recently been proposed [144]. This initiative will make genetics or genomics-related ADR more predictable if adequate statistical power is available. Incorporation of genomics data into a pharmacoepidemiology database was, in fact, also proposed three years ago as a means of identifying genetic risk factors and correlating them with the corresponding drug responses [145]. Bioinformatics is facilitating database integration of these two distinct research disciplines.

15.6 Conclusions

Although the origins of the discipline go back to the late 1960s [146, 147], the importance of bioinformatics became apparent in the 1990s as a direct result of the human genome project. In the past decade the roles of bioinformatics in the pharmaceutical industry have changed profoundly, from sequence analysis and gene identification in the early days to systemic database integration, data mining, prediction of compound toxicity, and evaluation of drug efficacy and adverse reactions currently. Figure 15.8 summarizes the impacts of bioinformatics on the different stages of drug discovery and development. The more we understand biology and medicine, the more important is the iterative and integrative study of biological systems as systems. Modern drug research certainly has to use a systemic approach to study the global effects of novel chemical compounds on the human body, to maximize the efficacy and minimize the side effects of a drug. It is a great opportunity and also a large challenge for bioinformatics to integrate and to analyze data from sequence-oriented, macrostructure-oriented, pathway-oriented, and patient-oriented studies and hence to facilitate drug discovery.

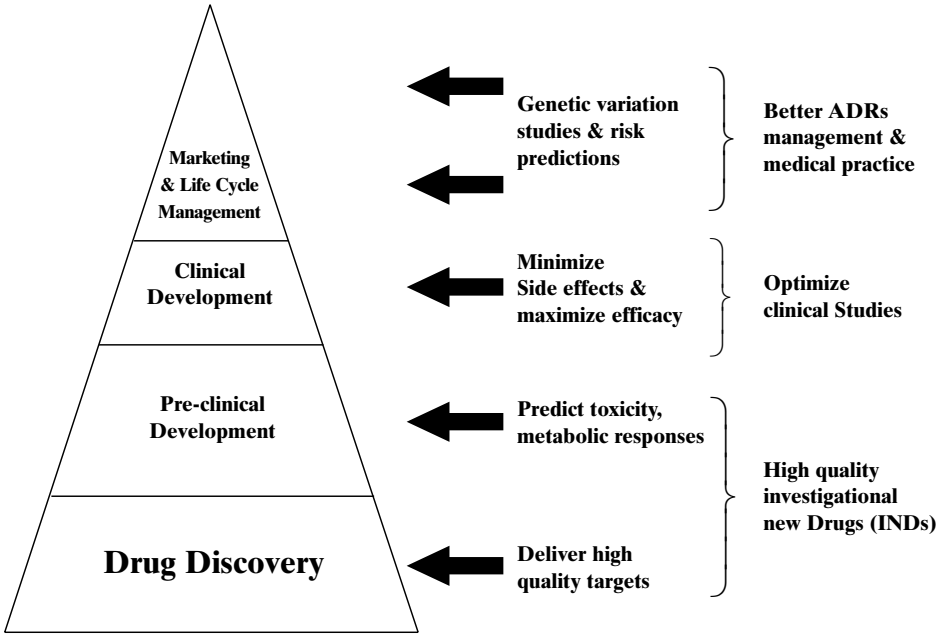


Fig. 15.8 Bioinformatics is being applied to the different stages of drug discovery and development in the pharmaceutical industry. By integrating genomic, genetic, transcriptomic, proteomic, and metabonomic studies with clinical data, bioinformatics is playing a key role in predicting drug response and adverse reactions in patient populations and in looking for new drugs with maximum therapeutic benefit and minimum side effects.

References

1 Lander EL, Birren LM, Nusbaum B, Zody C, Baldwin MC, Devon J, Dewar K, Doyle K, FitzHugh MW et al. (2001) Nature 409:860–921.
2 Venter JC, Adams MD et al. (2001).
3 Celniker SE, Rubin, G. M. (2003). Annu Rev Genomics Human Genet 4:89–117.
4 The Rat Genome Consortium (2004).
5 Waterston RH, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati

RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES; Mouse Genome Sequencing Consortium (2002) Nature 420:520–562.

6 Consortium, T. C. E. S. (1998). Science 282:2012–2018.

7 Zeng FY, Chan CWM et al. (2003). Exp Biol Med 228:866–873.

8 Celniker S, Wheeler D et al. (2002). Genome Biol (Research) 3:0079.1–0079.14.

9 Harris TW, Chen N et al. (2004) Nucl Acids Res 32:D411–D417.

10 Goffeau A, Barrell BG et al. (1996) Science 274:546–567.

11 Blattner FR, Plunkett G III et al. (1997) Science 277:1453–1462.

12 Benson DA, Karsch-Mizrachi I et al. (2004) Nucl Acids Res 32:D23–D26.

13 Stoesser G, Baker W et al. (2003) Nucl Acids Res 31:17–22.

14 Miyazaki S, Sugawara H et al. (2004) Nucl Acids Res 32:D31–D34.

15 Peterson JD, Umayam LA et al. (2001) Nucl Acids Res 29:123–125.

16 Karolchik, D, Baertsch, R et al. (2003) Nucl Acids Res 31:51–54.

17 Lennon G, Auffray C et al. (1996) Genomics 33:151–152.

18 Strausberg RL, Buetow KH et al. (2000) Trends Genet 16:103–106.

19 Mammalian Gene Collection Program Team*, Strausberg RL et al. (2002) PNAS 99:16899–16903.

20 Furuno M, Kasukawa T et al. (2003) Genome Res 13:1478–1487.

21 Boguski M, Lowe T et al. (1993) Nat Genet 4:332–333.

22 Wheeler DL, Church DM et al. (2003) Nucl Acids Res 31:28–33.

23 Burke J, Davison D et al. (1999) Genome Res 9:1135–1142.

24 Thanaraj TA, Stamm S et al. (2004) Nucl Acids Res 32:D64–D69.

25 Faustino NA, Cooper TA (2003) Genes Dev 17:419–437.

26 Pospisil H, Herrmann A et al. (2004) Nucl Acids Res 32:D70–D74.

27 Yoshida AH, Ikawa M (1994) Proc Natl Acad Sci USA 81:258–261.

28 Jaruzelska J, Abadie V et al. (1995) J Biol Chem 270:20370–20375.

29 Cox NJ, Hayes MG et al. (2004) Diabetes 53:S19–S25.

30 Capon DB, Kerry F, Mikus N, Danz G, Somogyi C (1996) Clin Pharmacol Ther 60:295–307.

31 Mulder AvL, Bon HJ, van den Bergh MA, Touw FA, Neef DJ, Vermes C (2001) Clin Pharmacol Ther 70:546–551.

32 Caporaso NL, Audrain C, Boyd J, Main NR, Issaq D, Utermahlan HJ, Falk B, Shields RT (2001) Cancer Epidemiol Biomarkers Prev 10:261–263.

33 Martinez FD, Graves PE et al. (1997) J Clin Invest 100:3184–3188.

34 Kuivenhoven JA, Jukema JW et al. (1998) N Engl J Med 338:86–93.

35 Fredman D, Munns G et al. (2004) Nucl Acids Res 32, D516–D519.

36 Gabriel SB, Schaffner SF et al. (2002) Science 296:2225–2229.

37 Gaasterland T, Sczyrba A et al. (2000) Genome Res 10:502–510.

38 Buchan DWA, Rison SCG et al. (2003) Nucl Acids Res 31:469–473.

39 McGuffin LJ, Street SA et al. (2004) Nucl Acids Res 32:D196–D199.

40 Stitziel NO, Binkowski TA et al. (2004) Nucl Acids Res 32:D520–D522.

41 Chen J, Zhao P et al. (2004) Nucl Acids Res 32:D578–D581.
42 Hill DP, Begley DA et al. (2004) Nucl Acids Res 32:D568–D571.
43 Velculescu VE, Zhang L et al. (1995) Science 270:484–487.
44 Divina P, Forejt J (2004) Nucl Acids Res 32:D482–D483.
45 Wang X, Seed B (2003) Nucl Acids Res 31:e154.
46 Karp PD, Riley M et al. (2002) Nucl Acids Res 30:56–58.
47 Krieger CJ, Zhang P et al. (2004) Nucl Acids Res 32:D438–D442.
48 Kanehisa M, Goto S et al. (2004) Nucl Acids Res 32:D277–D280.
49 Weinstein JN, Myers TG et al. (1997) Science 275:343–349.
50 Fang XS, Zhang L, Wang H (2004) J Chem Inf Comput Sci 44:249–257.
51 Mortishire-Smith RJ, Skiles GL, Lawrence JW, Spence S, Nicholls AW, Johnson BA, Nicholson JK (2004) Use of metabonomics to identify impaired fatty acid metabolism as the mechanism of a drug-induced toxicity. Chem Res Toxicol 17:165–173.
52 Lindon JC, Nicholson JK, Holmes E, Antti H, Bollard ME, Keun H, Beckonert O, Ebbels TM, Reily MD, Robertson D, Stevens GJ, Luke P, Breau AP, Cantor GH, Bible RH, Niederhauser U, Senn H, Schlotterbeck G, Sidelmann UG, Laursen SM, Tymiak A, Car BD, Lehman-McKeeman L, Colet JM, Loukaci A, Thomas C (2003) On temporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. Toxicol App1 Pharmacol 15:137–146.
53 He Q-Y, Chiu J-F (2003) Proteomics in biomarker discovery and drug development. J Cell Biochem 89:868–886.
54 Kasprzyk A, Keefe D et al. (2004) Genome Res. 14:160–169.
55 Gene Ontology Consortium (2004) Nucl Acids Res 32:D258–D261.
56 Gerstein M, Lan N et al. (2002) Science 295:284–287.
57 Ge H, Walhout AJM et al. (2003) Trends Genet 19:551–560.
58 Hood LGD (2003) Nature 421:444–448.
59 Drews J u r (2000) Science 287:1960–1964.
60 Smith TW (1981) J Mol Biol 147:195–197.
61 Altschul SF, Gish W et al. (1990) J Mol Biol 215:403–410.
62 Pearson WL (1988) Proc Natl Acad Sci USA 85:2444–2448.
63 Altschul S, Madden T et al. (1997) Nucl Acids Res 25:3389–3402.
64 Eddy S (1998) Bioinformatics 14:755–763.
65 Bateman A., Birney E et al. (2002) Nucl Acids Res 30:276–280.
66 Madera M, Gough J (2002) Nucl Acids Res 30:4321–4328.
67 Nahoum V, Gangloff A et al. (2003) FASEB J 17:1334–1336.
68 Shah M, Passovets S et al. (2003) Bioinformatics 19:1985–1996.
69 Manning G, Whyte DB et al. (2002) Science 298:1912–1934.
70 Chopra A (2000) J Am Osteopath Assoc 100:S1–S4.
71 Ishida N, Hayashi K et al. (2002) J Biol Chem 277:41147–41156.
72 Rho JA., Socci CR, Merkov ND, Kim L, So N, Lee H, Takami O, Brivanlou M, Choi AH (2002) DNA Cell Biol 21:541–549.
73 Heaney C, Shalev H et al. (1998) Hum Mol Genet 7:1407–1410.
74 Cleiren E, Benichou O et al. (2001) Hum Mol Genet 10:2861–2867.
75 Matthews DJ, Kopczynski J (2001) Drug Discovery Today 6:141–149.
76 Giaccia AJ, Kastan MB (1998) Genes Dev 12:2973–2983.
77 Hainaut P, Hernandez T et al. (1998) Nucl Acids Res 26:205–213.
78 Ollmann MYL, Di Como CJ, Karim F, Belvin M, Robertson S, Whittaker K, Demsky M, Fisher WW, Buchman A, Duyk G, Friedman L, Prives C, Kopczynski C (2001) Cell 101:91–101.
79 Lee C-Y, Clough EA et al. (2003) Current Biol. 13:350–357.
80 Yamauchi TK, Ito J, Tsuchida Y, Yokomizo A, Kita T, Sugiyama S, Miyagishi T, Hara M, Tsunoda K, Murakami M, Ohteki K, Uchida T, Takekawa S, Waki S, Tsuno H, Shibata NH, Terauchi Y, Froguel Y, Tobe P, Koyasu K, Taira S, Kitamura K, Shimizu T, Nagai T, Kadowaki R (2003) Nature 423:762–769.
81 Hussain MM, Strickland DK et al. (1999) Annu Rev Nutr 19:141–172.
82 Chouchane LS, Bousaffara I, El Kamel R, Sfar A, Ismail MT (1999) Int Arch Allergy Immunol 120:50–55.
83 Schmid EF, Smith DA (2004) Is pharmaceutical R & D just a game of chance or can strategy make a difference? Drug Discov Today 9:18–26.

84 Morgan KT (2004) Gene expression analysis reveals chemical-specific profiles. Toxicol Sci 67:155–156.

85 Guerreiro N, Staedtler F, Grenet O, Kehren J, Chibout SD (2003) Toxicogenomics in drug development. Toxicol Pathol 31:471–479.

86 Hamadeh HK, Bushel PR, Jayadev S, DiSorbo O, Bennett L, Li L, Tennant R, Stoll R, Barrett, JC, Paules, RS, Blanchard K, Afshari CA (2002) Prediction of compound signature using high density gene expression profiling. Toxicol Sci 67:232–240.

87 Waters MD, Olden K, Tennant RW (2003) Toxicogenomic approach for assessing toxicant-related disease. Mutat Res 544:415–424.

88 Whittaker PA (2003) What is the relevance of bioinformatics to pharmacology. Trends in Pharmacol Sci 24:434–439.

89 Golub et al. 1999

90 Perou et al. 2000

91 Heijne WH, Stierum RH, Slijper M, van Bladeren PJ, van Ommen B (2003) Toxicogenomics of bromobenzene hepatotoxicity: a combined transcriptomics and proteomics approach. Biochem Pharmacol 65:857–875.

92 Viskin S (1999) Long QT syndromes and torsade de pointes. Lancet 354:1625–1633.

93 Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH (2002) Timing of new black box warnings and withdrawals for prescription medications. JAMA 287:2215–2220.

94 Crumb W, Cavero I (1999) Pharm Sci Technol Today 2:270–280.

95 Keseru GM (2003) Prediction of hERG potassium channel affinity by traditional and hologram QSAR methods. Bioorg Med Chem Letters 13:2773–2775.

96 Splawski I, Shen J, Timothy KW, Lehmann MH, Priori S, Robinson JL, Moss AJ, Schwartz PJ, Towbin JA, Vincent GM, Keating MT (2000) Spectrum of mutations in long-QT syndrome genes. KVLQT1, HERG, SCN5A, KCNE1, and KCNE2. Circulation 102:1178–1185.

97 Larsen LA, Andersen PS, Kanters J, Svendsen IH, Jacobsen JR, Vuust J, Wettrell G, Tranebjaerg L, Bathen J, Christiansen M (2001) Screening for mutations and polymorphisms in the genes KCNH2 and KCNE2 encoding the cardiac HERG/MiRP1 ion channel: implications for acquired and congenital long Q-T syndrome. Clin Chem 47:1390–1395.

98 Lees-Miller JP, Duan Y, Teng GQ, Thorstad K, Duff HJ (2004) Novel gain-of-function mechanism in K(+) channel-related long-QT syndrome: altered gating and selectivity in the HERG1 N629D mutant. Circ Res 86:507–513.

99 Laitinen P, Fodstad H, Piippo K, Swan H, Toivonen L, Viitasalo M, Kaprio J, Kontula K (2000) Survey of the coding region of the HERG gene in long QT syndrome reveals six novel mutations and an amino acid polymorphism with possible phenotypic effects. Hum Mutat 15:580–581.

100 Pietila E, Fodstad H, Niskasaari E, Laitinen PJ, Swan H, Savolainen M, Kesaniemi YA, Kontula K, Huikuri HV (2003) Association between HERG K897T polymorphism and QT interval in middle-aged Finnish women. J Am Coll Cardiol 40:511–514.

101 Anson BD, Ackerman MJ, Tester DJ, Will ML, Delisle BP, Anderson CL, January CT (2004) Molecular and functional characterization of common polymorphisms in HERG (KCNH2) potassium channels. Am J Physiol Heart Circ Physiol [Epub ahead of print].

102 Wilson AG, White AC, Mueller RA (2003) Role of predictive metabolism and toxicity modeling in drug discovery – a summary of some recent advancements. Curr Opin Drug Discov Devel 6:123–128.

103 Beresford AP, Segall M, Tarbit MH (2004) In silico prediction of ADME properties: are we making progress? Curr Opin Drug Discov Devel 7:36–42.

104 Shoeb F, Malykhina AP, Akbarali HI (2003) Cloning and functional characterization of the smooth muscle ether-a-go-go-related gene K+ channel. Potential role of a conserved amino acid substitution in the S4 region. J Biol Chem 278:2503–2514.

105 Crociani O, Guasti L, Balzi M, Becchetti A, Wanke E, Olivotto, M, Wymore RS, Arcangeli A (2003) Cell cycle-dependent expression of HERG1 and HERG1B isoforms in tumor cells. J Biol Chem 278:2947–2955.

106 Priori SG, Barhanin J, Hauer RNW (1999) Genetic and molecular basis of cardiac arrhythmias: impact on clinical management. Part I and II. Circulation 99:518–528.

107 Paulussen AD, Gilissen RA, Armstrong M, Doevendans PA, Verhasselt P, Smeets HJ, Schulze-Bahr E, Haverkamp W, Breithardt G, Cohen N, Aerssens J (2004) Genetic

variations of KCNQ1, KCNH2, SCN5A, KCNE1, and KCNE2 in drug-induced long QT syndrome patients. J Mol Med 82:182–188.
108 Pirmohamed M, Park BK (2001) Genetic susceptibility to adverse drug reactions. Trends Pharmacol Sci 22:298–305.
109 Rushmore TH, Kong AN (2001) Pharmacogenomics, regulation and signaling pathways of phase I and II drug metabolizing enzymes. Curr Drug Metab 3:481–490.
110 Nelson DR (1998) Cytochrome P450 nomenclature. Methods Mol Biol 107:15–24.
111 Ingelman-Sundberg M, Oscarson M, McLellan RA (1999) Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. Trends Pharmacol Sci 20:342–349.
112 Ingelman-Sundberg M, Oscarson M (2002) Human CYP allele database: submission criteria procedures and objectives. Methods Enzymol 357:28–36.
113 Vermeulen NP (2003) Prediction of drug metabolism: the case of cytochrome P450 2D6. Curr Top Med Chem 3:1227–1239.
114 Lamba JK, Lin YS, Schuetz EG, Thummel KE (2002) Genetic contribution to variable human CYP3A-mediated metabolism. Adv Drug Deliv Rev 54:1271–1294.
115 Matsumura K, Saito T, Takahashi Y, Ozeki T, Kiyotani K, Fujieda M, Yamazaki H, Kunitoh H, Kamataki T (2004) Identification of a novel polymorphic enhancer of the human CYP3A4 gene. Mol Pharmacol 65:326–334.
116 Zheng P, Pennacchio LA, Le Goff W, Rubin EM, Smith JD (2004) Identification of a novel enhancer of brain expression near the apoE gene cluster by comparative genomics. Biochim Biophys Acta 1676:41–50.
117 Evans WE, Relling MV (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. Science 286:487–491.
118 Tirona RG, Kim RB (2002) Pharmacogenomics of organic anion-transporting polypeptides (OATP). Adv Drug Deliv Rev 54:1343–1352.
119 Marzolini C, Paus E, Buclin T, Kim RB (2004) Polymorphisms in human MDRI (P-glycoprotein): recent advances and clinical relevance. Clin Pharmacol Ther 75:13–33.
120 Kuhlmann J (1999) Alternative strategies in drug development: clinical pharmacological aspects. Int J Clin Pharmacol Ther 37:575–583.
121 Savage DR (2003) FDA guidance on pharmacogenomics data submission. Nature Rev Drug Dis 2:937–938.
122 Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR (2001) Chemosensitivity prediction by transcription by transcriptional profiling. Proc Natl Acad Sci USA 98:10787–10792.
123 Agrawal D, Chen T, Irby R, Quackenbush J, Chambers AF, Szabo M, Cantor A, Coppola D, Yeatman TJ (2003) Osteopontin identified as colon cancer tumor progression marker. C R Biol 326:1041–1043.
124 Yeatman TJ (2003) The future of cancer management: translating the genome, transcriptome, and proteome. Ann Surg Oncol 10:7–14.
125 Tanne JH (1998) The new word in designer drugs. British Med J 316:1930.
126 Roses AD (2000) Pharmacogenetics and the practice of medicine. Nature 405:857–865.
127 de The H, Chomienne C, Lanotte M, Degos L, Dejean A (1990) The t(15;17) translocation of acute promyelocytic leukaemia fuses the retinoic acid receptor alpha gene to a novel transcribed locus. Nature 347:558–561.
128 Huang ME, Ye YC, Chen SR, Chai JR, Lu JX, Zhoa L, Gu LJ, Wang ZY (1988) Use of all-trans retinoic acid in the treatment of acute promyelocytic leukemia. Blood 72:567–572.
129 Takayama N, Kizaki M, Hida T, Kinjo K, Ikeda Y (2001) Novel mutation in the PML/RARalpha chimeric gene exhibits dramatically decreased ligand-binding activity and confers acquired resistance to retinoic acid in acute promyelocytic leukemia. Exp Hematol 29:864–872.
130 Samuelsson B, Dahlen SE, Lindgren JA, Rouzer CA, Serhan CN (1987) Leukotrienes and lipoxins: structures, biosynthesis, and biological effects. Science 237:1171–1176.
131 Drazen JM, Israel E, Obyrne PM (1999) Treatment of asthma with drugs modifying the leukotriene pathway. N Engl J Med 340:197–206.
132 In KH, Asano K, Beier D, Grobholz J, Finn PW, Silverman EK, Silverman ES, Collins T, Fischer AR, Keith TP, Serino K, Kim SW, DeSanctis GT, Yandava C, Pillari A, Rubin P, Kemp J, Israel E, Busse W, Ledford D, Murray JJ, Segal A, Tinkleman D, Drazen JM

(1997) Naturally occurring mutations in the human 5-lipoxygenase gene promoter that modify transcription factor binding and reporter gene transcription. J Clin Invest 99:1130–1137.
133 Silverman ES, Du J, De Sanctis GT, Radmark O, Samuelsson B, Drazen JM, Collins T (1998) Egr-1 and Sp1 interact functionally with the 5-lipoxygenase promoter and its naturally occurring mutants. Am J Respir Cell Mol Biol 19:316–323.
134 Drazen JM, Yandava CN, Dube L, Szczerback N, Hippensteel R, Pillari A, Israel E, Schork N, Silverman ES, Katz DA, Drajesk J (1999) Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. Nat Genet 22:168–170.
135 Chandrasekharan NV, Dai H, Roos KL, Evanson NK, Tomsik J, Elton TS, Simmons DL (2002) COX-3, a cyclooxygenase-1 variant inhibited by acetaminophen and other analgesic/antipyretic drugs: cloning, structure, and expression. Proc Natl Acad Sci USA 99:13926–13931.
136 Lu AY (1998) Drug-metabolism research challenges in the new millennium: individual variability in drug therapy and drug safety. Drug Metab Dispos 26:1217–1222.
137 Palkimas MP, Skinner HM, Gandhi PJ, Gardner AJ (2003) Polymorphism induced sensitivity to warfarin: a review of the literature. J Thromb Thrombolysis 15:205–212.
138 Rost S, Fregin A, Ivaskevicius V, Conzelmann E, Hortnagel K, Pelz HJ, Lappegard K, Seifried E, Scharrer I, Tuddenham EG, Muller CR, Strom TM, Oldenburg J (2004) Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature 427:537–541.
139 Li T, Chang CY, Jin DY, Lin PJ, Khvorova A, Stafford DW (2004) Identification of the gene for vitamin K epoxide reductase. Nature 427:541–544.
140 Basile VS, Masellis M, Badri F, Paterson AD, Meltzer HY, Lieberman JA, Potkin SG, Macciardi F, Kennedy JL (1999) Association of the MscI polymorphism of the dopamine D3 receptor gene with tardive dyskinesia in schizophrenia. Neuropsychopharmacology 21:17–27.
141 Wangsness M (2000) Pharmacological treatment of obesity. Past, present, and future. Minn Med 83:21–26.
142 Mallal S, Nolan D, Wit C, Masel G, Martin AM, Moore C, Sayer D, Castley A, Mamotte C, Maxwell D, James I, Christiansen FT (2002) Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. Lancet 359:727–732.
143 Hetherington S, Hughes AR, Mosteller M, Shortino D, Baker KL, Spreen W, Lai E, Davies K, Handley A, Dow DJ, Fling ME, Stocum M, Bowman C, Thurmond LM, Roses AD (2002) Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. Lancet 359:1121–1122.
144 Brazell C, Freeman A, Mosteller M (2002) Maximizing the value of medicines by including pharmacogenetic research in drug development and surveillance. Br J Clin Pharmacol 53:224–231.
145 Jones JK (2001) Harmacogenetics and pharmacoepidemiology. Pharmacoepidemiol Drug Saf 10:457–461.
146 Dayhoff MO (1969) Computer analysis of protein evolution. Sci Am 221:86–95.
147 Rybak B (1978) A program for teaching bio-informatics. Biosci Commun 4:158–159.

16
Genome Data Representation Through Images: The MAGPIE/Bluejay System

Andrei Turinsky, Paul M. K. Gordon, Emily Xu, Julie Stromer, and Christoph W. Sensen

16.1 Introduction

Graphical display systems for complex data, which can be used to analyze what would otherwise be overwhelming amounts of data, are becoming increasingly important in many scientific fields. Molecular biology and genomics are key areas for this phenomenon, because of the continually increasing complexity and number of genome databases and analysis tools. In 1996, Gaasterland and Sensen introduced a system for automated analysis of biological sequences, called MAGPIE (multipurpose automated genome project investigation environment) [1]. This tool-integration system executes and integrates the analysis from multiple bioinformatics tools into an easily interpretable form. A typical microbial genome has up to 1000 open reading frames (ORFs) per megabase of genome sequence. MAGPIE typically runs 20 tools on a genome, if each tool results in hits against 100 database entries (which is not an uncommon number), approximately two million pieces of evidence would need to be recorded and mapped along a megabase of genome sequence. When dealing with this amount of information, the saying "a picture is worth a thousand words" is certainly very true.

MAGPIE has been used for the analysis of complete genomes, genome DNA fragments, EST, proteins, and protein fragments. Initially, all MAGPIE output was in tabular format, but the need for rich visual representations of genome analyses became evident early on. MAGPIE summarizes information about evidence supporting functional assignments for genes, information about regulatory elements in genome sequences, including promoters, terminators and, Shine Dalgarno sequences, metabolic pathways, and the phylogenetic distribution of genes. MAGPIE sorts and ranks the evidence by strength and is able to display the level of confidence associated with database search results.

MAGPIE was the first genome analysis and annotation system to add graphical representations to the results. On the basis of continual user feedback from a variety of installations, the images have been refined over time, allowing annotators to process the quantity of relevant information quickly and efficiently. Because MAGPIE has one of the most comprehensive graphical capa-

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

bilities, we will use MAGPIE as the example of genome annotation supported by imaging. We describe the meaning of the various images, the algorithms used to create them, and the data types from which they are derived. MAGPIE mainly produces precomputed results which are stored on the server side. On request, the information is served over the Internet.

The MAGPIE image types reflect the fact that sequence data is stored and presented in a hierarchical manner. A MAGPIE project usually consists of, but is not limited to, related sequences of an organism. To make browsing and data maintenance easy, sequences are organized into logical groups. For example, all sequences from a single clone are normally placed in one group, because they will be joined as sequencing progresses, or all ESTs from a tissue could be grouped together. Although it is convenient for the user to subdivide datasets, there is no limit on the number of sequences in a group, because some EST sets have 50,000 sequences in a group, which MAGPIE paginated for Web browsers. The images presented in this text are from the *Sulfolobus solfataricus* P2 sequencing project [2]. This project was chosen as the example because it includes all the graphical representations which can be produced by MAGPIE. The complete *Sulfolobus* MAGPIE project is available at http://magpie.ucalgary.ca/magpie/Sulfolobus_solfataricus_P2/private/.
An analysis of all publicly available genomes and several EST sets can be accessed through http://magpie.ucalgary.ca.

The MAGPIE images can be classified into three categories – representation of evidence, summary of genome features, and biochemical assay simulation, which supports the design of follow-up experiments. We discuss how these images aid the researcher in genome sequencing and annotation through pattern recognition and information filtering and how they can be used to support the validation of genome data.

In 1999 we began to work on a more advanced system, called Bluejay, which is capable of integrating many types of genome data. The system is Java-based and capable of handling plain text and XML data sources. Using exported MAGPIE XML data and XML documents served by other providers it is capable of handling any genome, even the human genome. Bluejay is an extremely flexible system, which can be configured by the users to display any level of detail necessary to answer complex questions, in a way that HTML views such as MAGPIE simply cannot.

16.2 The MAGPIE Graphical System

The MAGPIE user interface, which was initially implemented with a Web-based display that supported text and table output, has been developed over time into a Web-based graphical display system. As described previously [1], the tools used in MAGPIE include, but are not limited to, the FastA [3] family of programs (including the protein fragment analysis tools *fastf* and *tfastf*), the BLAST [4] family of programs (ungapped, gapped, and Position Specific Iterative), BLOCKS [5], ProSearch [6], Genscan [7], Glimmer [8], GeneMark [9], Paracel GeneMatcher (http://www.paracel.com/), and Decypher TimeLogic (http://www.timelogic. com/). To link image representations to alignments, individual tool "hits" and other related information, the original tool outputs (e.g. BLAST of BLOCKS responses) are stored as HTML files after data processing. Hit length and scores are extracted during the response processing using dedicated parsers. The hits are sorted into user-defined confidence levels.

The parts of the modular computer code used for the MAGPIE input and output are modules for Web standards written in the Perl5 (http://www.perl.com) programming language: HTML (Hypertext Markup Language), CGI (Common Gateway Interface), and PNG (Portable Network Graphics). The text reporting system is based on a combination of pre-computed HTML pages and CGI programs producing HTML dynamically. A graphics library module called GD.pm (http://stein.cshl.org/WWW/software/GD), which is dynamically loaded into the Perl5 system, is used to generate the MAGPIE graphics. GD.pm provides functionality for drawing and filling lines, basic geometric shapes, and arbitrary polygons on a two-dimensional canvas. The canvas can be translated into a Web browser-readable form such as PNG. The drawing functions, when applied to linearly encoded data such as DNA or proteins, lend themselves to particular succinct representations. Rulers, along which features are positioned, are drawn as straight lines. Simple ticks (small lines perpendicular to the ruler) generally represent position-specific features that occur frequently, e.g. stop codons. Unique polygons that occupy more space are used for position-specific features, which occur less frequently, e.g. promoter sites. Data that cover a range, e.g. open reading frames, are displayed as boxes. These boxes may be subdivided when additional information needs to be encoded.

MAGPIE's graphics fulfill three main needs in a genome project once data is collected – display of the genome features at different levels of detail in the genome context, evaluation of the evidence supporting a feature's functional annotation, and quality control. Four types of information are used to generate images in MAGPIE – user preferences, sequence information, data from both external tool output and internal analysis, and manual user annotations (verifications). User-based annotations are stored as text files, usually created via the Web interface, but they can also be imported from files using standards such as GenBank flat format or the General Feature Format (http://www.sanger.ac.uk/Software/formats/GFF/). User-configurable visualization parameters (e.g. the bases-per-pixel scale and the maximum image width) are stored in plain text configuration files, similar to the previously described [1] configuration files for confidence level criteria and other configurable parts of MAGPIE. By default, images in MAGPIE are defined with a maximum width of 1000 pixels. This enables landscape mode printing of the images on $8.5'' \times 11''$ paper.

Although most data in MAGPIE is hierarchical and stored as text files, cross-referencing of equivalent sequence identifiers is done using binary Berkeley Database Manager (DBM) files. This exception to text-file storage enables rapid execution of text-based searches for sequence descriptions and identifiers.

The graphical reporting system in MAGPIE has two distinct user-configurable modes: static or dynamic, respectively. In the static graphical mode all images are precomputed for viewing after the analysis is finished. This requires considerable disk space, but it is computationally and temporally efficient when the sequence and analyses do not change much, e.g. when a completely finished genome is analyzed. In the dynamic mode, images are created on demand, using the data extracted from the analysis. Although this requires more *ad hoc* computation, it is appropriate when the underlying sequences or the analyses are frequently updated, for example in an ongoing genome-sequencing project.

Two key features for viewing data in context are hierarchical representation of the

data and consistent display idioms. Idioms such as bolder coloring for stronger evidence (darker text, brighter hues, respectively), and using blue and red to indicate the forward or reverse DNA strand, respectively, pervade the images and enable complex information to be encoded in the graphics without cluttering them. Good idioms need be learned once only [10]. For example, by surveying the annotation and evidence strength status from Fig. 16.3, and the overlaps, it becomes evident how the clone relates to its genome neighbors, and how much information has been gathered about the nonredundant part of the sequence.

Consistent color use improves the delivery of information in the MAGPIE images. Blue bars and borders indicate information located on the positive DNA strand (forward strand) whereas red bars and borders represent information located on the negative DNA strand. Black generally indicates information that is not strand-specific. Coloration of analysis data is specified in a user-definable color preference file. Consistent coloring can group the evidence from similar tools by color range. For example and shown later in the text, FastA [3] hits against an EST collection will always be shown in a particular green hue, and FastA hits against a protein database might be colored in a different green hue.

Shading is used throughout the MAGPIE interface to denote the confidence level of the evidence. Stronger evidence is always displayed in darker shades. For example, description text in reports is black when the evidence is good, gray when it is moderate, and white when it is only marginally useful. As shown below, this is true for the graphical evidence displays also. It is easy to filter out information related to potential genome function by following this simple concept. These representational consistencies also reduce visual clutter. Information is implicitly conveyed in the color used instead of requiring explicit depiction or labeling of the displayed features.

Three key features of the individual ORF display are succinctness, pattern display, and data linking. Succinct representation of the tool responses is essential to enable the annotator to survey the salient information quickly. Some of the information, for example subject description and discriminant score, remains in textual form within the images whereas details such as the location of the hits are graphically mapped on to the images. The exact positioning of hit patterns is important to the annotator, because it can determine the relevance of the match. The comparative match display is maximized in MAGPIE by displaying results from tools, which are based on similar algorithms, data sets, and evidence types (i.e. amino acid, DNA, and motif) atop each other.

Although a succinct representation of the responses can be very useful, it is equally important to be able to access the original responses, database entries, associated metabolic pathways and other related information, thus most image types contain user-configurable hyperlinks to Web-based information, including SRS-6, ExPASy, or the NCBI Web site.

Finally, simulation images assist the wet-lab researcher during the verification process. MAGPIE analyses, like all results generated by automated genome analysis and annotation systems, are only computer models, which often need verification by means of a biochemical experiment.

16.3 The Hierarchical MAGPIE Display System

In the paragraphs below we introduce the different graphical displays implemented in MAGPIE. The MAGPIE hierarchy is reflect-

ed in the set of graphical images. The resolution of the images is increased over several levels until almost single-base-pair resolution is reached. Figure 16.1 shows the hierarchical connection between the different images. Depending on the state of the analyzed sequence, not all images are present, and some images are mutually exclusive.

16.4 Overview Images

16.4.1 Whole Project View

MAGPIE can track sequencing efforts from single sequence reads to fully assembled clones and genomes. Based on mapping information, which identifies the relationship of clones in the sequencing project, MAGPIE can automatically generate and display nonredundant sequence(s) and the nonredundant gene set. Figure 16.2 (a and b) shows examples of the images that display the summary "whole project view" page of the *Sulfolobus solfataricus* P2 genome. Acting as a starting point for the annotator, this single page contains hyperlinked images representing all of the contiguous sequences (contigs) in the MAGPIE project. Contigs are drawn to scale, and color-filled according to their MAGPIE state. The states used in the *Sulfolobus* MAGPIE project are primary, linking, polishing, and finished, but the user for other projects could define other states, such as an EST state. These states can be included in the overview graphics, or for large EST sets, simple textual links are provided, because a graphical overview provides no additional information.

The colors for the states can be set in the user preference files. Users can also define in which of the states the sequence is resolved well enough so that MAGPIE can assemble larger contigs from individual clones. In the example of the *Sulfolobus* MAGPIE project, this would be sequence in state "polishing". Overlapping contigs are appropriately positioned in the image, and the areas of overlap are grayed out to denote redundancy. In keeping with the color usage described earlier, blue outlines denote that the contigs are in their normal (forward) orientation whereas those outlined in red were reverse-complemented to fit into the genome assembly. White text on a black background denotes the presence of manual annotation (verification) on the labeled contig. In Fig. 16.2a, it is apparent from the black label text and gray fill that clone l910_127 is the only clone without annotation. This clone was not annotated because it is completely redundant.

Even though overlaps between MAGPIE contigs are calculated to remove redundancy, MAGPIE is not meant to be a full-fledged assembly engine. For contigs to be considered overlapping the user must specify that two clones are neighbors. This is information usually known from the clone-mapping phase or derived from self-identity searches in MAGPIE. The user specification avoids spurious assemblies that may be taken for granted. The extent and orientation of the sequence overlap is determined by running a BLAST similarity search. On the basis of the percentage similarity and the length criteria set for the project, the overlap is either accepted or rejected. If the overlaps do not occur at the very ends of the contigs, the match is also rejected. This avoids linking contigs based on repetitive regions. Based on the one-to-one neighbor information, larger contigs are formed using the logic:

1. Let S be a set of sets, each containing a single contig
2. Let N be the set of neighbor relationships

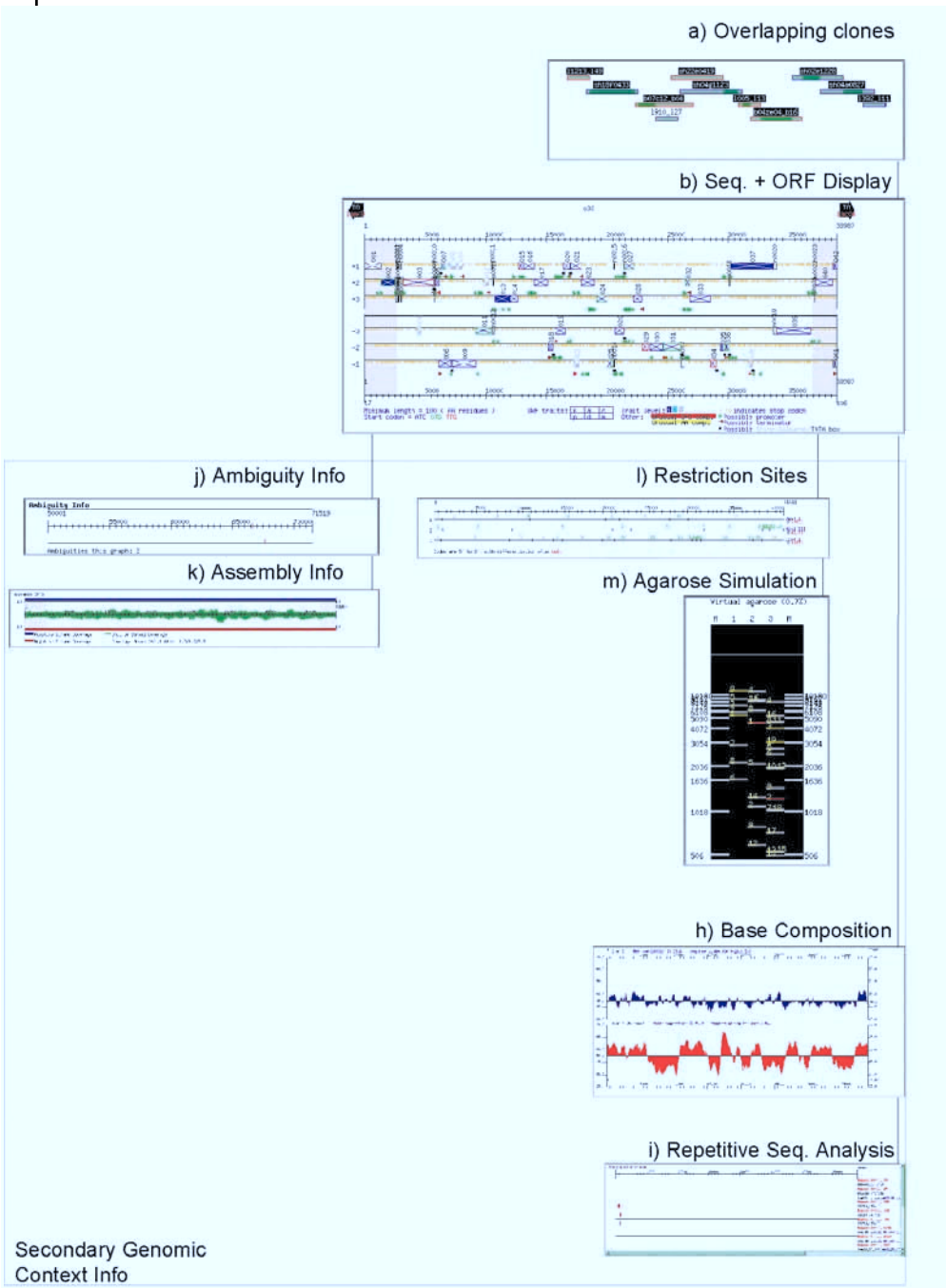
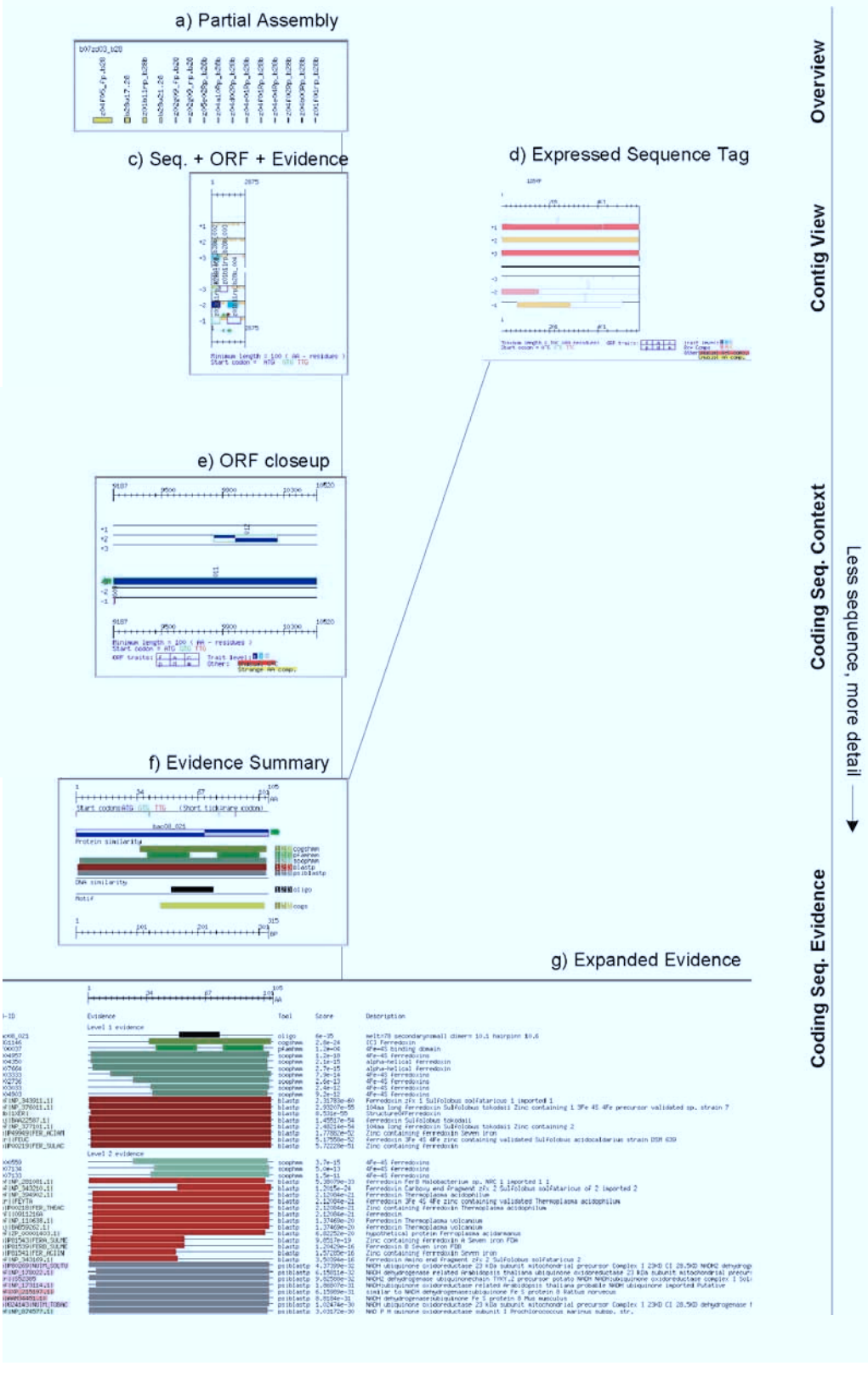


Fig. 16.1 Image hierarchy. From top to bottom more detail is shown about smaller subsequences. Where connecting lines occur the images are either juxtaposed or click-through. The images are labeled according to their figure numbers.


a) Partial Assembly
c) Seq. + ORF + Evidence
d) Expressed Sequence Tag
e) ORF closeup
f) Evidence Summary
g) Expanded Evidence
Overview
Contig View
Coding Seq. Context
Coding Seq. Evidence
Less sequence, more detail

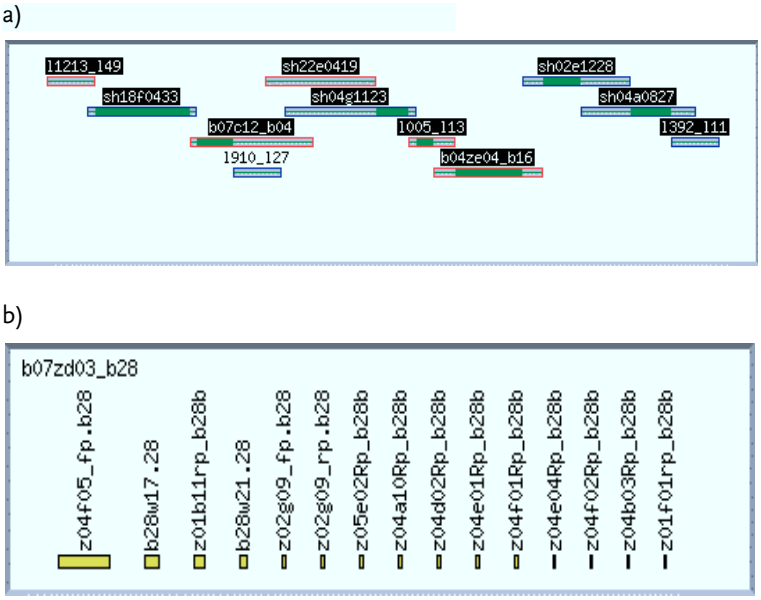


Fig. 16.2 a) Overlapping clone cluster in a MAGPIE project. Each sequence is hyperlinked to its MAGPIE report. All of the sequences are filled in with green, denoting finished sequence, and shaded where redundant. Red-outlined sequences are reverse-complemented in the assembly. White-letter labels denote the presence of annotations in the sequence. b) Partially assembled fragments of BAC b07zd03_b28. Fragments are sorted by size and hyperlinked to their respective MAGPIE reports. The yellow fill indicates that sequences are in the linking state.

3. While N is not the empty set
   a. Pick a relationship R(C1,C2) from N
   b. N = N – R
   c. Find S1, the set in S to which the contig C1 belongs
   d. Find S2, the set in S to which the contig C2 belongs
   e. S = (S – S1 – S2) ∪ (S1 ∪ S2)

A consensus sequence for each contig set in S is determined. When conflicts between overlapping contigs occur, the better-resolved sequence of the contig in a more complete state takes precedence. The ORFs on the consensus sequence are identified and a table of ORF equivalency across all the contigs is created subsequently.

At step 1, each contig is in its own set. Because the data set consists of nonredundant contigs, the sets in S are by default disjoint. Steps 2, 3a, and 3b iterate through all known connections. Two sets, which are determined in steps 3c and 3d can be combined when a contig from the first set is connected to one from the second set, which is valid because of the transitive property of contig connections. In step 3e, the two now connected contigs are removed from S and replaced with a combined set. The resulting sets of disjoint contigs are called “supercontigs” in MAGPIE, because they can consist of more than one clone.

The overlap and equivalency information is used in the images described below to

propagate ORF information across equivalent sequence feature displays. The algorithm has been used in the *Sulfolobus solfataricus* P2 genome project to successfully assemble 110 bac, cosmid, and lambda clones into a single nonredundant contig. Use of this sort of "tiling" technique enables large datasets to be managed and analyzed. For genomes that do not have natural subsections, such as random shotgun bacterial genome assemblies, MAGPIE provides utilities to splice the sequence into stretches (usually 100,000 bases for Web browser convenience).

16.5 Coding Region Displays

Several different image types represent the ORF evidence with increasing levels of detail. Sometimes it is necessary to see the actual evidence, e.g. for decision making during the manual annotation (verification) process. At other times the larger context of the ORF is more useful, in this case, a detailed display containing all MAGPIE evidence would be overcrowded. These needs necessitate multiple representations of the same data. The varying levels of evidence abstraction are the most powerful part of the MAGPIE graphical environment.

A user can specify a wide variety of tools to be run against all contigs in a particular MAGPIE state. The operation and analysis of these tools can take considerable time. Tools used by MAGPIE can also produce periodically updated results, adding to the dynamic nature of the evidence. A user might wish to store all of the MAGPIE-generated images or create them on demand. This depends on available disk space, CPU power, and the frequency at which contigs and tool outputs are updated.

All scripts that generate a graphical representation of a contig and its ORFs may generate more than one image for the sequence. If the combined contig length and scale factor exceeds the maximum image width, the image is split into multiple "-panes". The number and size of the images must be determined before any drawing takes place. This enables the drawing canvases to be allocated in the program. The information is also used to create the required number of image references in the HTML pages. The height of the image is fixed, because it depends entirely on fixed settings. The image width is variable because sequences are represented horizontally. The width is a function of the sequence length multiplied by a scale factor, plus constant elements such as border padding. When multiple panes are required, all but the last image have maximum width. If the last image has ten or fewer pixels it is merged into the preceding image. This slightly exceeds the maximum permitted width, but avoids an unintelligibly small sequence display.

16.5.1 Contiguous Sequence with ORF Evidence

Figure 16.3 displays a sequence and its features with the largest amount of data abstraction. Images of the type shown in Fig. 16.3 summarize the evidence against all the ORFs in a contig in all six open reading frames, providing a rich summary of the genome context for a set of genes. They also show additional genome features, which might be located around coding regions – promoters, terminators, and stop codons. Links to the corresponding ORF reports are provided via the HTML image map. ORFs are labeled sequentially from left to right. In our example from the *Sulfolobus* project the 100-amino-acid-residue cutoff is stated in

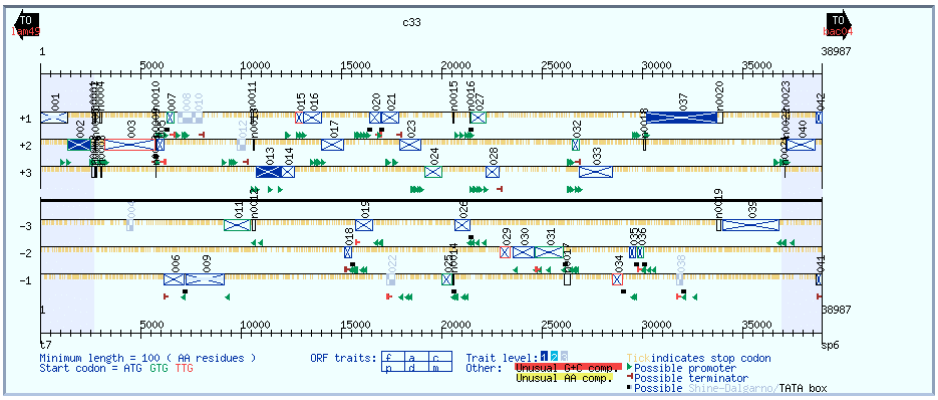


Fig. 16.3 Contiguous sequence with open reading frames displayed. Boxes on the six reading frame lines represent possible genes. The boxes are exed when annotated. Light exed boxes have annotations described as hypothetical or uncharacterized. Sub-boxes in unannotated genes indicate composition characteristics, plus the best level of protein, DNA and motif database hits. Grayed-out genes have been suppressed. Background shading and hyperlinked arrows in the corners indicate neighboring sequence overlaps. Boxes with labels that start with "n" are possible genes shorter than the specified minimum length.

the lower left-hand corner. Inter-ORF regions are analyzed separately by MAGPIE using a set of scoring criteria, which is different from that defined for large-ORF regions. The goal of inter-ORF analysis is to identify small coding regions (e.g. small proteins and RNA-coding regions). The names of the potential small coding regions meeting the user-defined criteria are denoted with the prefix "n". Clicking on a displayed small coding region brings up a screen that displays the evidence as shown in Fig. 16.8. The user must confirm that the small sequence segment is, indeed, a coding region. When confirmed, the identifiers of the small ORFs are displayed with the "s" prefix. This naming convention differentiates small ORF without the need to rename all downstream ORFs after a small ORFs is confirmed.

There are several aspects to ORF coloration. Grayed-shaded boxes indicate ORF suppression when users deem particular ORF to be non-coding. This might be when the ORF is more than a certain percentage inside another ORF (typically completely contained in an ORF on the opposite DNA strand), or the ORF has an unusual amino acid composition. MAGPIE can be configured to automatically suppress ORF for either reason, or for lack of evidence. ORF with Xs (so called Saint Andrew's crosses) through them have been annotated. A white background indicates that the assigned function is "putative", "hypothetical", or "uncharacterized". All three words stand for unknown function. These functional assignments (or lack thereof) can be carried over from equivalent ORF in the genome. The equivalencies used were determined in the process of creating Fig. 16.2a. Outline colors for ORF starting with ATG, GTG, and TTG are blue, green, and red, respectively. The different colors are used only if they are defined as valid start codons for the organism. When the ORF starts upstream of the current contig and the start codon is unknown, the outline is black. Black also indicates the use of alternative start codons, which can occur in some organisms.

ORFs not validated by a user are split into three by two isometric blocks. These blocks can be colored to indicate the presence and strength of certain evidence. This is indicated in the "ORF traits" section of the Fig. 16.3 legend. The blocks in the upper half denote calculated sequence characteristics. Blocks "f" and "a" indicate on/off traits. A blue "f" block denotes that the codon usage in the ORF fits a chi-square distribution test for frequencies observed in this organism. A blue "a" block denotes that the purine (A + G) composition of this ORF is greater than 50 %. This is known to be an indicator of a likelihood of good coding for many prokaryotes [2]. For organisms with a (G + C) % greater or smaller than 50 %, the "c" block in the upper right corner of the block represents another indicator of coding sequence, (G + C) % codon compensation. Calculation of this term is based on the third (and to a lesser extent second) position codon wobble. Compensation at confidence level 1 occurs when the combined frequency of (G + C) % compensation is highest in the third base of the codons. For Level 2 the second base (G + C) % compensation is the highest, followed by the compensation for the third base of the codons.

The blocks in the lower half denote database search results and the level of confidence in three levels, level one indicating the strongest evidence. The colors for the three levels are blue, cyan, and gray, respectively. The lower half trait levels are determined by comparing extracted similarity analysis scores with the user-specified criteria. The "p" block indicates the highest level of protein similarity found through the database searches. The "d" block indicates the highest level of DNA similarity found. The "m" block indicates the best level of sequence motifs found (e.g. scored Prosite hits). After learning the representation scheme, the annotator can quickly see the nature and strength of coding indicators for all ORFs in a sequence.

Other indicators for transcription include Shine–Dalgarno motifs [11], promoters, and terminators. These features are displayed in the appropriate reading frame as small black rectangles, green triangles, and red sideways Ts, respectively. In keeping with representational consistency, the candidate with the highest score for each of these features around any ORF is colored a darker shade. Shine–Dalgarno motifs are found by matching a user-defined subsequence, which represents the reverse complement of the 3′ end of the organism's 16s rRNA molecule. Promoter and terminator searching are available for archaeal DNA sequences (Gordon and Sensen, unpublished work).

Stop codons are marked with orange ticks within the reading frames. The location of stop codons is determined while the image is being created, as follows. In each forward translation frame the search for the next stop codon begins at the first in-frame base represented by the next pixel, thus increasing the calculation efficiency without sacrificing information in the display. For example, if a stop codon is found at base 23, and each pixel represents fifty bases, the search for the next stop codon in that frame starts at base 51. This is because 51 is the first in-frame triplet in the next pixel, representing the {51,100} range. By not repeatedly drawing stop ticks in the same pixel, no rendering effort is wasted. Another display shortcut is to search for the reverse complements of the stop codons on the forward strand in the same manner when rendering the negative DNA strand. Finding stop codon reverse complements saves the effort of reverse complementing the whole sequence and inverting the information again for graphical rendering.

On the ends of the lower ruler, labels can be added to indicate information about the

priming sites, which are flanking the insert of the clone. In Fig. 16.3 it is shown that the insert is in sp6 to t7 orientation. The information about the ends of a clone is stored in the configuration file that defines the clone relationship.

The image also displays the overlaps between contigs, which were determined during the generation of Fig. 16.2 (a and b). Overlaps are denoted with arrows in the upper left and right hand corners of the image. The extent and orientation of the overlaps is indicated by the shading of the background between the upper and lower rulers in the blue (for forward orientation of the neighbor) or red (for reverse complement orientation of the neighbor).

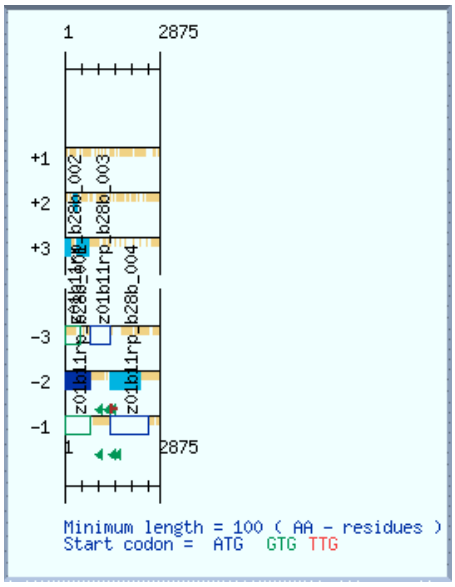


Fig. 16.4 Contiguous sequence with open reading frames and evidence displayed. Evidence and ORFs are displayed in their frames and locations. This facilitates easy recognition of frameshifts and partial genes. Other characteristics are the same as in Fig. 16.3.

16.5.2
Contiguous Sequence with Evidence

Figure 16.4 shows an example of a contiguous sequence display, with evidence. Images of this type are similar in layout to the image shown in Fig. 16.3, reducing the user's learning curve. This type of image is used when the sequence is in a primary or linking state, i.e. when multiple sequencing errors and ambiguities might still exist in the sequence. The key difference between Figs. 16.3 and 16.4 is that evidence in Fig. 16.4 is not abstracted into ORF traits. Evidence is displayed at its absolute position on the DNA strand. Figure 16.5 demonstrates the usefulness of this display. The cyan (level 2) and blue (level 1) evidence is on the –2 DNA strand, and a frameshift at the 3′ end of the sequence is likely, because of the neighboring ORF on the –3 strand. Higher ranked evidence always appears above lower ranked evidence. In this way, the more pertinent information is displayed in the limited screen area available. The evidence markings are hyperlinked to the pattern matches and alignments which they represent. The ORFs are linked to a Fig. 16.8 view of the evidence that falls within the same boundaries on the same DNA strand.

These images can also be used to find previously unrecognized assembly overlaps. To identify potential overlaps, a MAGPIE sequence database search of each contig against all contigs in the project is performed, and self-matches of contigs are suppressed in the subsequent analysis. Any matches potentially mating two contigs are then visible. This kind of information would not be represented in Fig. 16.3.

16.5.3
Expressed Sequence Tags

ESTs (Fig. 16.5) are displayed in a manner very similar to primary and linking sequences from a genome project, because

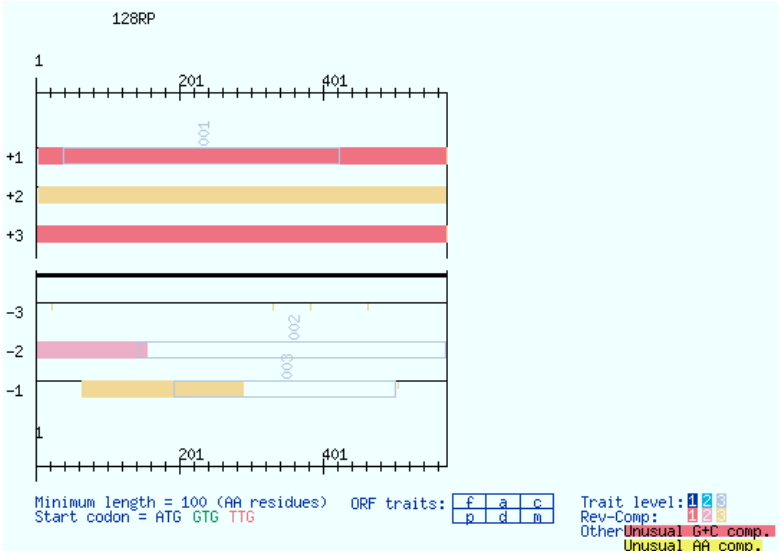


Fig. 16.5 *Mastigamoeba balamuthi* Expressed Sequence Tag (EST). The prevalence of red bars denotes that hits are to the reverse complement strand, indicating that the EST is in the 3′ to 5′ orientation. A sudden jump of evidence from one frame to another can clearly indicate a base-calling error (frameshift) in the sequence, as in Figure 16.4.

they share many of the same characteristics such as short length and error-proneness. EST additionally display evidence in frame as shades of either blue or red for forward and reverse complemented evidence. Because the EST is only translated on one strand, this evidence coloration can be used for rapid identification of EST that must be reverse complemented for further amino-acid-level analysis.

16.5.4 ORF Close-up

The main purpose of Fig. 16.6 is to display a close-up of the ORF and surrounding features. Links to the overlapping ORFs are provided, so that the user can check whether the inner ORF or the outer ORF is an artifact, and to provide indications for frame shifts, which might result in a slightly overlapping ORF such as ORF number 009 in the example. The coloration of the ORF is analogous to that in Fig. 16.3; this is useful for smaller ORFs which otherwise could be difficult to read. The green arrow indicates the orientation of the ORF. ORF on the negative DNA strand are reverse-complemented in Figs. 16.7 and 16.8 to display all evidence in 5′–3′ orientation, thus they can be displayed in orientations different from those in Fig. 16.6.

Because of the simplicity of this image, its width is fixed. Unlike in most other images, the scale factor is a function of the sequence length divided by the fixed width. The rulers have major markings every 100 pixels. For example, in Fig. 16.6 the scale is (9900 – 9500)/100, or four bases per pixel. Because of its similarity to the contiguous sequence images, this image provides a bridge between the genome context already described and the ORF-specific context of the functional analysis images that follow.

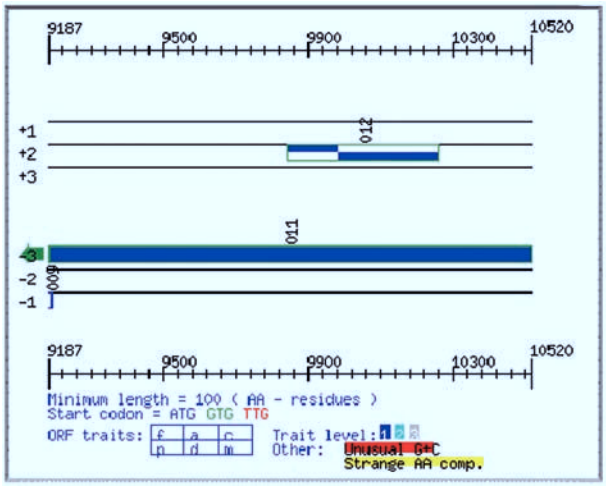


Fig. 16.6 ORF close-up. Displays the exact start and stop coordinates, and hyperlinked overlapping ORFs. Gene labels and trait sub-boxes are potentially easier to read than in Fig. 16.3.

16.6 Coding Sequence Function Evidence

When looking at the genome context, the user has a choice of linking either to information about the whole genome sequence or to information about a particular coding sequence. The coding sequence information has a common representation for assembled, partially assembled, and EST sequences. These functional evidence images form the most important visual component of MAGPIE, and usually the next image browsed in all projects. The discussion of secondary genome images is therefore left to a later section, although they are technically higher in the hierarchy than evidence graphics.

16.6.1 Analysis Tools Summary

Figure 16.7 contains a summary of the location of evidence from all tools along an ORF. This provides an overview of which parts of the ORF have evidence. The image has a fixed width, and it is usually shown next to the similarly sized Fig. 16.6. Its height depends on the number of tools that yielded valuable responses. This requires that all of the evidence information is loaded into memory before canvas allocation and drawing of the image.

The graphic can also be useful in determining the location of the real start codon. One can rule out the first start codon under some conditions by using the rare and start codon indicators at the top of the image, combined with the fact that supporting evidence might only exist from a certain start codon onwards. By default, rare codons are those that normally constitute less than ten percent of the encoding of a particular amino acid for the particular organism. Rare codons are colored according to the color scheme for start codons. Shine–Dalgarno sequences are denoted with a black rectangle near the start codon indicators. Similar to Fig. 16.4, highly ranked evidence is placed on top so that best results are always

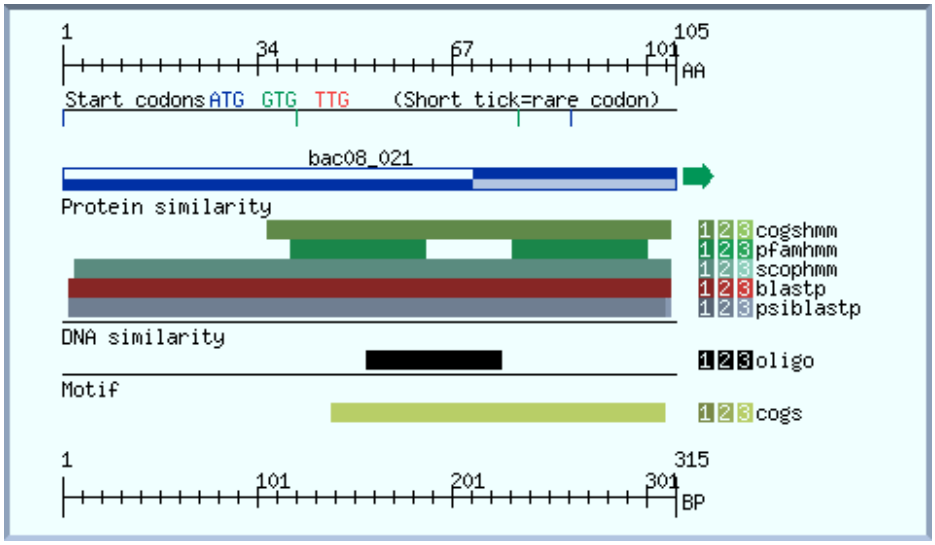


Fig. 16.7 Evidence Summary. Evidence is grouped by type, and displayed as one line for all hits from a tool. Better evidence is lighter and closer in the foreground.

shown for any region of the ORF. In the interest of efficiency, both Figs. 16.7 and 16.8 are generated from the same program at the same time, so that all the evidence to be displayed needs be loaded once only.

16.6.2 Expanded Tool Summary

Figure 16.8 displays in detail all of the evidence accumulated for the ORF during the MAGPIE analysis. The top ruler indicates the length of the translated ORF in amino acids. The bottom ruler indicates the position of the ORF within the contig. The rulers are numbered from right to left when the ORF is on the reverse-complement DNA strand. In this way the evidence is always presented in the same direction as the ORF translation. The evidence is separated into those database entries that have at least a single level-one hit, at least one level-two hit, and others. The evidence order of placement is identical to that used in Fig. 16.4 to highlight the best parts of hits.

This kind of image can be created for any sub-sequence. It is also used to display evidence in the confirmation pages for small inter-ORF features. For every database subject, the hit score is shown in the third text column, and the database subject description in the last column, thus consistently high scores and consistent descriptions are easy to spot.

The three types of representational display link to more in-depth information. The first linked data in text form are the accession numbers for the database subjects that hit against the query sequence in the first text column. The accession numbers are linked to the original database entries e.g. GenBank or EMBL in accordance with a link-configuration file. For example, the default MAGPIE links generally connect to information provided by the Sequence Retrieval System (http://www.lionbio.co.uk)

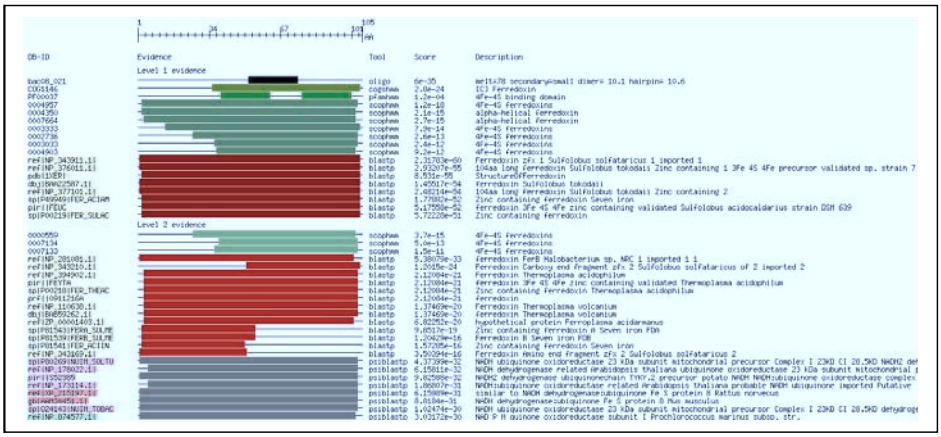


Fig. 16.8 Expanded Evidence. Evidence is sorted by level, tool, score, and length. The first column links to the database ID of the similar sequence. The second displays the similarity location. It is linked to the original tool report. The third names the tool used. The fourth displays the tool's scoring of the match. The fourth displays the Enzyme Commission number, hyperlinked to further enzyme information. The last column displays the matching sequence description.

of the Canadian Bioinformatics Resource (http://www.cbr.nrc.ca). Other sites might configure these links to point to Entrez (http://ncbi.nlm.nih.gov/Entrez) at NCBI. In addition to being hyperlinked, if an analysis result includes an NCBI GI identifier, a MAGPIE index that maps GI numbers to taxonomic information is used to color the identifier background. These colors and divisions are consistent with genes' "taxonomic distribution signatures" used elsewhere in MAGPIE: Viruses: brown, Bacteria: green, Archaea: yellow, Eukarya (other than those that follow): pink, Fungi: orange, Plants (viridiplantae): purple, Metazoa: red.

For the second display link, if an Enzyme Commission (EC) number is associated with the database subject, the number is placed in the fourth text column and linked to a MAGPIE page listing the metabolic pathways in which the enzyme occurs.

The last, but most informative, linked component is the colored match coverage. The quality and types of evidence are clear because of the different colors assigned to the respective tools and respective confidence levels. The color differentiation can also be used to display other differences such as BLASTs against different subsets of a database. Placing the mouse over the area with similarity causes a message to appear on the browser's status line. The message contains the exact interval of the similarity on both the query and subject sequences. The similarity display is hyperlinked to the original data in the text response.

The positioning of evidence rows in this image is more complex than in Fig. 16.7. The logic behind this display is:

1. Separate the evidence into sets where the tool and database id for the matches are the same.
2. Separate the tool/id hit sets into three sets where the top level match in the tool/id set is either 1, 2, or 3.
3. Order in descending sequence the tool/id match sets within each top match-level set by the user-specified tool rankings.
4. Order tool/id sets within a tool ranking by score. If all scores are greater than

unity, order in descending sequence, otherwise order in ascending sequence (e.g. expected random probability scores).
5. Within a score, rank in descending order tool/id sets by the total length of ORF intervals they cover, effectively giving longer hits higher priority.
6. Within a length, sort alphabetically by hit description.
7. Within a description, sort alphabetically by database identifier.

This fine level of sorting ensures predictability for the user. Evidence is sorted in terms of relevance and lexical ordering from top to bottom. In practice, the sorting is quite fast because differentiation is usually made between tool/id sets in step 4. Steps one and two are only executed once. The system uses Perl's built-in sort, which is an implementation of the all-purpose quicksort algorithm [12]. All information about database search tools, scores, and hit lengths is kept in hash tables for quick reference during sorting. The speed benefit of hash table lookup outweighs its space cost. MAGPIE is usually run on servers with the capacity to execute multiple analyses in parallel, therefore short-term requirements for large amounts of main memory are usually dealt with easily. The total height of the image can be calculated only after the number of tool/id sets and their ordering is determined. The image width is determined by keeping track of the longest value in each of the columns while the evidence is loaded. Drawing can only begin after the height and width are determined.

In addition to all of these links to more biological data in the graphic, various column headers are hyperlinked to help documents about the meaning and usage of those fields. For example, the DBID header links to an explanation of the identifier's taxonomy-based background coloration scheme. This follows a general usability pattern in MAGPIE where help links are displayed in the context in which they are used.

16.7 Secondary Genome Context Images

Images displaying characteristics of whole DNA sequences, e.g. the Base Composition or the Assembly Coverage Figures described below, are usually drawn to the same scale as Fig. 16.3. They can be juxtaposed on top of each other to view them in the context of the ORF locations, because they are normally used to identify ORF traits not represented in the images already described.

16.7.1 Base Composition

Figure 16.9 displays two base composition distributions along the DNA sequence. In both graphs the colored region shows the actual base composition, which is determined by a "sliding window" method used along the forward DNA strand.

The (G + C) % graph is configured with a mean (as indicated by a horizontal line) equal to the average of the complete genome sequence. Denoted on the left scale in the example, the *Sulfolobus solfataricus* P2 genome average is 35 %. The graph enables the rapid detection of areas of unusual base composition. Such aberrations might, for example, indicate the presence of transposable elements or other genome anomalies. In the example chosen, however, there is no great variability, indicating a low likelihood of the occurrence of transposable elements in this region of the genome. The 34.1 % average base composition for this sequence is denoted on the right hand scale.

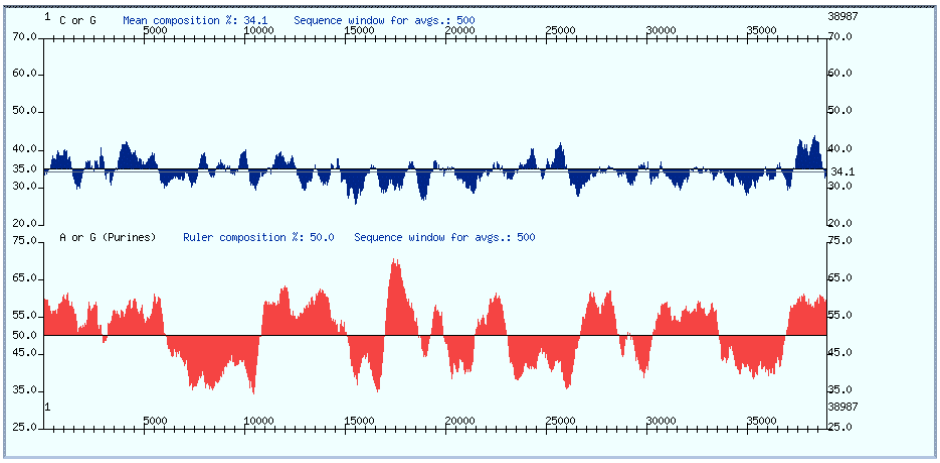


Fig. 16.9 Base Compositions. Average A + G and G + C compositions are calculated using a sliding window of 500 bases. The moving average is displayed as a filled graph, both above and below the centerline average. Unusual G + C might indicate the presence of transposable elements. Majority A + G indicates coding strand in many organisms.

As previously mentioned, the red purine (A + G) composition graph can be used for many species to predict the strand containing the coding sequence. Lined up against Fig. 16.3, ORFs that are most probably noncoding can be detected. When the purine composition is greater than 50 %, the coding ORFs are probably on the positive strand. They are likely to be on the negative strand when the composition is much below 50 %.

The composition percentages are smoothed out by calculating averages with a sliding window of 500 base pairs. When each pixel represents 50 base pairs on the scale, and the window for composition averaging is 500, we can use the previous five pixels' totals (50/pixel × 5 pixels = 250 bases) and the next five pixels' totals for each plotted pixel column to calculate the current average. At each pixel column, we add a new total and discard the total for the first pixel column in the sliding window. Calculating the average at any location requires only averaging 10 numbers. Otherwise, in our example, 500 would need to be averaged at the sequence ends where the look-ahead and memory about the values for previous columns do not exist. To avoid invading white space, the average peak is truncated when it is outside the user-defined scale ranges.

16.7.2 Sequence Repeats

Figure 16.10 indicates the portions of the sequence that are repeated in the project. MAGPIE calculates families of repeated sequences sharing a minimum number of contiguous bases. By default, the minimum number is twenty. Repeats are sorted into families of matching subsequences, which are further sorted by size in descending order. The matching sequence name is in the left right hand column whereas the location and size of the match are displayed as filled boxes under the ruler. Red, blue, and green boxes represent forward, reverse complement, and complement matches.

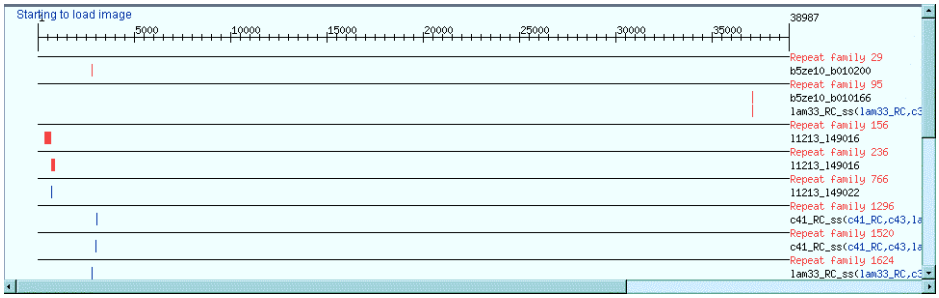


Fig. 16.10 Sequence Repeats. The image has scrollbars (as part of an applet) because of its large dimensions. Repeats are sorted into families of matching subsequences, which are further sorted by size in descending order. The matching sequence name is in the left right hand column, while the location and size of the match are displayed as filled boxes under the ruler. Red, blue, and green boxes represent forward, reverse complement, and complement matches.

When many repeats occur in a sequence, the images are very tall. In these instances, the dimensions might exceed the maximum image dimensions that can be shown by the browser. The stored image is loaded into a Java applet with scroll bars to overcome this browser limitation. The repeats finder is a wrapper around the MEGABLAST program. This wrapper exports the repeats information to enable specialized viewing of the data by Java applets accepting a special data format.

16.7.3 Sequence Ambiguities

Figure 16.11 displays the location of ambiguous bases in a contig. This image can be used in the polishing stage of the DNA sequencing project. It is usually viewed atop Fig. 16.12, which provides the assembly context. The generation of this image requires assembly information from the Staden package [13].

The scale in the center denotes the number of base pairs in the sequence. Red vertical ticks on the upper line represent ambiguities, where there is only positive strand coverage. Blue ticks on the lower bar represent ambiguities when only the negative strand is sequenced. When an ambiguity occurs in a region of double stranded coverage, the tick appears on the center scale. As an exception to the consistent use of color, this tick is colored red for better visibility. The total number of ambiguities is displayed in the lower-left corner. In our example the second pane of the display is shown.

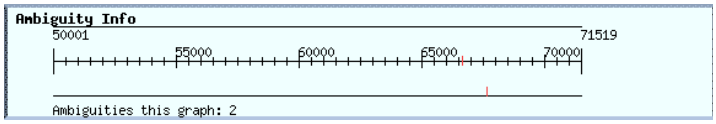


Fig. 16.11 Ambiguity Information. This graph starts at base 50,001 because it is the second pane for a 71,519 base sequence limited to 50,000 bases per image. If the ambiguous base has been sequenced on the forward or reverse strand, or both, the tick is displayed on the top, bottom, or center line. The ambiguities total displayed is for the pane, not the sequence as a whole.

The number of ambiguities (two) is valid for this pane only; there might be more ambiguities in the first pane.

16.7.4 Sequence Strand Assembly Coverage

The image shown in Fig. 16.12 summarizes the quality of the sequence assembly for emerging genomes in MAGPIE. MAGPIE can extract assembly information from the output of the Staden package assembler *gap4*. As in Fig. 16.3, background shading denotes the extent and orientation of the neighboring contigs.

Average coverage multiplicity is the average number of times any base of the contig has been sequenced. This number is calculated separately for each strand. The two values and the total for the contig are displayed at bottom-center of the image. The green area in the center of the image can be interpreted as two separate histograms; that above the ruler is quantifying the average coverage on the positive strand whereas that below the center ruler is quantifying the average coverage on the negative strand. The histogram bars are truncated if greater than tenfold coverage is reached. The histograms are used to determine the reliability of the data. This is useful where frame shifts or miscalled bases are suspected.

Further resolution of poorly covered regions is provided through the continuity of the large horizontal blue and red bars. A gap will appear in the blue or red bar if even a single base pair has not been sequenced on the forward or reverse strand, respectively. This enables the user to see how much DNA sequence polishing is required to double-strand the entire sequence assembly. The displayed overlap with other project sequences is once again useful. Gaps in the current sequence's assembly might be less worrisome in regions overlapping with other contigs.

To create this image the assembly information is read in chunks. The chunk size is equal to the scale factor for the image (50 bases/pixel by default). The base pairs on each strand are mapped onto a blank template string of fifty bases. Blanks in the template after all base pairs are mapped to indicate gaps.

The average sequencing coverage for that pixel is calculated at the same time. The sequence averages are calculated by summing up the pixel totals. This provides major savings over the much larger individual base totals.

16.7.5 Restriction Enzyme Fragmentation

Figure 16.13 displays the location of restriction enzyme cuts on the insert. The MAGPIE user can define the set of restriction enzymes. The cloning vector and the orientation of the insert in the clone can be specified when the contig is added to the

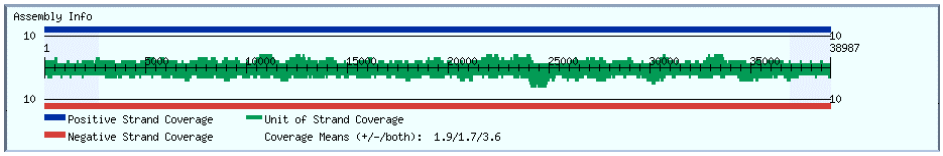


Fig. 16.12 Assembly Information. Histograms of average positive and negative strand assembly coverage are above and below the center line. Breaks in the blue and red bars indicate gaps in the positive and negative strand coverage. Genome neighbors are indicated by background shading as in Fig. 16.3.

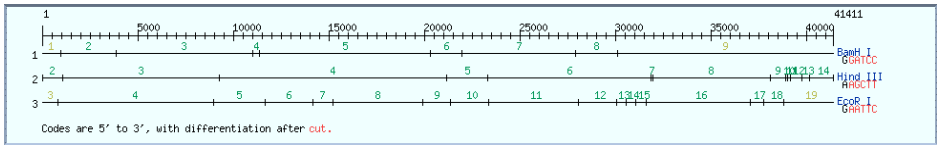


Fig. 16.13 Restriction sites. Ticks correspond to the location of cut sites for the restriction enzymes listed in the right-hand margin. The fragments produced are labeled 5′ to 3′, with the undisplayed cloning vector on the 5′ end. Fragments containing vector have yellow labels, otherwise they are green. The HindIII line has restriction sites at the ends of the sequence because the clone library was HindIII restricted. As a consequence it has no yellow-labeled fragments.

MAGPIE project. This information is taken into account during the fragment calculation. In combination with Fig. 16.14, these provide quality control data for the user by allowing wet lab experimentation to verify the assembly data.

In the figure, cut locations are denoted as vertical ticks on each enzyme's lines. The fragments are numbered from the 5′ end to the 3′ end, including the vector sequence (which is not shown in this display). The vector is always oriented to the 5′ end (i.e. left end) of the insert. Fragment numbers in green represent parts of the vector-free insert. Fragments labeled in yellow contain parts of the vector sequence. Fragments that contain only vector are not displayed in this figure, because the ruler only includes the range of the insert. Such fragments appear only in the agarose gel simulation described below. The numbering of fragments in the Fig. 16.13 display does not always start at number 1, because of the out-of-sight vector fragments. The example clone is from a *Hind III* restricted library. The enzyme cuts the sequence at the very start and at the very end of the insert (i.e. there are no yellow-labeled fragments).

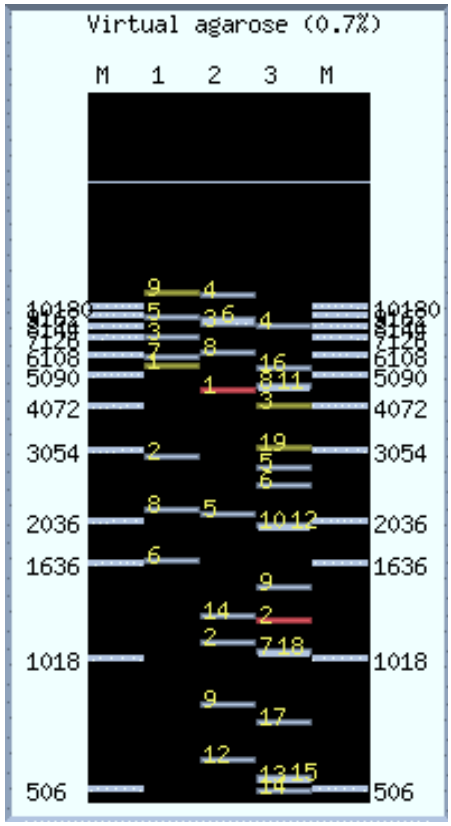


Fig. 16.14 Agarose gel simulation. Fragment lanes and labels correspond to those in Fig. 16.13. The fragment migration is calculated using the specified standard marker "M" lane migrations. Fragments containing vector are colored yellow. Fragments composed of only vector are colored red. Where fragments are very close, luminescence is more intense and labels are offset for readability.

16.7.6 Agarose Gel Simulation

Figure 16.14 is a computer simulation of an agarose gel with the same the restriction di-

gests as Fig. 16.13. The width of the image is based on number of restriction enzymes. The height depends on the user-configurable size of the agarose plate.

Given that we know the theoretical fragment lengths, the hypothetical fragment migration distances are calculated using the reciprocal method [14]. This is used instead of the less accurate logarithmic scale based on the sequence length. The reciprocal method calculates mobility as the inverse of the length. The exact relationship is governed by constants calculated from a least-squares fit of the marker mobility data. The reciprocal formula is:

$$(m - m_0)(L - L_0) = c$$

where m is the mobility in the agarose gel and L the fragment length. The constants from the least-squares fit are m_0, L_0, and c. Like most of the settings for the graphical displays in MAGPIE, the marker mobility data are specified in a configuration file. The regression is performed using the traditional summation method as opposed to the matrix multiplication method. This method was chosen because of the limited efficiency of array manipulation in Perl5.

Describing the features of the gel from top to bottom, the agarose percentage is displayed. This number is specified in the marker mobility configuration. The outside lanes marked "M" are the marker lanes. The inside lanes are numbered left to right in the top to bottom order of the restriction enzymes of Fig. 16.13. The horizontal gray line represents the location of the wells. The fragments in the marker lanes have their lengths displayed in the image margins. The bands of marker lane fragments are solidly colored to ensure they can be clearly observed. The bands in the enzyme lanes appear slightly diffused in order to resemble more closely the appearance of physical gels. These bands are also distinguishable from undifferentiated bands, which are more solid and overall brighter in color. When bands are close to each other in a lane, some labels for fragment numbers are offset in an attempt to maximize readability. As an example of these distinctions, observe fragments 16, 8, and 11 in lane 3 (an *EcoR I* digestion). Fragment 16 is isolated, and has a diffused band and a left justified label. On top of one another, fragments 8 and 11 are given a larger bright area. This highlights their overlap even though they occupy the same amount of space as fragment 16. The label for number 11 is offset to the right of number 8.

Fragments containing or entirely composed of vector sequence are also displayed. This is done to remain true to the physical manifestation. Partial vector fragments are colored yellow. Full vector fragments are colored red. This labeling is done so that fragments containing vector are not accidentally isolated from the real agarose gels. As would be expected, in the *Hind III* lane (number 2) there is a single fragment, number 1, which contains the entire vector.

16.8 The Bluejay Data Visualization System

The MAGPIE images shown in this text are extremely useful as a support tool for genome annotation. Without these images efficient scanning of genome evidence would be much harder and often probably impossible. More than 100 genome sequences have been annotated with MAGPIE to date and we are working on a complete survey of all public genomes.

Despite this, the system cannot satisfy the need for more complex data queries, which will be a theme for at least the next decade. For example, most prokaryotic ge-

nomes are circular in nature, a feature not displayed in any of the MAGPIE graphics. Although the administrator can configure particular features of the MAGPIE graphics, there are severe limits to the flexibility. For example, a graphical display of a query like: "Display all the tRNA coding genes and the tRNA-synthetase coding genes in a genome and show potential relationships between the two gene sets" cannot be served by the static MAGPIE environment. Refinement of the displays and the addition input forms for eukaryotic features [15] in MAGPIE will continue as more analysis and genome annotation is performed, but the technology used for the implementation of MAGPIE limits the system to precomputed graphical outputs with little possibility of variation.

A key to the development of more flexible systems is the use of the extensible markup language (XML) format, addition of Java-based display tools, and incorporation of scalable vector graphics (SVG) into the toolkit. MAGPIE is now capable of exporting XML-formatted information about genome features. This enables the rendering of MAGPIE information by XML-compliant clients, which have more display options, enabling more or less indefinite viewing possibilities for genome data.

The system capable of using XML-formatted information is called Bluejay [16]. Bluejay is a Java-based genome-visualization system which enables the user to customize the types of information displayed, zoom dynamically into the data, and perform queries like that described above. We consider systems like Bluejay to be the next step to a genome display environment that will enable genome researchers of the future to perform the *in silico* biology envisaged for the 21st century.

16.9 Bluejay Architecture

The Bluejay system is designed as a client–server architecture. Any XML-based server, including MAGPIE, and TIGR or GenBank services, can be used as input for the client. In addition a proxy server, installed at the Sun Center of Excellence for Visual Genomics in Calgary enables utilization of non-XML based data sources, which are reformatted to XML by the proxy server.

The Java-based client provides the user interface to the Bluejay system. It is designed with the look and feel of a Web browser, creating an interactive visual model for biological sequences. The display is dynamic and highly customizable, with sequence landmark sites, annotations, and related information available for visual exploration. Most image elements are linked to external bioinformatics sources, for example MAGPIE gene-annotation pages and BIOMOBY services [17]. The visual model of the sequence thus serves as a backbone for mapping other relevant data, which can then be explored in a rich genome context. Bluejay supports several levels of visual manipulation – sequence-wide operations, queries based on selected functional categories, and exploration of individual elements.

One of the best practices in software development is design for change. The Bluejay architecture facilitates this practice through modularity and extensive support of open standards. The development of Bluejay focuses on two tasks – creation of the core Bluejay package, and making it suitable for interaction with publicly available software. External software is either plugged directly into the Bluejay as a component to implement the desired functionality, or accessed remotely using open standards and protocols. The key enabling tech-

nologies for this interaction are Java technology and XML-based standards.

The general information flow in Bluejay is presented in Fig. 16.15.

The browser side receives and visualizes the incoming XML data and parses these into an internal document object model (DOM) tree [18] in which nodes represent nested XML tags. Because these tasks are standard in XML processing, external software can be easily plugged in to perform parsing and creation of the DOM. Bluejay utilizes publicly available Java-based Apache software, such as the Apache Xerces parser and the Apache Batik implementation of the DOM [19]. By supporting the standard DOM interface several existing tools for XML processing can be used for genome data manipulation and retrieval.

After the data have been parsed, the visual model is created as a collection of backbone sequences, for example DNA or protein backbones, relative to which all other visual elements are placed. Bluejay logically separates painting into two stages. The high-level painting stage is abstracted and consists of placement of elements on to a sequence. For example, high-level painting of a gene is performed by specifying its positions on a DNA sequence. Note that sequence positions are intrinsic to the data and remain constant as the user manipulates the visual model. This process therefore has the same interface irrespective of the shape and relative position of the sequence backbone. The abstraction of the high-level painting makes Bluejay architecture more robust and easy to extend to new data types. The low-level painting stage follows, by using Java graphics to render images onto the canvas. The Java object representing the DNA sequence translates genome base pair positions into canvas coordinates.

Bluejay graphics is based on scalable vector graphics (SVG), a cross-disciplinary XML standard for imagery. This capability is provided by Apache Batik, a Java-based SVG visualization package [19]. In essence, the Java painting classes in Batik are adapted to create SVG tags in response to the usual drawing requests. The resulting SVG document is an explicit and portable XML representation of the visual model. An SVG-enabled canvas visualizes this model and enables the user to interact with it; Bluejay translates this into actions on the original sequence data. The canvas is plugged into the Bluejay main browser and connected to

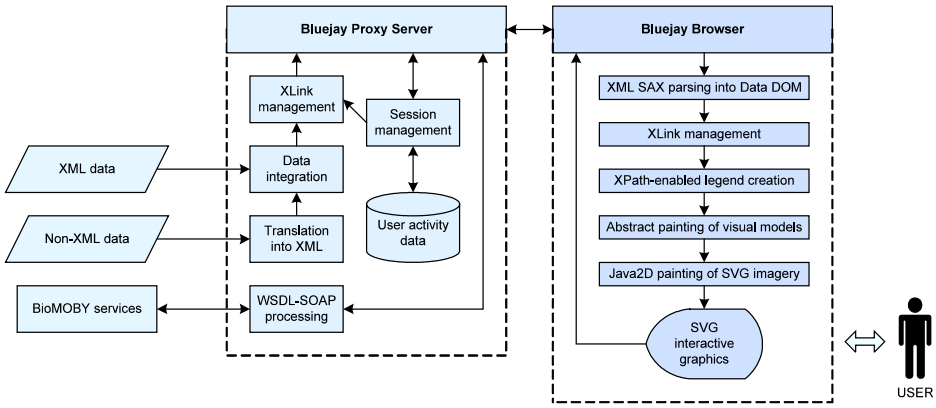


Fig. 16.15 Bluejay information flowchart and main components.

the data DOM tree. Data access is based on XLink, XPointer, and XPath standards for XML [18].

16.10 Bluejay Display and Data Exploration

When using bioinformatics tools in practice software complexity is a prominent issue. Ease of use is therefore an important design consideration. In Bluejay, data manipulation is entirely visual. No scripting, configuration files, or command line parameters are required to run the software and enjoy a fully enabled interaction. Most data manipulation is also visual and uses well-understood GUI idioms. In the paragraphs below we present some of the display features available to users in the current version of the browser. Bluejay is an actively developed project, therefore data types and visual data analysis capabilities are regularly being added.

16.10.1 The Main Bluejay Interface

Figure 16.16 shows the Bluejay window with its main components – the SVG-enabled interactive display, the context tree, and the interactive image legend. A full view of an *Escherichia coli* K12 genome is presented here. When a genome sequence is first loaded, it is shown at the highest level of detail. For example, the view in the figure displays only the main features on the genome. The visual model represents the DNA in a six open reading frame mode, with genes and other features identified pri-

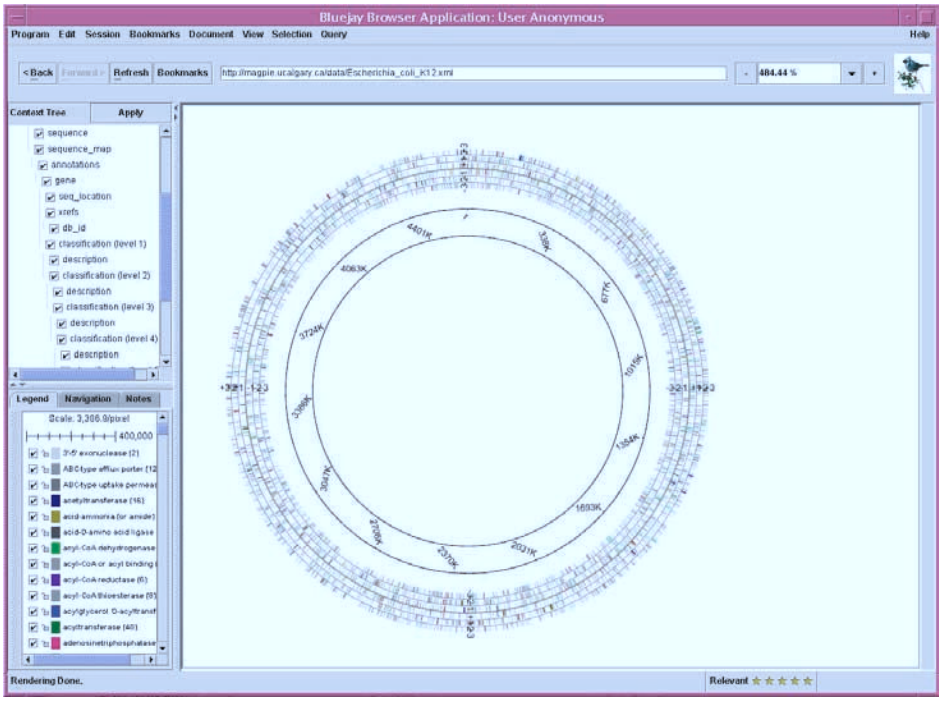


Fig. 16.16 Bluejay interface with a complete *Escherichia coli* K12 genome.

marily by colors. The color scheme represents the available gene ontology-based functional categories, listed in the legend. The legend itself is not only a very easy idiom to understand, but is also interactive – entire functional categories of genes can be queried by manipulating the corresponding legend items.

Another manipulation tool is the context tree, also shown. It summarizes the structure of the XML data, which is essentially a nesting of XML tags. The context tree is also interactive and enables the manipulation of entire data types. In this case, however, the data types are defined by the hierarchy of XML elements. Therefore, the legend and the context tree are two complementary ways to query the data – one based on a biological ontology the other on the specific XML schema used to represent the data. As an example, the same XML markup element <gene> can represent genes from a variety of functional categories; on the other hand, a gene ontology category "enzymes" might be represented differently in different XML schemas.

16.10.2 Semantic Zoom and Levels of Details

In visualization based on a scalable vector graphics, the ability to scale the image to any zoom level is a key feature. Bluejay also takes advantage of this – the zooming range in Bluejay is virtually unlimited. The zoom level can be chosen visually, either by selecting a region of interest with a mouse or by using global mouse-based scaling, both methods are provided by the Batik SVG visualization package. Alternatively, a desired zoom factor can be manually typed into the zooming widget at the top of the Bluejay screen.

But Bluejay goes much further by offering a semantic zoom, which is becoming a required feature for genome browsers [20]. Without the semantic zoom, all a user can get, e.g. by zooming in to a DNA image, is an enlarged copy of the same image, which is rather meaningless for visual data exploration. Instead, Bluejay internally translates the new SVG zoom scale into a higher level of detail, which enables seamless visualization of finer, previously unseen features of the DNA sequence. Figure 16.17 shows a region of interest on the *E. coli* K12 genome at a higher zoom level. This zooming resulted in a transition to the more detailed view, and a number of new genome features are now available for display, such as annotations for individual genes, (A + G) % and (G + C) % levels.

The level of granularity can also be defined by choosing a desired view mode. These range from a pie-chart summary of available functional categories, to two-strand view of the DNA, to six open reading frame view, to a text view of the sequence showing individual base pairs. As the viewer switches between different view modes, Bluejay maintains consistent legend color scheme, visibility settings for data types, navigation settings, and other features.

16.10.3 Operations on the Sequence

Bluejay enables a range of visual manipulations of the whole genome. Sequence-wide operations include switching between linear and circular models of the sequence (in bacterial genomes), switching between normal and reverse-complement views of a double-stranded sequence, rotation by a desired angle, and "cutting" at a specified start position, e.g. at the beginning of a gene. The operations are performed visually using specialized GUI tool panels, which can be brought up through the menu.

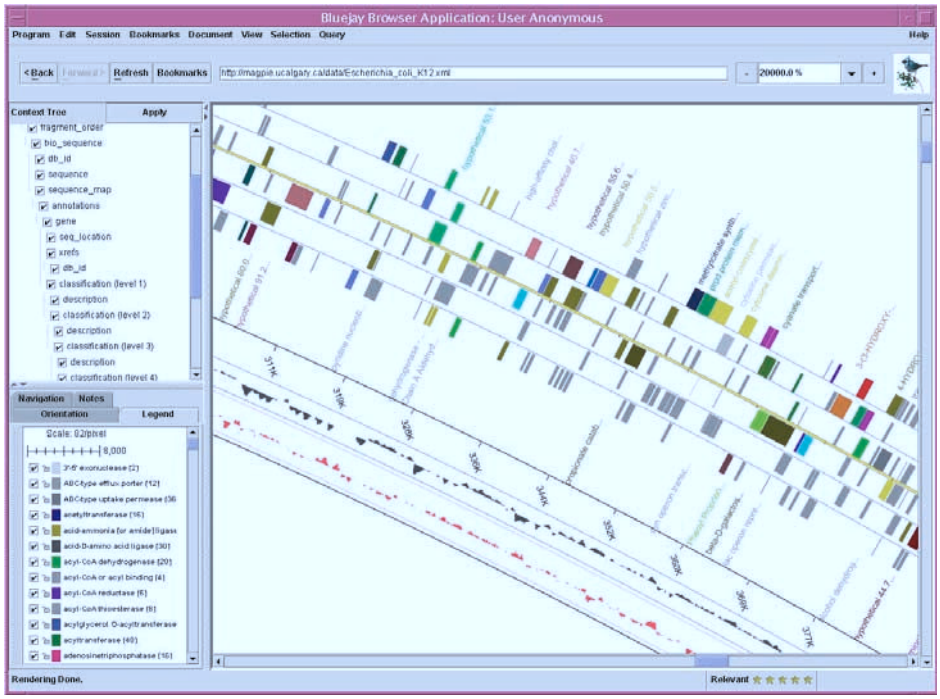


Fig. 16.17 A zoomed-in fragment of the E. coli genome. Semantic zoom automatically reveals finer details, for example functional annotation labels for genes, A + G and G + C percent composition levels.

An interactive legend enables operations on specific functional categories. For example, the user may click on the legend item "Enzymes" and either ghost out or completely remove all enzyme-producing genes from the visual model of a DNA. Internally, clicking on a legend item issues a customized XPath query to the XML DOM data, which then retrieves XML elements that fit the query description. A required visual operation then follows, performed on the retrieved data elements. To implement XPath access to DOM, an existing XPath module from the Apache Xalan–Java project [19] was plugged into Bluejay. The availability of the functional categories also depends on the current settings of the context tree – for example, the user might choose only the top levels of gene ontology hierarchy represented in the XML data. In this case legend items will be grouped in broader categories, thus making the legend list shorter.

Figure 16.18 shows the same region of the *E. coli* K12 genome in reverse complement view. Only the top gene ontology level is displayed, selected in the context tree. The legend is now much shorter. Some of the functional categories related to enzymes are "ghosted" by clicking the corresponding legend items.

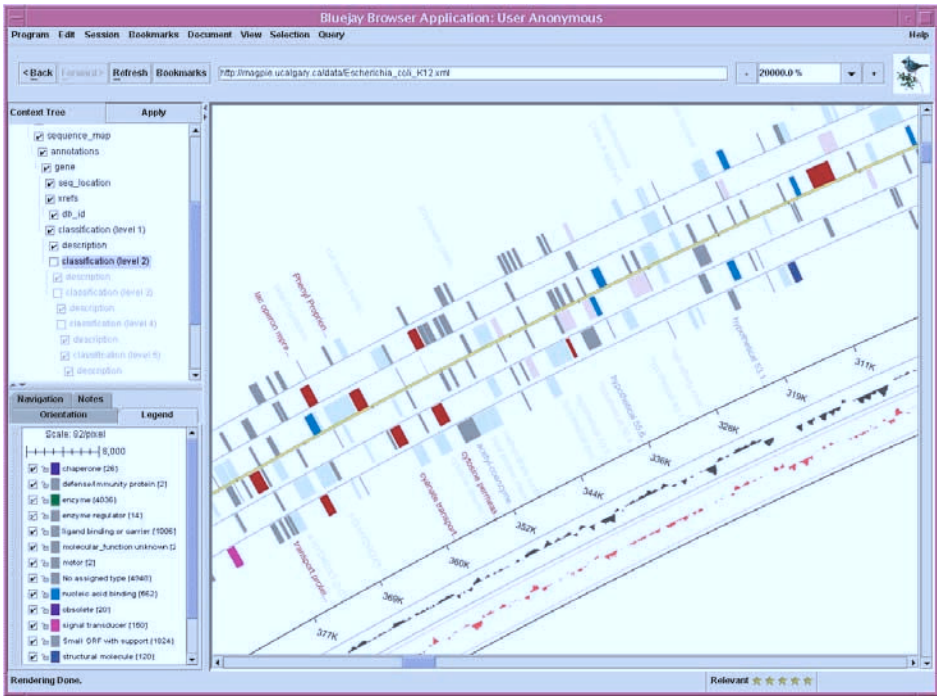


Fig. 16.18 Manipulation of gene ontology functional categories through the interactive legend, where all enzyme-related genes are shown "ghosted".

16.10.4 Interaction with Individual Elements

The user can access any individual element on the sequence through mouse clicks. Bluejay makes an extensive use of the XLink standard for hyperlinks. For example, by clicking on an image of a gene the user can access that gene's data and hyperlinks. In Bluejay, each visual element is typically linked to multiple external resources and services. For example, a link from a gene to MAGPIE launches a web browser window with a MAGPIE annotation page for that gene. A list of links to MOBY services is also presented. When a user clicks on a gene, a query is issued to MOBY Central, which provides the necessary data about the gene, e.g. its accession number or the gene ontology category ID. The MOBY central returns a collection of currently available MOBY services that are relevant to the information provided. These are immediately displayed in the list of other links, and the user can launch a desired MOBY service by a mouse click. Bluejay then queries the remote service, retrieves the results in an XML format, and uses XSLT stylesheet transformations [18] to display them as an HTML page. Figure 16.19 shows a sample view of MOBY results.

Another type of hyperlink is an embedded link. For example, only the top view of a eukaryotic chromosome can be present in an XML data file, with embedded links to more detailed information about the genes in the chromosome. This feature enables the browser to save memory and load the

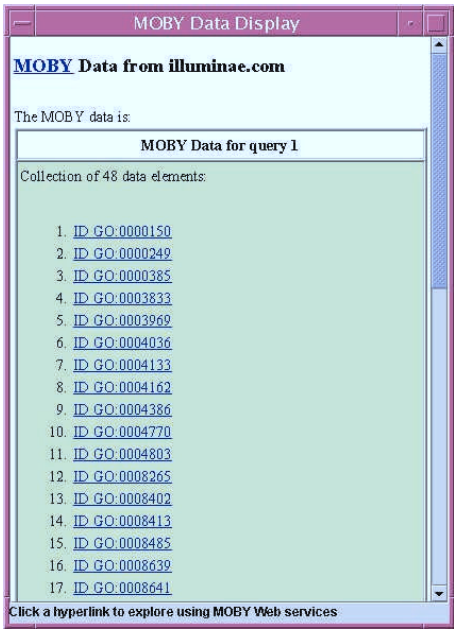


Fig. 16.19 A sample of data returned by a BioMOBY service after a user's visual query from the main Bluejay interface.

data faster. When the user zooms to a higher level of detail, the embedded links are activated (Bluejay supports activation of links both on load and on request). The detailed information is then fetched in by the proxy and inserted into the visual model, so that finer gene features become available for display and analysis.

The XLink standard enables such advanced features as multiple outgoing links originating from the same element, inbound and third-party links, and linkbase libraries, which are external resource files that define collections of links. Some of this behavior is mandated by the XLink standard while other behavior is application-specific. Bluejay recognizes, extracts, and activates XLinks from the incoming data to other genome data and web services. Architecturally the XLink module acts as an optional SAX-compliant filter on top of a SAX-compliant XML parser (for example Apache Xerces) and extracts all relevant XLinks into a link pool [21]. Bluejay's internal structures then handle dynamic link recognition and activation. The uses of XLinks in Bluejay will probably be expanded to the tasks of visual integration of data and services.

16.10.5 Eukaryotic Genomes

Visual manipulation of eukaryotic genomes is very similar to that for prokaryotic genomes, but the visual model becomes more complex. The eukaryotic display depends on the specific XML scheme, therefore Bluejay contains specialized Java painter classes for each supported scheme. Figure 16.20 shows the visual display of the *Arabidopsis thaliana* genome in Bluejay, encoded in TIGR XML. This XML format defines a range of XML elements to describe the eukaryotic genome structure. A transcriptional unit (TU) is a top element colored according to the functional category. It contains a gene MODEL element (blue outline) that describes coding and noncoding structures of a splicing isoform. The MODEL encodes several mRNA EXONS (shown in black). Protein coding portions of exons (CDS, outlined in green) and untranslated regions (UTR, outlined in red) are also shown. Bluejay display follows the general structure of the TIGR XML format specifications but enhances the view with color-coded information and interactive behavior of individual elements.

16.11 Bluejay Usability Features

Usability is an important design issue in bioinformatics, and several usability features in Bluejay enhance interaction with the

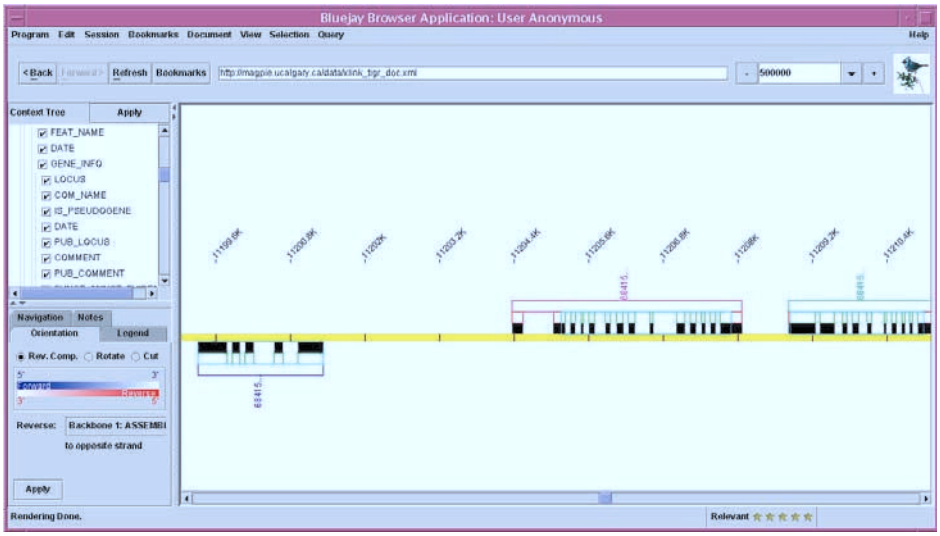


Fig. 16.20 A fragment of the eukaryotic genome, for example that of *Arabidopsis thaliana*, displayed in Bluejay.

browser. Bluejay supports a growing range of commonly used XML standards for biology. Already implemented are BioML, GenBank XML, TIGR XML, AGAVE XML, and support for new formats is added, enabling users to visualize and manipulate external data. This contrasts favorably with many other existing browsers created solely to showcase specific datasets.

Bluejay is written in Java, and is therefore portable to most existing platforms. Aside from having a current version of Java, Bluejay does not require any additional configuration or maintenance. This simplicity is one of the key usability features of Bluejay – life scientists are typically well familiar with the Java environment but are averse to dealing with less known or more complicated computer technology. Bluejay is available in two versions – a VeriSign-certified secure applet and a downloadable standalone application. Unlike regular web applets, the secure applet enables local saving and uploading of data files. The applet is especially suitable for new users who can learn most of the features simply by running it on the Bluejay website. Downloading the application version, packed as a JAR (Java archive) file, could be the next step. The application can utilize more system resources, which is helpful when dealing with large genomes. Both the applet and the application versions of Bluejay have an important advantage – many other genome browsers are web-based and thus have to communicate with a remote web server for each visual operation. In contrast, the Bluejay browser runs on a local Java virtual machine on the user's system and is not subject to communication latency.

Bluejay provides session management, whereby users may return to their previous data exploration sessions, possibly from a different computer environment. Bluejay also allows users to save visualization results as scalable vector graphics files, which enables the images to retain their full publication quality irrespective of the scale, for

example, preserving finest details in wall-size posters of genomes.

16.12 Conclusions and Open Issues

The analysis and annotation of genomes has progressed so much over the last ten years that it is almost time to put this issue to rest. Many tools exist which are capable of handling the task, and their flexibility, as shown in the Bluejay example, is enabling execution of complex queries and the display the results in many different ways.

The next step is to connect genome information to gene expression, Proteomics and other large scale life sciences results. The better analysis engines for this kind of experiment recognize the value of genome annotation and connect to the existing knowledge base. The task for the next few years will be to fully integrate the entire knowledge space and provide meaningful visual representations of the results.

References

1 Gaasterland T., Sensen C.W. (1996) Fully automated genome analysis that reflects user needs and preferences – a detailed introduction to the MAGPIE system architecture. *Biochimie* 78, 302–310.

2 Charlebois R.L., Gaasterland T., Ragan M.A., Doolittle W.F., Sensen C.W. (1996) The *Sulfolobus solfataricus* P2 genome project. *FEBS Letters* 389, 88–91.

3 Pearson W.R., Lipman D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* 85, 2444–2448.

4 Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

5 Henikoff S., Henikoff J.G., Pietrokovski S. (1999) Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15, 471–479.

6 Kolakowski L.F. Jr., Leunissen J.A.M., Smith J.E. (1992) ProSearch: fast searching of protein sequences with regular expression patterns related to protein structure and function. *Biotechniques* 13, 919–921.

7 Burge C., Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94.

8 Salzberg S., Delcher A., Kasif S., White O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 26, 544–548.

9 Borodovsky M., McIninch J. (1993) GeneMark: Parallel Gene Recognition for both DNA Strands. *Comput. Chem.* 17, 123–133.

10 Cooper A. (1995) About Face: The Essentials of User Interface Design. IDG Books Worldwide, Foster City, CA, USA.

11 Shine J., Dalgarno L. (1974) The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl Acad. Sci. USA*, 71, 1342–1346.

12 Hoare C.A.R. (1962) Quicksort. *Comput. J.* 5, 10–15.

13 Staden R., Beal K.F., Bonfield J.K. (1998) The Staden Package, 1998. *Comput. Methods Mol. Biol.* 132, 115–130.

14 Schaffer H.E., Sederoff R.R. (1981) Least squares fit of DNA fragment length to gel mobility. *Anal. Biochem.* 115, 113–122.

15 Gaasterland T., Sczyrba A., Thomas E., Aytekin-Kurban G., Gordon P., Sensen C.W (2000) MAGPIE/EGRET Annotation of the 2.9 Mb *Drosophila melanogaster* ADH Region. *Genome Res.* 10, 502–510.

16 Gordon P., Sensen C.W. (1999) Bluejay: A Browser for Linear Units in Java. *Proc. 13th Annu. Int. Symp. High Performance Computing Systems and Applications*, pp. 183–194.

17 Wilkinson MD, Links M (2002) BioBIOMOBY: an open source biological web services proposal. Brief Bioinform. Dec 3(4):331–341.

18 World Wide Web Consortium (W3C), http://www.w3.org.

19 Apache Software Foundation, http://www.apache.org.

20 Loraine AE, Helt GA. (2002) Visualizing the genome: techniques for presenting human genome data and annotations. BMC Bioinformatics, 3(1):19.

21 Simple API for XML, http://www.saxproject.org.

17
Bioinformatics Tools for Gene-expression Studies

Greg Finak, Michael Hallett, Morag Park, and François Pepin

17.1 Introduction

Microarrays have certainly generated a tremendous amount of excitement over the past decade. The ability to interrogate expression levels on a genome-wide scale has opened up broad new experimental possibilities. The noise and bias seemingly inherent to gene expression data, the large quantity of data, and the difficulties associated with interpreting results have necessitated improvements in the underlying biotechnology, statistical methods, and computational tools. Some understanding of each of these areas is necessary for the effective design and performance of experiments involving microarrays. It is hard to believe there are many researchers with, *a priori*, a deep understanding of so many diverse fields.

The purpose of this chapter is to provide an intuitive description of the major problems and challenges that must be addressed throughout the course of gene expression analyses. Our exposition focuses on existing bioinformatics tools and strategies that the reader can explore in more detail independently. In lieu of providing a complete bibliography of related work, we focus on a few important references we feel are most useful for developing the skills and computational infrastructure necessary for research in this area. With the proper background, these references should be sufficient for the reader to gain a working knowledge of these frameworks.

We begin by providing a sketch of microarray technology. Beyond this intuition however, we do not explore the underlying biotechnology, construction, or laboratory procedures of microarrays – we refer the reader elsewhere [1]. The organization of the rest of the chapter follows the progress of a typical microarray project. We discuss issues concerning experimental design and explain the sources of error in experiments. Section 17.2 gives an outline of microarray standards, commonly used database systems, and general-purpose analysis packages. Particularly for large-scale microarray projects, these tools form the "backbone" of the computational infrastructure. Section 17.3 covers many of the fundamental problems and solutions associated with the data-acquisition phase of individual microarray hybridization and the normalization of the data. Readers lacking a basic knowledge of

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

statistics are, again, referred elsewhere [2]. Section 17.4 discusses the fundamental problem of reliably finding differentially expressed genes between two or more classes. Sections 17.5, 17.6, and 17.7 cover in detail many of the common questions, approaches, and tools used to analyze gene-expression data. Readers who are interested in techniques from statistical machine learning are referred to Hastie et al. [3].

17.1.1 Microarray Technologies

There are many types of microarrays including DNA and protein arrays. For gene expression studies the most commonly used technology is the DNA microarray. DNA microarrays (referred to as simply *microarrays* or *arrays*) are assays for estimating the amount of different species of mRNA transcripts present in a sample. The underlying assumption of this technology is that the number of mRNA transcripts present is a good approximation of the level of expression of the corresponding gene. Typically, in a microarray experiment total RNA is extracted from the specimen and isolated. The mRNA is then reverse transcribed to a form of labeled DNA referred to as a *target*.

A microarray consists of a solid surface to which strands of DNA are attached in specific, pre-defined *features*. Microarrays can contain hundreds of thousands of such features. Each strand of DNA is referred to as a *probe* and consists of either short oligonucleotides or (full length) cDNA. Millions of copies of one *species* of probe are placed in a single feature. The ordered nature of the microarray enables us to associate a feature with a known species of mRNA from the specimen.

Under appropriate experimental conditions, the targets hybridize to probes on the array when the target and probe share sufficient sequence complementarity. This is referred to as *hybridization*. Ideally, probes are chosen so that exactly one target from the mRNA of the specimen will hybridize to that probe and so that the strength of the target–probe hybridization is sufficiently strong. After this hybridization, each location on the microarray should contain probes that have hybridized with their corresponding targets and the number of such probe–target hybrids should be proportional to the level of expression of the gene represented by that probe. Because targets are labeled (that is, they fluoresce), the intensity of each feature can be measured and the resulting value provides an estimate of expression for the associated gene.

Each step in this procedure is extremely important and crucial to the success of an experiment. Procedures change according to the nature and goals of a particular experiment [1] and are very dependent on the specific microarray platform in use. The two main microarray platforms are cDNA microarrays and high-density oligonucleotide microarrays (for example, Affymetrix GeneChips). We describe each of these briefly.

17.1.1.1 cDNA Microarrays

With cDNA microarrays, mRNA is reverse-transcribed to complementary DNA (cDNA) and then labeled fluorescently. cDNA microarrays consist of DNA probes robotically printed on a glass slide in features (termed *spots*). Robotic printers have multiple pins that "spot" the glass slide simultaneously. This effectively partitions the slide into a set of rectangular grids, each serviced by one pin. Each such partition is referred to as a *pin group*. cDNA microarray hybridizations are almost always *dual-label* hybridizations.

In other words, two cDNA samples are labeled with different fluorescent dyes. These are typically Cy3 (green) and Cy5 (red). For

example, the two samples might correspond to mRNA from healthy tissue and mRNA from a tumor. Both samples are simultaneously applied to the glass slide. The differentially labeled targets competitively hybridize to their corresponding probes on the array. The intensities of both the red and green *channels* are read independently and a red to green ratio (R/G) is produced for each probe. The underlying assumption is that this ratio is proportional to the relative amount of each gene present in the red and green samples.

17.1.1.2

Oligonucleotide Microarrays

These microarrays use oligonucleotide probes that are synthesized on a silicon chip by lithography. Samples are washed over the array during the hybridization step. Probes are short oligonucleotides of length 25 bp. Short probes tend to have a smaller hybridization potential with targets when compared with the hybridization potential of the longer probes used in cDNA array platforms. To compensate for poor binding affinity and potential cross-hybridization between probes, Affymetrix uses a so-called *probe pair*. The first element of this probe pair is the *perfect match* oligonucleotide that corresponds to a complementary 25-mer in the exon of a gene. The second element is the *mismatch* oligonucleotide. This is identical to the perfect match probe with the exception of the middle (13th) base. Intuitively, the purpose of the mismatch probe is to control experimental variability due to nonspecific binding of mRNA from other parts of the genome. Each feature (termed a *cell*) of an oligonucleotide array contains millions of copies of a single species of probe. Whereas it is common in most cDNA microarray platforms to represent each gene by one probe, oligonucleotide microarrays use between 11 and 16 probes to represent each gene (Fig. 17.1). (Therefore, a set of 11 to 16 cells is dedicated to one specific gene.) Last, whereas cDNA microarray hybridizations are dual-labeled, some oligonucleotide arrays are single-labeled. In other words, a single sample is hybridized against one oligonucleotide array at a time and the intensity of fluorescence of the probes is an absolute measure of the amount of a target present.

There are several other microarray platforms, for example that of Agilent Technologies (Palo Alto, CA, USA), which combine aspects of cDNA and oligonucleotide arrays. In the Agilent system oligonucleotides are spotted on a glass slide by ink jet technology but remain dual-label systems. Occasionally the software and analysis strategies discussed throughout this chapter can be used in these alternative settings with little or no modification. One should, however, be extremely careful and not assume this to be true without experimental evidence.

17.1.2

Objectives and Experimental Design

The number of different applications of microarrays is growing rapidly. *Gene profiling* via microarrays involves determining the function of genes under specific conditions. Similarly, microarrays have enabled genome-wide *class comparisons*. Several studies in the literature have identified genes that are consistently differentially expressed between two or more classes. There are several classic examples for various types of cancer in which comparison is between normal versus tumor versus metastatic specimens. Such approaches hope to find the complete set of genes that differentiate between cellular states and shed light on the underlying differences between these classes at the molecular level. *Class predic-*

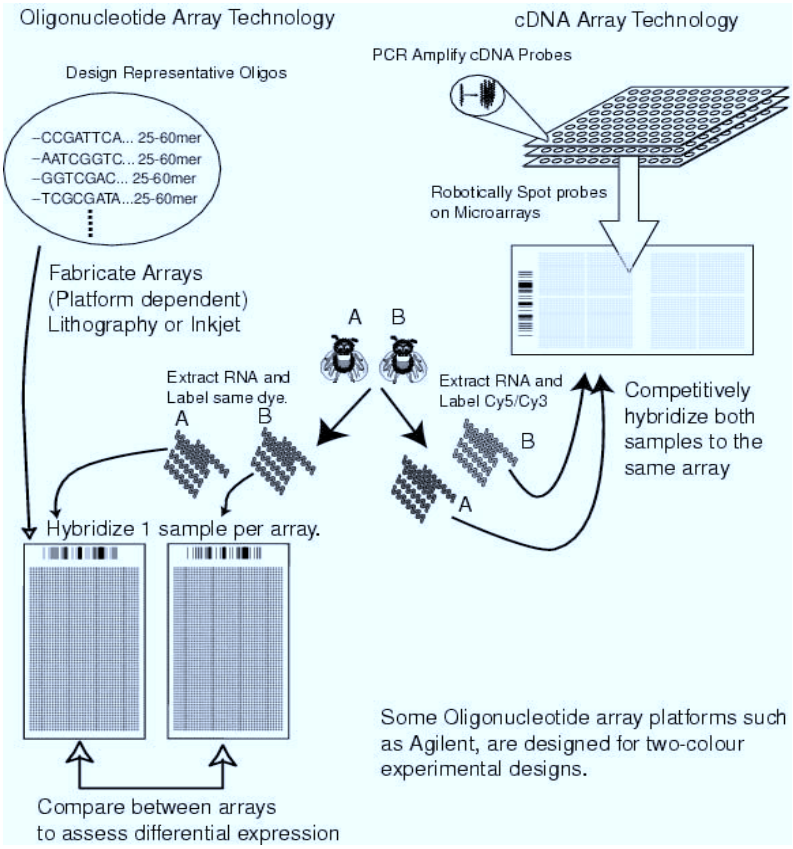


Fig. 17.1 Diagram depicting the two major types of microarray technology – oligonucleotide and cDNA arrays.

tion experiments aim to find sets of genes that act as good "markers" for a particular class. The ability to predict with high probability the subclass of cancer for a patient at an early stage has clear social and economic benefits. *Class discovery* experiments attempt to determine biologically relevant subclasses of a particular cellular state. For example, the goal of such an experiment might be to classify different mRNA responses of individuals within a population to a particular type of cancer. A more complete description of the uses of microarrays is given elsewhere [4].

Not surprisingly, the types of bioinformatics strategies and tools required for a project are strongly affected by the basic objectives and general hypotheses of the research program. When cDNA microarrays are used and the project entails many subjects, it is necessary to decide on an *experimental design*. For example, because cDNA microarrays are dual-label systems, each hybridization requires comparison of two RNA samples. The expression profiles of all subjects often need to be compared with each other. When the number of samples is large, it might be infeasible, extremely expensive, or simply a waste of resources to

perform hybridizations comparing every pair of subjects directly. Several alternative designs have been proposed in the literature including the *reference design*, the *balanced block design,* and the *loop design*. This topic is discussed in more detail elsewhere [5].

The most challenging design phase decisions are due to the sources of variation. These include but are not limited to:

1. intra-gene variation (differences between intensities of probes representing the same gene);
2. intra-array variation (systematic differences between intensities of probes representing different genes on the microarray);
3. inter-sample variation (different expression levels for a gene over different RNA samples from the same specimen);
4. inter-specimen variation (different expression levels for a gene over different specimens from the same subject); and
5. inter-subject (different expression levels for a gene among subjects within the same class).

Sources 1–3 are usually considered different forms of *technical variability* and can be addressed by performing additional hybridizations, consistent laboratory procedures, and via the normalization tools described below. Several strategies are used to estimate the number of replicates. One particularly useful tool is the S-PLUS implementation [6]. Further information about these statistical frameworks is available [7]. Sources 4 and 5 are different forms of *biological variability* and are considered more difficult to address. We note that at potentially every step of a microarray experiment (from the manufacture of the array itself through to the image analysis necessary to extract intensities) bias is introduced into our measurements. For example, in cDNA arrays bias arises because of differences between physical properties such as the rate of incorporation and rate of decay of the dye. Tools and strategies to address issues of bias are discussed below.

17.2 Background Knowledge and Tools

This section describes the basic computational infrastructure and tools needed for microarray experiments.

17.2.1 Standards

The multitude of different microarray platforms has given rise to a series of inter-operability problems. The basic question here is how to describe gene expression data so that researchers (and software) can share and compare this information seamlessly. A step in the right direction is the *minimum information about a microarray experiment (MIAME)* standard [8]. The purpose of MIAME is to state explicitly all relevant information required to unambiguously define and reproduce a microarray experiment. Several journals are now requiring that microarray data be made available in this standard before articles presenting the data are allowed to appear. MIAME is built by the *microarray gene expression data (MGED,* http://www.mged.org) society and is based on their microarray ontology.

Although MIAME describes the information which needs to be documented, it does not standardize the format of this description. The *microarray gene expression markup language (MAGE-ML)* [9] addresses this need. MAGE-ML is based on XML and is a standard way of representing information related to microarrays. For example, if details regarding an expression experiment

are expressed in MAGE-ML, they can be directly imported into databases such as ArrayExpress (EBI), GEO (NCBI), and yMGV.

The *microarray gene expression object model (MAGE-OM)* provides an object model representation of microarray expression data that facilitates the exchange of microarray information between different software systems and organizations. Objects can be translated automatically to and from the MAGE-ML data format. A list of microarray repositories can be found at http://www.nslij-genetics.org/.

17.2.2 Microarray Data Management Systems

Because of the size and complexity of microarray data, they are typically stored in a relational database. Commercial databases can be quite expensive but for many projects open source efforts such as MySQL and postgresql suffice. Projects such as BioPerl (http://www.bioperl.org), RmySQL, and general languages such as Python and Perl can be used to transfer data between the database and analysis software.

There are several publicly available and MIAME-compliant databases tailored for microarrays including *bioarray software environment (BASE)* [10]. In addition to a relational database for microarray data and a web-based graphical user interface that can be installed locally, BASE includes a *laboratory information management system (LIMS)*. LIMS enable individual hybridizations to be tracked through all phases from sample preparation to analysis of data. When a project involves many hybridizations and many individuals, LIMS become invaluable. The web site http://genome-www5.stanford.edu/resources/regtech.shtml maintains an up-to-date list of additional efforts in this direction.

17.2.3 Statistical and General Analysis Software

A cornerstone to any microarray project is a set of tools to perform general statistical analyses (for example, *t*-tests, filtering) and more general microarray analyses (for example, clustering, time series analyses). Many open source systems address these needs. We highly recommend the statistical programming language R [11] (http://cran.r-project.org), an open source version of the commercial package S-Plus. Although there is a substantial learning overhead, fluency in R gives researchers a high-quality and robust set of statistical and graphical packages that can be tailored to specific needs, and, moreover, the Bioconductor project is developed in R. The mandate of the project is to provide a complete toolbox for analysis and comprehension of genome data. The project encourages users to build upon existing modules and make the enhanced functionality available to the community. Currently, release 1.5 of BioConductor contains some 96 packages addressing issues from microarray preprocessing and normalization for several different microarray platforms to techniques for finding differentially expressed genes, visualization classification, and class prediction. Subprojects aim to produce graphical user interfaces for various packages within R and Bioconductor.

For researchers who do not want to invest the time learning R or who do not need the flexibility that it offers, there are several general-purpose microarray analysis packages with well-developed graphical user interfaces. For example, open source packages such as ArrayViewer (http://www.tigr.org/software/), TM4 (http://www.tigr.org/software/tm4/), and MAExplorer (http://maexplorer.sourceforge.net/) provide func-

tionality for normalization/clustering, and support tailored database access. It is beyond the scope of this chapter to cover all the commercial packages available. It is important to note, however, that there is much variance in the quality and completeness of functionality offered by these systems, and not all systems provide flexible methods for communicating with a database system. This is a particularly salient point if microarray projects involve many hybridizations and other types of data beyond gene-expression measurements.

17.3 Preprocessing

This section describes the "low-level" techniques and software associated with the data-acquisition phase of a microarray experiment. The first step – quality control – is arguably the most important aspect of a successful microarray experiment. There are three principal areas – image quality, spot quality, and array quality. The next step is to compute a summary of the expression level of a gene when multiple probes are used. Last, arrays must be normalized to remove bias and to enable inter-array comparison.

17.3.1 Image, Spot, and Array Quality

The quality of the image is judged by the presence/absence of many sources of noise including imaging artifacts, dust, scratches, or misprinted features on the original array. Given the importance of quality control, any microarray analysis should begin with visual inspection of the scanned image to assess overall quality. Most statistical packages provide routines to visualize log transformed expression values.

The evaluation of spot quality involves several steps including (1) *gridding*, (2) *segmentation*, (3) *feature extraction*, and (4) *background correction*. Gridding involves the localization of features in the image. Many commercial image-analysis packages automate the process by locating features on the array provided by the manufacturer that specify the position, diameter, and distance between the features on the array (this makes the software microarray platform-dependent). Segmentation involves discerning background regions from foreground regions. Feature extraction is the process of summarizing the pixel intensities for each feature on the array and converting these intensities to measures of RNA abundance. This process usually considers a ratio of foreground to (local) background intensities. Depending on the platform, the methodology of background correction varies.

Array-level quality control is aimed at identifying arrays that have hybridization, labeling, manufacturing, or scanning problems. The final estimates of quality are largely a function of the number of problems identified with the array. Some platforms, however, for example Affymetrix, contain spiked controls with known levels of concentration providing an additional means of judging quality.

Many packages are available for performing feature extraction from cDNA arrays; the choice of software can significantly affect results [12]. A well-maintained list of tools is located at http://www.genomicshome.com. With oligonucleotide arrays such as the Affymetrix GeneChip platform, image analysis is performed by Affymetrix MicroArray Suite (MAS) software. Increasing efforts outside Affymetrix are being made to address these problems [13] and several packages are available in R/Bioconductor. Agilent's software (http://www.

chem.agilent.com) is quite complete with solid outlier detection schemes.

17.3.2 Gene Level Summaries

When multiple probes are used to represent the same gene, it is necessary to convert the individual intensity measurements into one expression level for the gene. This is especially true for the Affymetrix GeneChip platform. The expression summary for GeneChip arrays contained in the MAS package has evolved over recent years. Several summary measures of expression have been proposed including average difference, Tukey biweight, model based expression index (MBEI), and robust multichip average (RMA) [13]. MBEI (available in dChip, [14]) uses both perfect match and mismatch probes, and models a baseline probe pair response with a multiplicative increase because of the mismatch probes, plus a multiplicative increase due to the perfect match probes, plus a random error term. RMA is a combination of normalization techniques and a model that accounts for inter-probe variability, background, and random error. GCRMA improves upon RMA by correcting for hybridization differences caused by probe GC content. This improves performance for probes with low intensity levels [15].

17.3.3 Normalization

Normalization of microarrays involves correcting for different sources of systematic bias. These effects can interfere with and mask the biological variability that one is interested in measuring. Normalization is also aimed at making hybridizations "comparable". The basic assumption behind all normalization techniques is that most of the genes on a microarray will not show any differential expression.

A principle analogous with Occam's razor should be applied when normalizing microarray data. It is, in general, better to under-normalize than to over-normalize the data. If an array is difficult to normalize, it might be best to verify that the experiment was performed properly, and repeat it if necessary. Including a bad array in an analysis can spoil the whole batch if one is not careful.

Often, a linear transformation will be insufficient, and the data will show non-linear effects. These non-linear effects might be because of differences in the quantum efficiency of the fluors, differential dye incorporation, print tip effects, or plate effects. Dye effects are typified by stronger fluorescence in the green (Cy3) channel, especially visible at the low intensity end in an MA plot (intensity vs. log ratio).

Such non-linearity can be corrected by a variety of techniques, including lowess (locally weighted sum of squares) regression of the log ratio on the average intensity, and variants thereof. For an excellent in-depth treatment of normalization for two-color microarrays, we recommend Yang et al. [16].

Spatial bias is another source of systematic variability that can result from uneven hybridization or washing. Such effects are usually visible in the raw image file. Some normalization packages, for example SNOMAD (available online) and YASMA (available in R), correct for spatial biases by fitting a loess surface to the average intensity, or log ratios, over the row and column coordinates of the array [17, 18].

To deal with heteroscedastic variance, a novel generalized log transformation was developed concurrently by the groups of Rocke and Huber. The transformation is calibrated from the data to produce expression measures with constant variance

across the whole range of fluorescence intensities. The algorithm is implemented in the package VSN for the Bioconductor project [19].

To make arrays comparable, some form of scaling between them might sometimes be required. The median absolute deviation (MAD) scaling outlined in Yang et al. is the most commonly applied. This scaling ensures that the distribution of log-ratios on all arrays has the same interquartile range [16].

GeneChip arrays are subject to slightly different consideration when it comes to normalization. Currently, the leading technique is RMA. RMA is a series of different steps and goes beyond gene level summaries. It also involves background estimation and a fit of an additive model on the log scale that accounts for probe specific effects.

17.4 Class Comparison – Differential Expression

Perhaps the simplest type of microarray experiment is to search for sets of genes that are differentially expressed between two or more predefined classes. For example, we might be interested in finding all genes that are significantly over or under-expressed in tumor tissue compared with normal tissue. Identification of differently expressed genes with known function can lead to insight into the biological differences between the classes. Alternatively, this method might enable us to identify differently expressed genes for further study.

Initial strategies for detecting differently expressed genes between two classes are straightforward (Fig. 17.2). Essentially they involve two-sample comparison of the differences between mean log expression of the classes. This is nothing more than a *t*-test or its nonparametric analogs. When more than two classes are involved, *F*-statistics and nonparametric analogs can be applied. Almost all gene expression and statistics software contain routines for performing these analyses.

At least three problems are associated with such simple tests. First, within-class variation of expression and the magnitude of different expression tend to be gene specific. This necessitates the use of methods that weigh inter-class variation against intra-class variation. Second, if we decide on a *P*-value of α and test n genes in this manner, we expect, under the null hypothesis that no genes are differently expressed, $\alpha \times n$ false positives. For $\alpha = 0.01$ and $n = 10{,}000$, this gives 100 false positives. Because it is common for these sets of distinguished genes to be further interrogated by use of (costly) low-throughput techniques, for example Southern blotting and RT-PCR, it is important to minimize the number of false positives. Third, such univariate gene-by-gene tests of different expression ignore important underlying dependencies between gene expression levels. Because all genes are measured by the same device (and are therefore subject to the same sources of error) and the genes are related by complex underlying biological networks (and therefore expression levels are not independent), it makes sense to apply joint estimation techniques.

Standard statistical approaches to addressing these concerns include Bonferroni or Fisher's LSD adjustments to *P*-value calculations. More advanced sampling strategies have been proposed [20] although such strategies require a relatively large number of samples. The ArrayAnalyzer software (in S-PLUS) provides an alternative strategy called *local pooled error* (this algorithm is also available in R). The intuition is to pool variance estimates for genes with similar

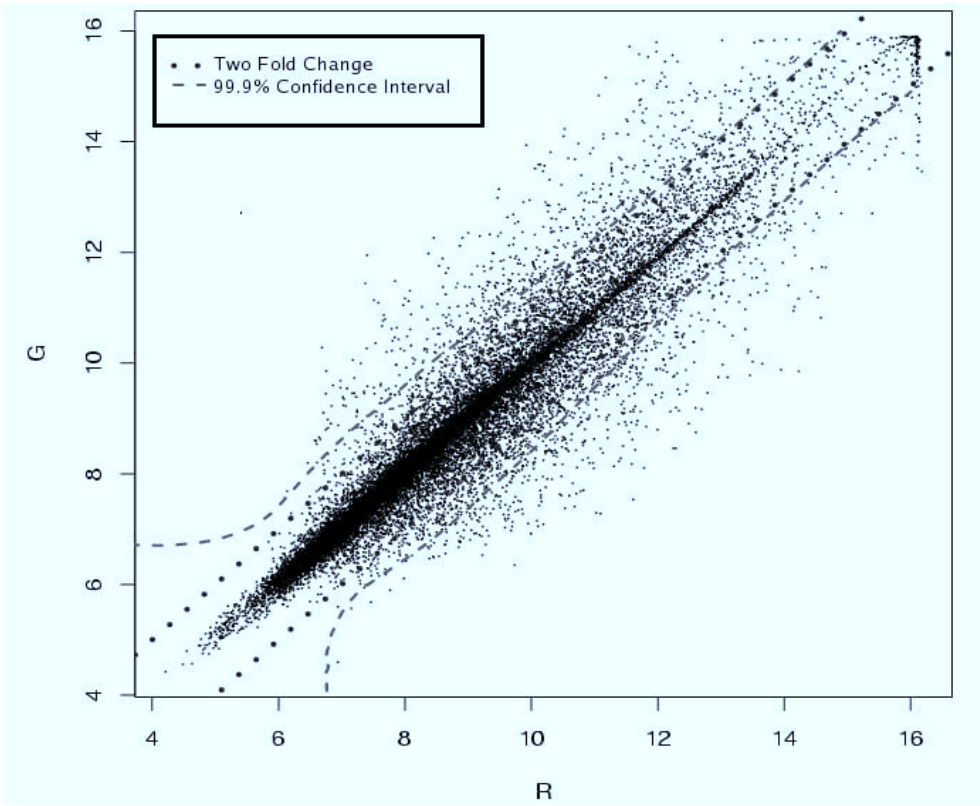


Fig. 17.2 Scatterplot of average log red vs. green channel for replicate two-color microarrays. The dotted line indicates a twofold change threshold for different expression. The dashed line indicates the 99.9% confidence interval for different expression. Notable is the increased variance at low intensity, which is not accounted for by the fold change threshold.

expression intensities and to construct a smooth variance function in terms of average expression intensity.

Another popular and useful tool is *significance analysis of microarrays (SAM)* [21]. SAM is a tool for correlating gene expression with an outcome value (class, treatment, survival time). One of the main uses of the software is to detect differently expression genes, and recent implementations [22] contain more modern treatments of controlling the proportion of false positives. Essentially, they adjust their test statistic for the within-gene variance to take into account the dependencies between different expression and absolute intensity.

Intuitively, Bayesian approaches relax the assumption that the variances in expression levels are equal and model them explicitly. They enable specification of prior distributions for the model parameters. For example, the log expression for a given gene and class are normally distributed with an unknown mean and variance drawn from a specific prior distribution. These techniques are particularly useful when the number of samples per class is small. The Cyber-T package (in R) from Baldi and Long

[23] provides an implementation of this strategy that enables users to provide a broad prior distribution for the mean of expression for genes. Within-class variability is estimated as a weighted combination of the within-class variability of the distinguished genes and the estimated within-class variability of genes with similar expression levels.

Using an empirical Bayes approach, Efron [24] present a framework that estimates the probability that a gene is differently expressed whilst avoiding the need to specify distributions on the parameters. Essentially, means and variances are determined empirically as a mixture of two distributions; the first models differently expressed genes and the second models non-differently expressed genes. The EBarrays (in R) empirical Bayes approach [25] shares many similarities to the Efron [24] system. Smyth et al. developed a hierarchy-based model called *linear models for microarray analysis (Limma)* (in Bioconductor) for an arbitrary number of samples [26, 27].

17.5 Class Prediction

Class prediction methods in microarray analysis are aimed at discovering subsets of genes that can best distinguish between two or more classes of sample. These techniques have been widely applied to discriminate between different classes of leukemia or between tumor and normal breast tissue, and for prognostic prediction. A wide variety of statistical techniques is available for class prediction, including LDA (linear discriminant analysis), weighted voting, nearest-neighbor classifiers, support vector machines, neural nets, and Bayesian methods. At the center of all of these methods is the issue of feature selection. The goal is to select a subset of features (genes) that best distinguish between known classes and predict well new, unseen samples.

Numerous prediction methods are available. Class prediction methods generally require some form of pre-selection of genes to limit the noise present in microarray data. The usual approach is to select genes with different expression between the classes being compared. Linear and quadratic discriminant analysis (LDA and QDA) are well-known classification methods [28]. LDA and QDA can take into account correlation between the genes, whereas diagonal linear or quadratic discriminant analysis assume independence.

A well-known variant of diagonal linear discriminant analysis is Golub's weighted voting method, a restricted form of a naïve Bayes classifier [29]. This method calculates the correlation of each gene vector with an "ideal" expression profile between two classes of sample. The correlation for a gene is defined as the difference between the mean expression of each class, divided by the sum of the standard deviations for each class. Permutation testing can identify genes with statistically significant correlation. The top n significantly correlated (and anti-correlated) genes are selected for use in the predictor. Classification is performed on unseen samples by having each gene cast a vote according to a discriminant function. The sum of the votes for a sample will be either positive or negative, indicating that the sample belongs to one class or the other.

As with support vector machines (SVM), a common technique is to perform a dimensionality reduction. This reduces the vector of expression values for a sample to a single scalar through a weighted combination of the elements. The differences between methods lie in how those weights are calculated. In each case, the weights are

selected to maximize or minimize some target function. SVM use kernel functions that reduce the sample vector into a simpler kernel. The algorithm then finds a hyperplane that separates the classes and maximizes the distance between the instances and the hyperplane.

Nearest neighbor classification requires one to select a distance metric to measure the similarity of expression profiles. The choice of distance metric can have a dramatic impact on the results of nearest neighbor. Correlation or Euclidean distances are most commonly employed. The *k*-nearest neighbors method calculates the distance from the expression profile to be classified to every sample of known class. The *k* nearest profiles are selected and the unknown sample is classified to the majority vote.

PAM (prediction analysis for microarrays) is a variant of the nearest centroid method. Nearest centroids classifies unknown samples by their distance from the centroids of samples in known classes. PAM modifies this methodology by calculating "shrunken" centroids that give higher weight to genes with low variance within a class. For further details, we recommend Tibshirani et al. [30].

The question of prediction error arises in any classification situation. In microarray analysis there are often insufficient samples to furnish an explicit training and test dataset. To assess the accuracy of a classifier that has been trained on the entire dataset, the classic methods of cross validation or bootstrapping can be applied. Both involve separation of the data into training and testing sets. A new classifier is trained on the training data and tested on the testing data. Because the classifier has not seen the testing data, this is an unbiased test of its ability to classify unknown data. Bootstrap estimation of prediction error involves sampling data with replacement from the original data set and performing feature selection and training on that sampled data. This classifier is used to predict the class of the data not present in the training set for that iteration. This procedure is repeated for many iterations and an overall prediction error can be estimated.

Most of these algorithms can be found in add-on packages for R and in most software packages.

17.6 Class Discovery

17.6.1 Clustering Algorithms

It is often informative to try to discover underlying, unknown classes in microarray data. For example one may want to search for classes of cancer with a particularly invasive phenotype. Several methods are used for this purpose; the most popular techniques are possibly *k*-means clustering, hierarchical clustering, self-organizing maps (SOM), and principal component analysis (PCA). A good review of these techniques can be found elsewhere [31].

The goal of clustering is to organize similar sample data together. Microarray data is typically clustered in any of two dimensions. Either the gene or the array dimension can be clustered according to some measure of similarity. Clustering can be performed independently on both the arrays and the genes, enabling one to see both genes with similar expression profiles across arrays, and arrays that have underlying similarities across genes (Fig. 17.3).

Clustering on both genes and arrays simultaneously can sometimes show cases where a subset of genes have similar behavior in a subset of the arrays. The biclustering algorithms introduced by Cheng and

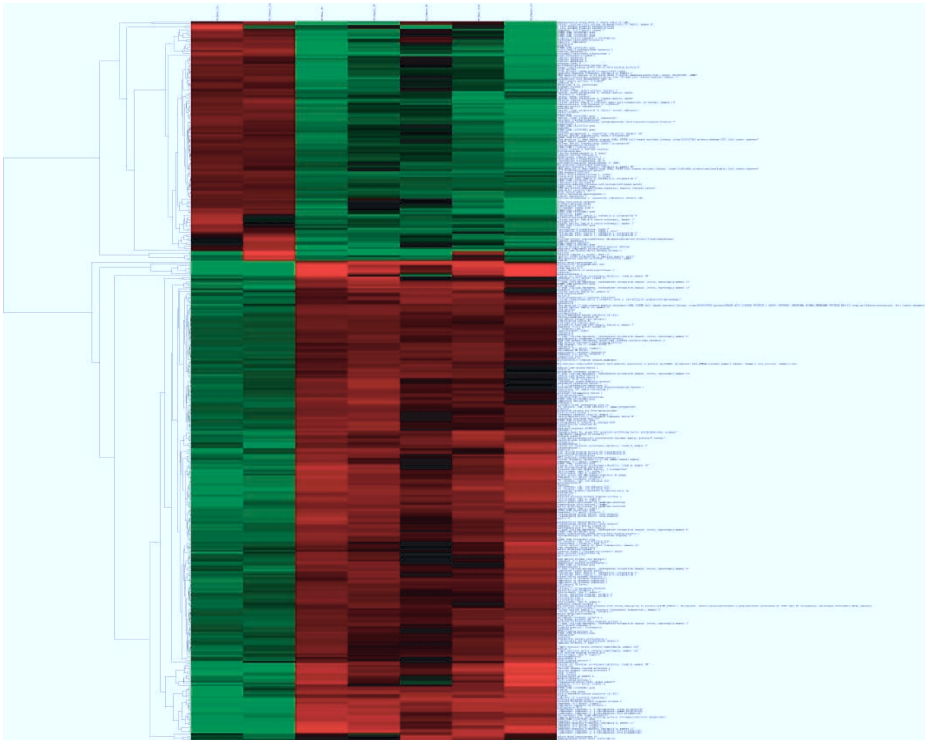


Fig. 17.3 Heatmap showing hierarchical clustering of seven microarray samples and several hundred differently expressed genes. Green indicates high expression, and red indicates low expression.

Church are systematic approaches to this problem [32]. Although clustering is not as sophisticated as some of the other model-based methods for class discovery, it has the advantage of being both easy to use and, usually, of giving a visual result that is easy to understand.

The framework described by Barash and Friedman [33] is of particular interest. This is an extension of the naïve Bayes classifier in which the goal is to discover dependencies between genes and class membership using the structural EM algorithm. Genes are allowed to depend upon the class variable with different distributions for each instance of the class variable.

17.6.2 Validation of Clusters

By clustering it is possible to verify the accuracy of the discovered classes computationally before proceeding to the wet laboratory. Once again, cross validation and bootstrapping can be applied for this purpose. Another useful technique is called parametric bootstrap clustering [34]. The intuition is to assume that all values are sampled on the basis of a specific distribution which can be estimated (Kerr et al. use ANOVA). Bootstrapping re-samples additional values from this distribution and re-clusters the data. If the bootstrap clusters are very close

to the original cluster, this might indicate the cluster is stable.

17.7 Searching for Meaning

It commonly occurs in gene-expression experiments that our analysis produces a set of "distinguished genes". This set might, for example, represent a list of differently expressed genes between two classes, or a list of genes found to be co-expressed over many experimental conditions. Irrespective of the origins of such a set, we are often interested in determining possible relationships between its members. The hope, of course, is that there is some underlying pattern or classification that leads to biological insight into *why* the genes in this particular set are in fact differently expressed or show co-expression. There are now several tools for exploring such questions.

The first such approach involves "mapping" the distinguished set of genes on to the gene ontology (GO). The mandate of the GO project [35] is to enumerate all biological objects and the relationships among these objects. These categories are recursively refined into more specific components, processes, and functions. Each gene is mapped on to one or more categories depending on the specific level of knowledge about the particular gene. Software such as *OntoMiner* [36] and *GoMiner* [37] enable users to map sets of distinguished genes onto this hierarchy in a graphical manner. At the visual level, the user can determine if some category has a surprising number of genes mapped to it. These applications also provide routines for estimating the significance of the over-representation (or, under-representation) of categories.

Although GO continues to expand its ontology into new domains and organisms, the main disadvantage of such approaches is its incompleteness in respect of some biological domains and for some organisms. It is common that the list of genes map to very broad, generic categories that lead to little insight. When genes from some organisms are not contained in GO or are categorized in very broad terms, the user must rely on sequence similarity (for example, BLAST) of these genes to better classify them.

In a manner similar to the use of GO, a second approach maps distinguished sets of genes to metabolic pathways [38]. For organisms like yeast, for which there are well-developed databases of metabolic networks, and signaing pathway such tools can work quite well. Occasionally one can readily see that specific pathways tend to contain many member enzymes that are differently expressed over one or more mRNA expression experiments. This might lead to more specific hypotheses regarding the significance of the distinguished set of genes.

Several tools for exploring mRNA expression data with known protein–DNA and protein–protein interaction databases have now appeared in the literature. Cytoscape [39] is currently one of the better known tools. The underlying assumption is that protein expression levels are well correlated with mRNA expression levels; the intuition is to map expression data on to the protein interaction network. This approach can yield important insights into the protein complexes perturbed in a given experiment and establish functional roles for the distinguished genes.

When co-regulation of a set of distinguished genes is suspected, because of patterns of co-expression of the genes, several tools can be used to identify common motifs in the regulatory regions of these genes. These common motifs presumably correspond to binding sites of transcription fac-

tors regulating the genes. Finding such common motifs can lend credence to the hypothesis of a biological role for the cluster, as opposed to members in a common cluster by chance. Several tools can be used to explore such questions, including INCLUSive [40]. The framework of the procedure introduced by Segal et al. [41] combines the search for transcription factor binding sites with the construction of clusters via gene expression data in a Bayesian setting. In a series of papers these authors progressively develop a framework that integrates many (heterogeneous) types of data including mRNA expression data, binding site motifs, cell-cycle information and so-called "ChIP on Chip" data via a Bayesian network. Their results for yeast suggest that this approach is powerful and the coherent statistical framework enables further types of data to be incorporated easily. The downside of such techniques, however, is that they are somewhat unproven for higher-order Eukaryotic organisms and there is a general lack of tools available for non-experts. Although there are general-purpose tools for Bayesian networks, these tools are still somewhat inaccessible to researchers without explicit knowledge in this area. Efforts such as BIAS (http://www.mcb.mcgill.ca/~bias) [42] promise to offer open source tools soon.

References

1 Schena M (2002) Microarray analysis. John Wiley and Sons, New Jersey
2 Wasserman LA (2004) All of Statistics. Springer, New York
3 Hastie T, Tibshirani R, Friedman J (2003) The Elements of Statistical Learning. Springer, New York
4 Kohane IS, Kho A, Butte AJ (2002) Microarrays for Integrative Genomics (Computational Molecular Biology). MIT Press, London
5 Yang YH, Speed T (2003) Design and analysis of comparative microarray experiments. In: Speed TP (Ed) Statistical Analysis of Gene Expression Microarray Data. Chapman and Hall/RCR, pp. 35–92
6 Pan W, Lin J, Le CT (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol 3(5):1–10
7 Yang MCK, Yang JJ, McIndoe RA, She JX (2003) Microarray experimental design: power and sample size considerations. Physiol Genomics 16(1):24–28
8 Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A , Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29(4):365–371
9 Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ, Brazma A (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3(9):1-0046.9
10 Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol 15:3(8)
11 Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comput Graph Stat 5(3):299–314
12 Kamberova G, Shah S (2002) DNA Array Image Analysis: Nuts and Bolts. Independent Publishers
13 Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 4(2):249–264
14 Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA. 98(1):31–36
15 Wu Z, Irizarry RA (2004) Stochastic models inspired by hybridization theory for short oligonucleotide arrays, In: Gusfield D, Bourne P, Istrail S, Pevzner P, Waterman M (Eds) 8th Annu Int Conf Computational Molecular Biology (RECOMB '04), pp 98–106
16 Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide

systematic variation. Nucleic Acids Res. 30(4):e15
17 Colantuoni C, Henry G, Zeger S, Pevsner J (2002) SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis. Bioinformatics 18(11):1540–1541.
18 Wernisch L, Kendall SL, Soneji S, Wietzorrek A, Parish T, Hinds J, Butcher PD, Stoker NG. (2003) Analysis of whole-genome microarray replicates using mixed models. Bioinformatics 19(1):53–61
19 Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 Suppl 1:S96–S104
20 Dudoit S, Speed TP (2000) A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. Biostatistics 1(1):1–26
21 Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA. 24;98(9):5116–5121
22 Storey JC, Tibshirani, R (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: Parmigiani G, Garrett ES, Irizarry RA, Zeger SL (Eds) The Analysis of Gene Expression Data, Springer, New York, NY, pp 272–290
23 Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. Bioinformatics 17:509–519
24 Efron B (2003) Robbins, empirical Bayes and microarrays. Ann Stat 31:366–378
25 Newton MA, Kendziorski CM, Parametric empirical Bayes methods for microarrays. In: Parmigiani G, Garrett ES, Irizarry RA, Zeger SL (Eds) The Analysis of Gene Expression Data, Springer, New York, NY, pp 254–271
26 Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3(1):Article 3
27 Lönnstedt I, Speed TP (2002) Replicated microarray data. Stat Sinica 12:31–46
28 Hastie T, Tibshirani, R, Friedman J (2001) Linear methods for classification. In: The Elements of Statistical Learning. Springer, New York, NY, pp 79–114
29 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537
30 R, Hastie T, Narasimhan B, Chu G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA. 99(10):6567–6572
31 Quackenbush J (2001) Computational analysis of microarray data. Nat Rev Genet 2(6):418–427
32 Cheng Y, Church GM (2000) Biclustering of expression data. Proc Int Conf Intell Syst Mol **Biol 8:93–103**
33 Barash Y, Friedman N (2002) Context-specific Bayesian clustering of gene expression data. J Comput Biol 9(2):169–191
34 Kerr MK, Churchill GA (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci USA 98(16):8961–8965
35 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2003) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 25(1):25–29
36 Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz S, Tainsky M (2003) Onto-Tools, The toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucleic Acids Res 31(13):3775–3781
37 Zeeberg B, Feng W, Wang G, Wang M, Fojo A, Sunshine M, Narasimhan S, Kane DW, Reinhold W, Lababidi S, Bussey K, Riss J, Barrett JC, Weinstein J (2003) GoMiner: A resource for biological interpretation of genomic and proteomic data. Genome Biol 4(4):R28
38 Bouton CM, Pevsner J (2002) DRAGON View: information visualization for annotated microarray data. Bioinformatics 18(2):323–324

39 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504

40 Thijs G, Moreau Y, De Smet F, Mathys J, Lescot M, Rombauts S, Rouze P, De Moor B, Marchal K (2002) INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. Bioinformatics 18(2):331–332

41 Segal E, Barash Y, Simon I, Friedman N, Koller D (2002) From promoter sequence to expression: A probabilistic framework in: Proc. 6th Int. Conf. on Research in Computational Molecular Biology (RECOMB), Washington, DC

42 Finak G, Goding N, Hallett M, Pepin F, Rajabi Z, Srivastava V, Tang Z (2004) BIAS: bioinformatics integrated application software. Bioinformatics (advanced access)

18
Protein Interaction Databases

Gary D. Bader and Christopher W. V. Hogue

18.1
Introduction

Given estimates based on the draft sequence of the human genome [1, 2] of between 30,000 to 80,000 human genes, it is apparent that a minority of these genes encode conventional metabolic enzymes or transcription–translation apparatus. Sequencing of the genomes of metazoans and, more specifically, vertebrates has uncovered large numbers of complex multi-domain proteins, many containing interacting modules, such as SH3 domains, which generally bind proline-rich protein regions. The complexity of the DNA blueprint is augmented exponentially when one considers the possibility that these multi-domain proteins could bind to several other biomolecules either simultaneously or at different points in the cell cycle or in different cell types. A molecular interaction is a specific binding event resulting from atomic-level physicochemical forces [3]. Multiple binding events among many different proteins in a cell form "interaction networks". These networks form conventional signaling cascades, classical metabolic pathways, transcription activation complexes, vesicle mechanisms, and cellular growth and differentiation systems, indeed all of the systems that make cells work [4].

The ultimate manifestation of gene function is through intermolecular interactions. It is impossible to disentangle the mechanistic description of the function of a biomolecule from a description of other molecules with which it interacts. One of the best forms of the annotation of a gene's function, from the perspective of a machine-readable archive, is information linking specific molecular interactions together, an interaction database. Thus, interactions, defining molecular function, and interaction databases are critical components as we move toward complete and dynamic functional description of the cell at a molecular level of detail. Interaction databases are essential to the future of bioinformatics as a new science. In this review, what can be achieved through integration of current interaction information into a common framework is broadly considered, and a number of databases that contain interaction information are examined.

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

18.2 Scientific Foundations of Biomolecular Interaction Information

Interaction information is based on the experimental observation of a specific interaction between two or more molecules. For the purposes of this discussion, natural, biological molecules spanning the entire range of biochemistry are spoken of, including proteins, nucleic acids, carbohydrates, and small molecules, both organic and inorganic and even light. Interaction information could also be considered for genes, as in a synthetic lethal genetic interaction, although this interaction is less direct. Interaction information is an inference that two or more molecules have a preferred specific affinity for each other, within a living organism, and that inference is based on experimental evidence using conventional experimental molecular and cell biology techniques.

The number of experimental types that can provide this evidence is large. Primary interaction experiments can be broadly described as being based on direct observation of two molecules directly interacting or of a measurable phenomenon directly related to that interaction. This might be *in vivo*, for example a yeast two-hybrid assay, or *in vitro* as in a fluorescence polarization experiment using purified reagents in a cuvette. Experimental genetic evidence provides another type of information. For example a tandem gene knockout in an organism might cause a particular phenotype to appear, for example a growth defect or lethality. This phenotypic readout provides evidence that the two genes are involved in pathways affecting the phenotype, and might imply molecular interaction between the two gene products or that the two genes act in redundant pathways. These data are indirect and possibly dependent on other genes in the background of the experimental system. Nonetheless all this information is important in helping us to understand gene and protein function in dynamic molecular interaction networks. Storage of primary interaction data in a common machine-readable archive, as is currently done for gene sequence and molecular structure information, would give us a tremendous resource for research biology and data retrieval.

Interaction databases should, ideally, contain information in the form of a correlated pair or group of molecules, some link to the experimental evidence that led to the interaction, and machine-readable information about what experimental interaction data are known. For example, did the interacting molecules undergo a chemical change during the interaction? Was the binding reversible? What are the kinetic and thermodynamic data, if they were measured in the experiment? Were the forms of the molecules in the experiment wild type or mutated variants? What are the binding sites on the molecules?

18.3 The Graph Abstraction for Interaction Databases

Consider the collection of molecules in a cell as a graph. Each molecule is a vertex or node, and each interaction is an edge. Classical bioinformatics databases hold protein sequence, DNA sequence, and small molecule chemistry databases, i.e. collectively hold molecules which are the vertices of this graph. In contrast, the ideal interaction database will hold the edge information – which two molecules come together, under which cellular conditions, location and stage, how they interact, and what happens to them in the course of the interaction. This concept is referred to as the graph-theory abstraction for interaction databases. It

is a powerful data abstraction as it simplifies the underlying concepts and allows one to apply algorithms that are well understood from the field of computer science to the larger problems of data mining and visualization. Having a clear picture of a general graph abstraction for interaction databases is the key to the integration of data into a universal framework.

Nodes in a graph do not have to represent single parts of a cell; a single node can represent multiple related parts. For example, a protein and its orthologs could be represented by one node in a graph of a metabolic pathway. The graph would thus correspond to a generalized pathway across more than one organism. Edges in graphs can have direction from cell-signal information flow (e.g. from cell surface to nucleus) or from chemical action direction (e.g. kinase phosphorylates a protein substrate). Nodes and edges can be assigned a weight that could be mapped from information in a database. For instance, a node could be assigned a higher weight if it is a larger protein and this could be translated as a larger node in a graph visualization system. An edge can be assigned a weight based on the confidence in an interaction. This probability value could be derived from a function of the type and number of experiments that were performed to conclude that two molecules interact.

The Biomolecular Interaction Network Database (BIND) [5–7] seeks to create a database of interaction information around a generalized graph theory abstraction of interaction data.

18.4 Why Contemplate Integration of Interaction Data?

In building the BIND data model, a prototyping approach was pursued, which is very different from the way many biological databases are created. A comprehensive data model was designed that enabled interaction information to be represented in a machine-readable format, spanning all type of molecular interaction, including protein, RNA, DNA and small molecules and the biochemical reactions, complexes and pathways they are involved in. The BIND data specification was created in accordance with the NCBI ASN.1 architectural model [8] and the NCBI software development toolkit for implementing early versions of the BIND database and its tools. A substantial amount of time was spent focusing on designing the data model for BIND, contemplating the way molecular interaction and molecular mechanism information would be stored, from inferences as broad as a genetic experiment to as precise as the atomic level of detail found in a crystal structure of an interacting complex. The hypothesis was that there is a plausible universal computer readable description of molecular interactions and mechanisms that can suffice to drive whole cell visualization, simulation and data-retrieval services. The design phase involved asking ourselves and others questions such as: What data should be represented? What abstractions should be used? How can interactions be described together with chemical alterations to the interacting molecules? The outcome of this hypothesis testing exercise is described in detail in the BIND specification [5].

18.5 A Requirement for More Detailed Abstractions

Molecular interaction data must be abstractly represented so that computations can be performed and data maintained in a

machine-readable archive more easily. This is a simple idea with an analogy in biological sequence information. Biopolymer molecules, DNA and proteins, are abstracted for the computer as strings of letters. This information tells us nothing about the conformation or structure of the molecule, just its composition and biopolymer sequence. The IUPAC single letter code for DNA and for amino acids are abstractions that contain sufficient information to reconstruct chemical bonding information, provided that a standard form of the biopolymer is being represented, and not a phosphorylated, methylated or otherwise modified form.

One cannot imagine a database of cellular biomolecular assembly instructions without first having an enumeration of the contents of the cell, the biomolecular parts list. Sequence databases only partially fulfill this parts list requirement, because precise information about post-translational modification of biopolymers is not encoded. Also, small molecules such as metabolites are not included in sequence databases. To encode exact information about biomolecules, one must have the capacity to describe the biopolymer both as a sequence and as an atom-and-bonds representation, the chemical graph.

A chemical graph description of a biomolecule is sufficient to recreate the atoms, bonds and chirality of the molecule, but without specifying the exact location of the atoms in 3D space. In other words, a chemical graph is an atomic structure without coordinate information. A chemical graph data abstraction exists within the NCBI MMDB data specification and database of molecular structure information [9]. This specification is the only example of a chemical graph based structure abstraction, and a complete chemical structure might be encoded in MMDB without knowing a single *X,Y,Z* atom coordinate. Neither the PDB (protein database) nor the newer mmCIF molecular structure file format has a chemical graph data structure that can describe the complete chemistry of a molecule without atomic coordinate information [10].

Sequence alphabet abstractions have been invaluable in bioinformatics, having enabled all computer-based sequence analysis. This would have been very difficult to compute had an exact database of atoms and bonds making up each biopolymer sequence been chosen as the abstraction. While this information might bog down sequence comparison, it is required for a more precise record of the chemical state of a biopolymer after post-translational changes. These chemical states, once accessible through a precise database query, are important to have recorded, because they form the control points for uncounted pathways and mechanisms for cellular regulation.

Abstractions are rarely universally applicable for all kinds of computation. As computing power increases, abstractions can be expanded in detail to fulfill the requirements of more kinds of computation. So far there has been resistance against expanding the abstractions of sequence information to more complete descriptions like a chemical graph, but it is clear these will be required to describe large and important parts of molecular biology such as phosphorylation, carbohydrate or lipid modification, and other post-translational changes upon which many molecular mechanisms depend.

Interaction databases can be contemplated now because it has been demonstrated that computer infrastructure can keep up with genomic information. The representational models selected must, however, be carefully chosen so that they not preclude computation that might be required in future research. It might be time to find an abstraction that can accommodate the most

complete description for molecular information one can imagine. With adequate standard data representations for molecules that are unambiguous for the purposes of general computation, specifying sequences, structures, and small molecule chemistry, it should be possible to move ahead with annotation of molecular function in a very complete fashion. Without this, machine-readable descriptions of knowledge will be ambiguous and will be limited in the precision which biological simulation, visualization, and data-mining tools will require.

18.6 An Interaction Database as a Framework for a Cellular CAD System

To achieve the goal of a computing and software system that can achieve whole-cell simulation, something like a CAD (computer-aided design) system must be built. CAD systems are used in engineering, for example, in the design of electronic circuitry. In biology, such a system could be used for representation and possible design of cellular circuitry. Unlike engineering, the biological CAD system could be used backwards as a reverse-engineering tool to enable understanding of the complexity of cellular life. This system would have a detailed representation of biochemistry sufficient to enable output of a data description of a snapshot of a living cell to a simulation, data-mining, or visualization system. In engineering, CAD systems are database-driven software, and the utility of a particular CAD system is proportional to the content of its database of parts, the symbols used to describe electronic components. Likewise, a biological CAD system will require a complete list of parts as an integrated software and database system. The fragmentation in the Bioinformatics parts list community must obviously be resolved to achieve such a list [11]. Federated databases with highly latent network interconnections and imprecise data models will not suffice for large cellular information systems. Interaction and parts information is best stored as an integrated system to meet the data demands of whole-cell simulation, visualization, and data mining. Overall, such a system requires a formal data model for molecular interactions that enables good abstraction of the data with precise computability without sacrificing the complexity of the information. The emergence of a standard will enable diverse groups to collaborate and work towards their common goals more efficiently.

18.7 BIND – The Biomolecular Interaction Network Database

The Biomolecular Interaction Network Database (BIND, http://bind.ca) has been designed as a system for storage of biomolecular interactions with the positive attributes of an interaction database discussed above. BIND is a web-based database system that is based on a data model written in ASN.1 (abstract syntax notation, http://asn1.elibel.tm.fr/). ASN.1 is a hierarchical data description language used by the NCBI to describe all of the data in PubMed, GenBank, MMDB, and other NCBI resources [12]. ASN.1 is also used extensively in air-traffic-control systems, international telecommunications, and Internet security schemes. The advantages of ASN.1 compared with other computer readable data description languages such as XML include being strongly typed and having an efficient cross-platform binary encoding scheme that saves space and CPU resources when

transmitting data. Disadvantages are that commercial ASN.1 tools are very expensive and that the ASN.1 standard process is closed. The NCBI, however, provides public domain cross-platform software development toolkits written in the C and C++ languages to deal with the NCBI data model and with ASN.1. Each toolkit can read an ASN.1 defined data model and generate C or C++ code that enables automatic reading (parsing), writing, and management of ASN.1 objects. Also supported is the ability to automatically translate ASN.1-defined objects to and from XML, and the automatic generation of an XML DTD for an ASN.1 data specification. These toolkits are currently available at ftp://ncbi.nlm.nih.gov/toolbox/. This powerful data-description language and toolkit combination enables us to circumvent the large and time-consuming problem in bioinformatics of parsing primary databases to integrate data for effective research use. With the toolkit, parsing is automatic, through use of machine-generated parsing code. The use of ASN.1 also allows the BIND specification to use mature NCBI data types for biological sequence, structure and publications.

Recently, the XML (extensible markup language, http://www.w3.org/XML/) language has gained popularity for data description. XML matches ASN.1 in its ease of use, although it does not provide strong types. For instance, ASN.1 recognizes integers and can validate them whereas XML treats numeric data as text. The advantages of XML are its open nature and familiarity, because it is similar to HTML. Many tools currently use XML, although free code-generation and rapid application development tools are only beginning to mature. XML also wastes space because it does not have a binary encoding scheme and because it is tag based (Fig. 18.1). An XML message will be many times

A)

```
Date ::= CHOICE {
     str VisibleString,
     std Date-std
}

Date-std ::= SEQUENCE {
     year INTEGER,
     month INTEGER OPTIONAL
}
```

B)

```
date
     std {
          year 1974 ,
          month 7 ,
          day 7
     }
```

C)

```
<Date>
     <Date_std>
          <Date-std>
               <Date-std_year>1974</Date-std_year>
               <Date-std_month>7</Date-std_month>
               <Date-std_day>7</Date-std_day>
          </Date-std>
     </Date_std>
</Date>
```

Fig. 18.1 Examples of ASN.1 and Equivalent XML. (A) An example of how a date data type is specified in ASN.1. (B) An example of how an instance of specific date data is represented in the print form of ASN.1. The BER binary encoded form of this ASN.1 would only take up less than 20 bytes. (C) An example of how the same date data as in (B) is represented in XML. XML does not have a binary encoded form. Note the excess of information required to specify a date.

larger than a binary encoded ASN.1 message. Recently, the XML Schema standard (http://www.w3.org/XML/Schema) has become popular and partially solves some of the problems of XML, such as lack of strong types. A wide range of commercial and open-source developer community support is available, thus XML Schema can be regarded as a replacement of ASN.1.

The BIND data specification describes biomolecular interaction, molecular complex, and molecular pathway data. Both genetic and physical interactions can be stored. Chemical reactions, photochemical activation, and conformational changes can be described down to the atomic level of detail. Everything from small molecule biochemistry to signal transduction is abstracted in such a way that graph theory methods can be used for data mining. The database can be used to study networks of interactions, to map pathways across taxonomic branches, and to generate models for kinetic simulations. The database can be visually navigated using a Java applet and queried using a text search or the BLAST against BIND service. BIND is an open effort; all records are in the public domain and source code for the project is made freely available under the GNU Public License. Users are encouraged to submit their favorite interactions. BIND has been used to manage and automatically discover new knowledge residing in large yeast protein–protein and genetic interaction networks in *Saccharomyces cerevisiae* determined using mass spectrometry, phage-display, yeast two-hybrid, and roboticized synthetic lethal screens. Recently, a large curation effort run by the non-profit Blueprint Initiative has been developed to keep BIND up-to-date with the growing amount of published molecular and genetic interaction data.

18.8 Other Molecular-interaction Databases

Most molecular interaction data resides in the scientific literature, in unstructured text, tables, and figures in thousands of papers in molecular and cellular biology. It is currently impossible to retrieve information from this archive using computational tools such as natural language query methods with the accuracy required by scientists. Many databases currently available, mainly over the Internet, contain interaction information, although most of these databases are not focused on storing biomolecular interactions. Most of these databases are small and have very select niches of interaction information, for example, the restriction enzyme database REBase [13] maintained by New England Biolabs, while not an interaction database *per se* does contain interactions between restriction enzymes and the specific patterns of DNA that they recognize and cleave. These protein–DNA interactions satisfy the node and edge criterion of the graph abstraction of interaction data and are thus a very valuable source of information.

18.9 Database Standards

The pathway informatics community has recently started to develop common data-exchange formats. The PSI (Proteomics Standards Initiative) has developed the first version of an XML-based format for exchanging protein–protein interactions, called PSI-MI (PSI molecular interactions) [14]. The data model of the format is simple, containing an "interaction" record comprising a set of proteins that interact, an experimental conditions-controlled vocabulary, and information about publication references and

protein features, for example binding sites and post-translational modification sites. The BIND, DIP, HPRD, MINT, and IntAct databases make their data available for download as PSI-MI files. Also network visualization tools, for example Cytoscape [15], can read and write PSI-MI-formatted XML.

An effort, called BioPAX, involving the BIND, EcoCyc, and WIT databases, among others, to develop an XML-based format for exchanging full pathways is nascent and might be available by the time this book is published (http://www.biopax.org).

18.10 Answering Scientific Questions Using Interaction Databases

The main purpose of building a computational infrastructure, such as an integrated molecular-interaction database, is to use it to answer scientific questions. For example, many of the protein–protein interactions of macromolecular signaling complexes are mediated by domains that function as recognition modules to bind specific peptide sequences found in their partner proteins [16]. SH3, WW, and EVH1 domains bind to proline-rich peptides [17–19], EH domains bind to peptides containing the NPF motif [20, 21], and SH2 and FHA domains bind to peptides phosphorylated at Tyr and Thr, respectively [22, 23]. For particular modules within the same family, specificity is determined by critical residues in the binding partner flanking the core peptide motif [24, 25]. A major challenge is to construct protein–protein interaction networks in which every module within the predicted proteome of a sequenced organism is linked to its cognate partner.

In budding yeast, the network of interactions mediated by SH3 domains has been mapped [26]. Yeast two-hybrid and phage display screens were used to gather protein interaction information about the 28 SH3 domains in 24 different proteins in the yeast proteome. Two-hybrid methods are known to have a high false-positive rate for interaction determination, and phage display results in a binding motif that can only be indirectly used to predict protein–protein interactions by scanning the yeast proteome for the binding motif for each SH3 domain. Conveniently, these two independently derived protein–protein interaction networks have orthogonal error profiles, thus an intersection of the networks should contain higher-quality interactions than are found by each method separately (Fig. 18.2). This was found to be true by comparing the intersection network to a gold standard and was shown to be statistically significant compared with a random model of the overlap procedure [26]. Throughout this work, an interaction database was used to organize the experimental information and was used as a base to build tools that performed the intersection, the random model, and the network visualization (Fig. 18.2) enabling this particular scientific question to be answered more efficiently.

18.11 Examples of Interaction Databases

Examination of the literature and the Internet results in a large and varied list of databases that contain interaction information covering proteins, DNA, RNA, and small molecules. In fact, over 100 Internet-accessible interaction-related databases have been catalogued in the pathway resource list (http://www.cbio.mskcc.org/prl). The number of projects indicates the importance of this data. The variety of data representation and file formats, data architecture, and license agreements is, however, a daunting

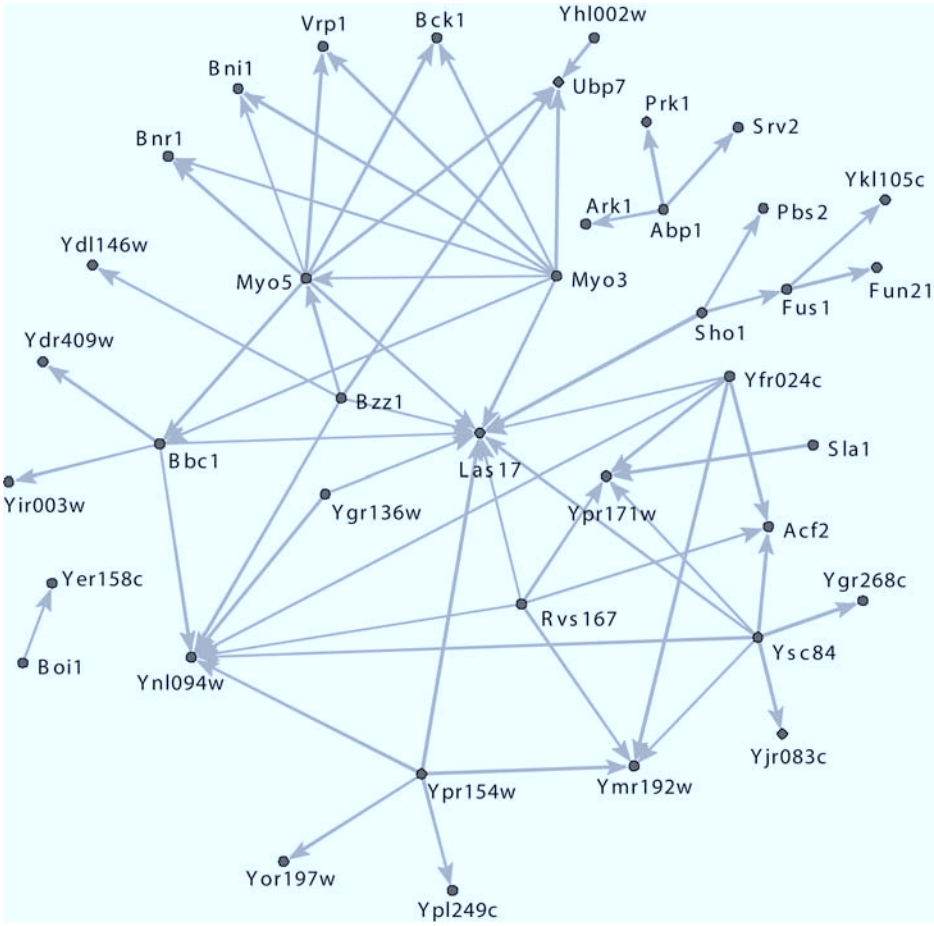


Fig. 18.2 The overlap of the protein–protein interaction networks derived from phage display and two-hybrid analysis is shown. All these interactions are predicted to be mediated directly by SH3 domains. The arrows originate from an SH3 domain protein and point to the target protein. The original phage display predicted a network containing 394 interactions and 206 proteins whereas the original yeast two-hybrid network contains 233 interactions and 145 proteins.

challenge to integration of this information into a common scheme. One can classify databases according to whether they are linked back to primary experimental data in the literature or are secondary sources of information based on review articles or the knowledge of expert curators. Databases based on primary information are few in this list, yet are among the most valuable.

The following database review highlights a number of representative resources and whether their data is present in a machine-readable form. Many databases are packed with information, but the information is entered in such a way that it cannot be unambiguously matched with those in other databases. For example, some databases are missing key data descriptors, for example

sequence accession numbers, CAS chemical compound numbers, PubMed identifiers for publication references, or unambiguous taxonomy information when data from multiple organisms are present. This limits the usefulness of the information, because it is difficult to tie it to other knowledge, which is required on a large scale if it is to be mined and more broadly understood. It is critical that these projects move toward sound database principles when describing data such that it can be manipulated unambiguously and precisely by computer. Where possible, the primary reference and license terms of the database to academic and industrial users of the data is listed to aid future data integrators when choosing databases to import into a data warehouse.

aMAZE
URL: http://www.amaze.ulb.ac.be/
Refs. [27, 28]

The aim of the aMAZE is to describe metabolism, gene regulation, molecular transport, and signal transduction. aMAZE is based on a formal object-oriented data model that will be integrated with CORBA. The database is chemical reaction-based, was started by describing metabolic pathways only, and was seeded from data from BRENDA [29]. aMAZE uses a graph abstraction for the interaction data and can describe chemical reactions and pathways. This has enabled pathway-finding and visualization tools to be implemented.

Aminoacyl-tRNA Synthetase Database
URL: http://rose.man.poznan.pl/aars/index.html
Ref. [30]

Contains aminoacyl-tRNA synthetase (AARS) sequences for many organisms. This database is simply a sequence collection, but collated pairs of AARS + tRNA can be used to create RNA–protein interaction records. The database is freely available over the web.

ASEdb (Alanine Scanning Energetics Database)
URL: http://www.asedb.org
Ref. [31]

ASEdb is a database of protein side-chain interaction energetics determined by alanine-scanning mutagenesis; it is curated manually by a single group. The database is not very large, but does provide valuable information on proteins binding with other molecules, mainly other proteins. This is derived from alanine scanning mutagenesis followed by a measurement of the change in free energy of binding that the mutation caused. The database is web-based and text searchable, but only contains a few specialized database fields.

BBID (Biological Biochemical Image Database)
URL: http://bbid.grc.nia.nih.gov/
Ref. [32]

The BBID is a searchable database of images from publications about cellular pathways and other biological relationships. It focuses on signal-transduction pathways. The molecules in the figures and the publications from which the figures are taken are indexed in a database that enables searching of the figures. Although molecular interaction information is available in the figures, it is not extracted in a machine-readable form, thus BBID remains a human reference only and is not machine-readable.

BindingDB (The Binding Database)
URL: http://www.bindingdb.org/
Refs. [33–35]

BindingDB is a public, web-based database containing kinetic and thermodynamic binding constants for interacting biomole-

cules. The data are only from isothermal titration calorimetry and enzyme-inhibition experimental methods, but may include data from other methods in the future. The database is rigorously designed and implemented using the latest database technology. The search interface is very advanced and even enables searching for small molecules that are similar to an input structure. Although it does contain information about biomolecular interactions, the data specification is focused on binding constant information and experimental method description for two specific methods.

Biocarta
URL: http://www.biocarta.com/

Biocarta is a commercial website which provides manually created clickable pathway maps for signal transduction as a resource for the scientific community, although the purpose of the site is to sell reagents. The presence of a standard set of symbols to represent various different protein components of pathways make the pathway maps clear and easy to understand. Proteins are linked to many different primary databases including PubMed, GenBank, OMIM, Unigene [36], KEGG, SWISS-PROT [37], and Genecard. Companies can sponsor genes and links to commercially available reagents are present. Biocarta invites volunteer users to supply pathways as figures, and Biocarta then creates clickable linked maps and makes them available via the web. The data model is not public and the database has not been published in peer-reviewed literature.

Biocatalysis/Biodegradation Database
URL: http://www.labmed.umn.edu/umbbd/
Ref. [38]

The UM-BBD contains data about microbial biocatalytic reactions and biodegradation pathways primarily for xenobiotic, chemical compounds. Approximately 140 pathways, over 915 reactions, 860 compounds, 580 enzymes, and 330 microorganisms are currently represented. The data model is chemical reaction-based with graph abstraction for pathways. Graph abstraction enables the "Generate a pathway starting from this reaction" function. PDB files for some of the small molecules are available. Graphics (clickable GIFs) are available for the different pathways. The work is funded by several organizations and is free to all users. Data are entered on a volunteer basis and records contain literary references to PubMed.

BRENDA
URL: http://www.brenda.uni-koeln.de/
Refs. [29, 39]

BRENDA is a database of enzymes. It is based on EC number and contains much information about each particular enzyme including reaction and specificity, enzyme structure, post-translational modification, isolation/preparation, stability and cross references to structure databanks. Information about a chemical reaction is extensive, but some of it is in free-text form and thus is not machine-readable. The database is copyright and is free to academics. Commercial users must obtain a license.

BRITE (Biomolecular Reaction pathways for Information Transfer and Expression)
URL: http://www.genome.ad.jp/brite/

BRITE is a database of binary relationships based on the KEGG system. It contains protein–protein interactions, enzyme–enzyme relations from KEGG, sequence similarity, expression similarity, and positional correlations of genes on the genome. The database mentions that it is based on graph theory, but no path-finding tools are present.

BRITE contains some cell cycle-controlling pathways that have now been incorporated into KEGG.

COMPEL (Composite Regulatory Elements)
URL: http://compel.bionet.nsc.ru/
Ref. [40]

Contains protein–DNA and protein–protein interactions for Composite Regulatory Elements (CRE) affecting gene transcription in eukaryotes including the positions on the DNA to which the protein binds. The database is organized in a fielded flat-file format and provides links to TRANSFAC. The data model does not use a graph theory abstraction. In January 1999 COMPEL 3.0 contained 178 composite elements. A professional version is available for purchase.

COPE (Cytokines Online Pathfinder Encyclopedia)
URL: http://www.copewithcytokines.de/

COPE is an encyclopedia of cytokines and related biological terms. COPE provides a free-text textbook-like entry describing each of the many terms and a dictionary for term definitions. Protein and other biomolecular interactions relating to the terms in the encyclopedia are described. The database can be browsed and searched using keywords but contains no formal data model and is thus not natively machine-readable.

CSNDB (Cell Signaling Networks Database)
URL: http://geo.nihs.go.jp/csndb/
Ref. [41]

CSNDB contains cell-signaling-pathway information for *Homo sapiens*. It has a data model that is specific to cell-signaling only and is constructed on ACeDB [42]. It is based both on interactions and reactions, and stores information mainly as unstructured text in fields within a structured record. The data model is sound and some fields contain controlled vocabulary. An extensive graph theory abstraction is present. It is probably one of the first databases to use a simple graph theory abstraction since its first publication in 1998. It can limit the graph to a specific organ and can mask subtrees for this feature. Fields have been added as they are needed and the system is not general. CSNDB contains interesting pharmacological fields for drugs, for example IC50. The database can represent proteins, complexes, and small molecules. It is linked to PubMed and TRANSFAC. TRANSFAC recently imported the CSNDB to seed its TRANSPATH database of regulatory pathways that link with transcription factors. An extensive license agreement limits corporate use. Free to academics. Funded by the Japanese National Institute of Health Sciences.

Curagen Pathcalling
URL: http://curatools.curagen.com/

The commercial Curagen Pathcalling program visualizes information from high-throughput yeast two-hybrid screening of the yeast genome along with other yeast protein–protein interaction from the literature. It contains only protein–protein interactions. Pathcalling uses a graph theory abstraction that enables the use of a Java applet to visually navigate the database. Each protein may be linked to SGD [43], GenBank, or SWISS-PROT. Because it is proprietary, the database does not make any of its information, software, or data model available.

DIP (Database of Interacting Proteins)
URL: http://dip.doe-mbi.ucla.edu
Ref. [44]

The DIP database stores only protein–protein interactions. It is based on a binary interaction scheme for representing interactions and uses a graph abstraction for its

tools. A visual navigation tool is present. DIP does not use a formal grammar for its data specification. The DIP data model enables the description of the interacting proteins, the experimental methods used to determine the interaction, the dissociation constant, the amino acid residue ranges of the interaction site, and references for the interaction. DIP contains over 40,000 protein–protein interactions representing approximately 110 different organisms. Academic users may register to download the database for free if they agree to the click-through license. Commercial users must contact DIP for a license.

DRC (Database of Ribosomal Cross-links)
URL: http://www.mpimg-berlin-dahlem.mpg.de/~ag_ribo/ag_brimacombe/drc
Ref. [45]

This database keeps a collection of all published cross-linking data for the *E. coli* ribosome. This is a database of hand-curated dBASE IV files with a web interface (last updated March 7th, 1998). Possibilities of machine-parsing of the database seem limited, because the field data are non-standardized and meant to be human-readable only.

DPInteract
URL: http://arep.med.harvard.edu/dpinteract/
Ref. [46]

DPInteract is a curated relational database of *E. coli* DNA binding proteins and their target genes. It provides BLASTN searching for DNA and has links to SWISS-PROT, EcoCyc, PubMed, and Prosite [47]. The database is text-based with a limited data specification. Interestingly, position-specific matrices are available to describe the DNA binding motif. Records are organized by protein structure family (e.g. Helix–turn–helix family proteins). Updating of the database continued from 1993–1997 and has now stopped. The database is copyright, but is freely available over the web and contains information about 55 *E. coli* DNA-binding proteins with known binding sites.

EcoCyc (and MetaCyc)
URL: http://biocyc.org/
Ref. [48]

EcoCyc is a database (freely available to academics) that contains metabolic and signaling pathways from *E. coli*. EcoCyc is one of the oldest pathway databases. It is based on an object-oriented data model. Chemical reactions are used to describe the data, which is intuitive in this case, because EcoCyc's main goal is to catalog metabolic pathways from *E. coli*. It is currently being retrofitted to deal with protein–protein interactions in cell-signaling pathways, although data are still described using a chemical reaction scheme. The fields of this database are mostly free-text based. All types of molecule from small molecules to molecular complexes can be represented and small molecule structures are present for common metabolites. EcoCyc uses a graph abstraction model that has enabled pathway traversing and visualization tools to be written. EcoCyc contains interactions of proteins with proteins and small molecules. MetaCyc contains EcoCyc and also pathways from over 150 other organisms. BioCyc was recently created to contain EcoCyc, MetaCyc, and computationally derived pathway databases for recently sequenced genomes, similar to the WIT project.

EMP (Enzymes and Metabolic Pathways Database)
URL: http://www.empproject.com/
Ref. [49]

EMP is an enzyme database that is chemical reaction-based. It stores information as

detailed as chemical reaction and k_m. Over 300 fields are stored as semi-structured text that might enable most of the database to be easily machine-readable. The database is part of the WIT project and can also be accessed from the WIT system. GIF and SVG images of many pathways are available and the project is heavily curated. Recently this project underwent a major website reorganization and is now very user friendly and easily searchable. Some source code is available for the project via a CVS server and the database is freely available over the web.

ENZYME
URL: http://www.expasy.ch/enzyme/
Ref. [50]

This database contains enzyme, substrate, product and cofactor information for over 4200 enzymes. It has been a crucial resource for metabolic databases including EcoCyc. It is chemical reaction-based. This database can be translated to an interaction model by breaking down the chemical reactions into substrate–enzyme, product–enzyme, and cofactor–enzyme groups. ENZYME links to BRENDA, EMP/PUMA, WIT, and KEGG. The database is free and is run by the not-for-profit Swiss Institute of Bioinformatics. There are no restrictions on its use by any institutions as long as its content is not modified in any way.

FIMM (Functional Molecular Immunology)
URL: http://sdmc.krdl.org.sg:8080/fimm/
Ref. [51]

The FIMM database contains information about functional immunology. It is primarily not an interaction database but contains information about major histocompatibility complex (MHC)/human leukocyte antigen (HLA) associated peptides, antigens, and diseases. The database contains information about more than 1400 peptides and almost 1400 HLA records at time of writing. It is linked to GenBank, SWISS-PROT, MHCPEP, OMIM, and PubMed, among others. This data provides records of protein–peptide interactions that are important immunologically and some records contain HLA class I structure models. The database is provided "as-is" by Kent Ridge Digital Labs in Singapore.

FlyNets
Ref. [52]

FlyNets is now defunct, but originally stored information about molecular interactions (protein–DNA, protein–RNA, and protein–protein interactions) and genetic interaction networks in the fruit fly, *Drosophila melanogaster*, focusing on developmental pathways. Information was linked to PubMed and FlyBase. Version 3.0 was available in May 1999 and contained 200 interactions. FlyNets was based on a graph abstraction and provided a visual graph navigation tool to draw networks from the database.

GeneNet (Genetic Networks)
URL: http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/
Refs. [53, 54]

GeneNet describes genetic regulatory networks from gene through cell to organism level using a chemical reaction based formalism, i.e. substrates, entities affecting course of reaction and products. The database is based on a formal object-oriented data model. GeneNet contains 23 gene network diagrams and over 1000 genetic interactions (termed relations in GeneNet) from a variety of organisms including *Homo sapiens*. The database is current and is regularly updated. Visual tools are present for examining and querying the pathway data in the context of a simple diagram of a cell, but are plagued by network latency problems that can prevent complete loading.

GeNet (Gene Networks Database)
URL: http://www.csa.ru/Inst/gorb_dep/inbios/genet/genet.htm
Ref. [55]

GeNet curates genetic regulatory networks for a few example species. It provides Java visualization tools for the genetic networks. The database contains extensive information about each example network in free-text form. This database is not machine-readable, although is a good genetic interaction resource.

GRID (General Repository for Interaction Datasets)
URL: http://biodata.mshri.on.ca/grid
Ref. [56]

GRID, or the general repository for interaction datasets, contains protein–protein and genetic interactions, currently for budding yeast (*Saccharomyces cerevisiae*), YeastGRID, fruit fly (*Drosophila melanogaster*), FlyGRID, and *Caenorhabditis elegans*, WormGRID. GRID is actively being expanded and other species might be available in the near future. GRID provides a simple summary of each interaction along with gene names, gene ontology (GO) annotation, and the experimental system used to determine the interaction. A network visualization tool called Osprey is also available for visualizing, browsing, and analyzing the interaction networks of the various GRIDs.

HIV Molecular Immunology Database
URL: http://hiv-web.lanl.gov/immunology/index.html
Ref. [57]

This database contains information about binding events between HIV and the immune system including HIV epitope and antibody binding sites that could provide data for an interaction database. HLA binding motifs are included and enable prediction of HLA–peptide interactions. This information is freely available from the database's FTP site.

HPRD (Human Protein Reference Database)
URL: http://hprd.org
Ref. [58]

HPRD, or the human protein reference database, is a recently released database of human proteins, but also contains a significant amount of information about protein– protein interactions. Information about the domain and region of interaction, if available, is present, as is the type of experiment performed to detect the interaction. Expression, domain architecture, and post-translational modifications are also curated for each protein. Several curated pathways created from the interaction data are available as images. HPRD data can be browsed and searched by a number of common database fields and by BLAST over the Web. Data are freely available to academics in the PSI-MI format but commercial users require a license.

HOX Pro
URL: http://www.iephb.nw.ru/hoxpro
Ref. [59]

The main purpose of this database is to provide a curated human-readable resource for homeobox genes. It also stores extensive information about genetic regulatory networks of homeobox genes for a few model organisms. Clickable pictures and a Java applet are available for visualizing the networks. The visualization system is the same as that used for GeNet.

InBase (The Intein Database)
URL: http://www.neb.com/neb/inteins.html
Ref. [60]

The main purpose of this database is as a curated resource for protein splicing. The

database contains descriptions of intein proteins (self-catalytic proteins) that are good examples of intramolecular interactions. The database records are present in a machine-readable format. Each record could be used by an interaction database to generate intramolecular interaction records containing chemical reaction description using information about the mechanism of protein splicing present on the InBase website.

Indigo
URL: http://195.221.65.10:1234/Indigo/

Indigo contains information about codon usage, operons, gene neighbors, and metabolic pathways for *Escherichia coli* and *Bacillus subtilis*. The metabolic pathway information contains information about enzymatic reactions and can be accessed using clickable images in a Java applet. Enzyme names in the pathway map are linked back to primary sequence databases.

IntAct
URL: http://www.ebi.ac.uk/intact
Ref. [61]

IntAct is a relatively new database of freely available protein interactions maintained by the European Bioinformatics Institute (EBI). An initial implementation available from mid-2003 focuses on protein–protein interactions collected from large-scale published studies and some literature. It enables searching by protein name and browsing using a graphical network interface. One difference between the IntAct data model and those of many other protein–protein interaction databases is that interactions are not necessarily binary, but rather are sets. The advantage of using sets to store interactions is that they can represent certain types of protein complex data where information is not known about the exact physical interactions in the complex, but only that the set of proteins co-purifies. Representing information this way has become more important since the release of large-scale biochemical co-purification studies [62, 63]. Data are available in the PSI-MI XML format.

Interact
Ref. [64]

Interact is an object-oriented protein–protein-interaction database based on Java and the POET database (www.poet.com) that is now defunct. It has a formal data-model that describes interactions, molecular complexes and genetic interactions. It stores information about experimental methods and is based on an object-oriented description of proteins and genes. The database does not provide other details about the interaction and the underlying description of genes and proteins is simplified compared with that of GenBank. The database is not publicly available, but the object-oriented design approach has been described in the literature. The database contains over 1000 interactions.

ICBS (Inter-Chain Beta-Sheets)
URL: http://www.igb.uci.edu/servers/icbs/
Ref. [65]

ICBS contains protein–protein interactions mediated by beta-sheets taken from the PDB database. The database contains over 3600 PDB structures that contain protein complexes mediated by this type of interaction. Basic information about each PDB file is provided, as is detailed physical and structural information about the beta sheets at the interaction interface. This database is similar to MMDBind, but is a more highly curated subset.

JenPep
URL: http://www.jenner.ac.uk/JenPep/
Ref. [66]

JenPep is a peptide binding database that contains more than 8000 peptide–protein interactions for MHC Class I, II, CD8, and CD4 T cells and TAP (transport of antigen) complex. All information in JenPep, for example IC50 and peptide origin, is from published experiments. Peptide epitopes can be searched over the web using a simple query interface.

KEGG (Kyoto Encyclopedia of Genes and Genomes)
URL: http://www.genome.ad.jp/kegg/
Ref. [67]

KEGG depicts many known metabolic pathways and some regulatory pathways for many different species as graphical diagrams that are manually drawn and updated. Each of the metabolic pathway drawings is intended to represent all chemically feasible pathways for a given system. As such, these pathways are abstractions on to which enzymes and substrates from specific organisms can be mapped. KEGG stores reactions mediated by each enzyme in the database and these are linked to from the pathway maps. The database is machine-readable, except for the pathway diagrams. Each enzyme entry contains a substrate and a product field that can be used to translate between the chemical reaction description scheme and a binary interaction scheme. The KEGG project distributes all databases freely for academics via FTP. KEGG is one of the best freely available resources of metabolic and small molecule information (the LIGAND database).

Kohn Molecular Interaction Maps
URL: http://discover.nci.nih.gov/kohnk/interaction_maps.html
Ref. [68]

Kohn molecular maps are one researcher's attempt to create a standard for representing biochemical pathways and molecular interactions using a symbolic language similar to electronic circuit diagrams. Kohn created detailed maps of the mammalian cell-cycle control and DNA repair systems as examples. The maps are pictures only and thus are not machine-readable, although they do have a grid system as in normal street maps. A separate annotation list is provided that enables mapping of molecules from the list of the map using the coordinate system. The ideas represented in these maps are useful for further research on pathway visualization systems and the first two maps provide a resource for manual extraction of molecular interaction information.

MDB (Metalloprotein Database)
URL: http://metallo.scripps.edu/
Ref. [69]

MDB contains the metal-binding sites from entries in the PDB database. The database is based on open-source software and is freely available. The data are present down to the atomic level of detail. An extensive Java applet is available to query and examine the data in detail. *Ad-hoc* queries of the database using SQL are available and tools are being developed to predict a metal binding site in a given protein structure.

MHCPEP
URL: http://wehih.wehi.edu.au/mhcpep/
Ref. [70]

MHCPEP is a database containing over 13,000 peptide sequences known to bind MHC molecules compiled from the literature and from direct submissions. It has not been updated since mid-1998. Although this database is not a typical interaction database, it provides peptide–protein interaction information relevant to immunology. The database is freely available via FTP in a text-based machine-readable format.

MINT (Molecular-interaction database)
URL: http://mint.bio.uniroma2.it/mint/
Ref. [71]

MINT is a database of molecular interactions gathered from the literature and manually input. Apart from a simple relational scheme for storage of set relationships among proteins, MINT can store some protein post-translational modifications, experiments, cellular location, pathways, and complexes. MINT contains more than 42,000 protein interactions and only a handful of complexes. An extensive graph abstraction is present which enables the use of a graphical Java viewer for the interactions. Interestingly, the size of the molecules is represented relative to each other in the visualization, so heavier proteins are drawn as larger circles.

MIPS Comprehensive Yeast Genome Database
URL: http://mips.gsf.de/proj/yeast/
Ref. [72]

The MIPS comprehensive yeast genome database (CYGD) summarizes current knowledge about the more than 6200 ORF encoded by the yeast genome. This database is similar to SGD and YPD and is not primarily an interaction database. The MIPS center, however, makes available large tables for direct protein–protein interactions and genetic interactions in yeast free for download at http://mips.gsf.de/proj/yeast/tables/interaction/index.html. Each interaction contains an experimental method used and, usually, a literature reference. Manually created clickable pathway maps are also available for a variety of metabolic and regulatory pathways in yeast. The MIPS yeast genome database uses a relational model, but most fields use unstructured text. For example, the experimental method used to determine the interaction field is unstructured and the same experimental type can be represented in many different ways. Although this makes the database difficult to parse with a computer, the CYGD is an extremely useful resource for yeast protein–protein interaction information. Recently, MIPS has made available a protein–protein interaction, complex, and genetic interaction query tool for searching this data.

MMDB (Molecular Modeling Database)
URL: http://www.ncbi.nlm.nih.gov/Structure/
Ref. [73]

This database is an NCBI resource that contains all the data in the PDB database in ASN.1 form. The MMDB validates all PDB file information and describes all atomic level detail data explicitly and in a formal machine-readable manner. Although this database is not an interaction database, it does contain atomic level detail of molecular interactions present in some records that describe molecular complexes. Sequence linkage is improved and MMDB is readily accessed by machine-readable methods that can obtain information about molecular interactions. MMDB is in the public domain and all software and data are freely available to academics or corporations.

NetBiochem
URL: http://medlib.med.utah.edu/NetBiochem/NetWelco.htm

NetBiochem is primarily an education resource that focuses on teaching detailed biochemistry of specific metabolic pathways, for example fatty acid metabolism, at the level of an introductory biochemistry course at a university. There is no formal data model, but the available pathways represent a good collection of different ways of presenting biochemical pathway data to an untrained audience. This site would, therefore, be useful as a resource for curators to

enter data into a molecular-interaction database and as a source of ideas for pathway-visualization research.

ooTFD (Object Oriented Transcription Factors Database)
URL: http://www.ifti.org/
Ref. [74]

The ooTFD contains information on transcription factors from a variety of organisms, including transcription factor binding sites on DNA and transcription factor molecular complex information. Thus it contains protein–DNA and protein–protein interactions. The database is based on a formal machine-readable object-oriented format and is available in numerous forms. The database contains thousands of sites and transcription factors and is freely available (including software) from http://ncbi.nlm.nih.gov/repository/TFD/.

ORDB (Olfactory Receptor Database)
URL: http://senselab.med.yale.edu
Ref. [75]

The ORDB is primarily a database of sequences of olfactory receptor proteins. It contains a section on small molecule ligands that bind to olfactory receptors. About 100 ligand–protein interactions are present in the database with about 130 small molecules. Structures of these small molecules are also available, in the associated OdorDB.

PATIKA (Pathway Analysis Tool for Integration and Knowledge Acquisition)
URL: http://www.patika.org/
Ref. [76]

PATIKA is a combination of a Java pathway modeling tool and an object-oriented pathway database. A data specification is present using a state and transition notion for pathway descriptions. This data model combines elements from BIND, EcoCyc, and Petri Nets. Interestingly, the data model allows multiple levels of abstraction to enable the description of cellular events when not all of the details are known. For instance, a transition can describe the change of one state to another and that state can be very detailed chemically or can be a very general cellular state. The Java tool enables one to build pathways and query the database remotely over the Internet. The data model is currently quite simple and is only designed to store human pathway information.

PhosphoBase
URL: http://www.cbs.dtu.dk/databases/PhosphoBase/
Ref. [77]

This database contains information on kinases and phosphorylation sites. The phosphorylation sites are stored with kinetic information and references for each kinase. Although this is not an interaction database directly, information is present about protein–protein interactions involved in cell signaling and their chemistry. A neural network-based phosphorylation site-prediction tool has recently been made available.

PIMRider (Protein Interaction Map – Hybrigenics)
URL: http://pim.hybrigenics.com/

PIMRider is a graphical Java applet-based protein interaction network visualization tool driven by a database of protein–protein interactions. All interactions have been determined using the sequence fragment (domain)-based two-hybrid screen experimental approach by the Hybrigenics company. All of the data and the data model are proprietary and only partially publicly available. The PIM database contains information on *Helicobacter pylori*, HIV (human immunodeficiency virus), HCV (hepatitis C virus) and *Homo sapiens*.

PIMdb (Drosophila Protein Interaction Map Database)
URL: http://proteome.wayne.edu/PIMproject1.html

PIMdb is a collection of two-hybrid generated protein–protein interactions for *Drosophila melanogaster*. A single laboratory is generating these data, which are currently unpublished. A simple binary interaction data model is used to store the information. Currently, PIMdb does not make available any query tools, but is rather just a manually created list of experimental results from one academic research group. Without peer review, the quality of the data is in question. The group asks to be contacted if any of their data are used for other projects.

ProChart (Axcell)
URL: http://www.axcellbio.com/products.asp

The ProChart database is sold by Axcell Biosciences and contains proprietary data on protein–protein interactions garnered using Axcell's proprietary experimental methods. No part of the database or data model is publicly accessible or has been published.

ProNet (Myriad Genetics)

This commercial database provides protein–protein interaction information to the public from Myriad Genetics proprietary high-throughput yeast two-hybrid system for human proteins and from published literature. Each protein record describes interacting proteins and a Java applet is available for navigating the database. The database stores only protein-interaction information with links to primary sequence databases and PubMed. It uses a graph abstraction to display the interactions. The database is fully proprietary and has not been published.

REBASE
URL: http://rebase.neb.com
Ref. [13]

REBASE is a comprehensive database of information about restriction enzymes and related proteins, for example methylases. Although it is not an interaction database, restriction enzymes and methylases participate in specific DNA–protein interactions. REBASE describes the enzyme and the recognition site, and thus can be used to create binary interaction records with chemical actions. Useful links are present to commercially available enzymes. REBASE is freely available to the academic community in many different formats.

Relibase
URL: http://relibase.ebi.ac.uk/
Ref. [78]

Relibase is a software query tool that enables powerful searches to be conducted on PDB entries containing protein–ligand interactions, where a ligand is anything that is not a protein. DNA and RNA are also considered ligands, but are ignored in searches. The purpose of Relibase is to help examine small molecules, for example therapeutics, that are currently in the PDB as binding to proteins. Full crystal structure and binding sites of ligands are available. The database can be searched by text, sequence, SMILES strings, and 2D/3D small molecule structures. The Relibase project is currently run by the Cambridge Crystallographic Data Centre, which makes the tool available over the web.

RegulonDB
URL: http://www.cifn.unam.mx/Computational_Genomics/regulondb/
Ref. [79, 80]

RegulonDB is mainly an *E. coli* operon database, although it does contain protein–DNA interactions (e.g. ribosome binding sites and promoters) and protein complexes. The database is free for non-commercial use. Commercial users require a license.

SELEX_DB
URL: http://wwwmgs.bionet.nsc.ru/mgs/systems/selex/
Ref. [81]

SELEX_DB is a curated resource that stores experimental data for functional site sequences obtained by using SELEX-like random sequence pool technologies to study interactions. The database contains interactions, including binding sites, between random DNA sequences and various types of ligand, most of which are proteins. It is available over the web and via SRS and the records are available in a machine-readable flat-file format.

SoyBase
URL: http://soybase.ncgr.org/

SoyBase is an ACeDB [42] database that contains information about the soybean, including metabolism. Metabolic pathways are based on a chemical reaction abstraction. SoyBase contains over 850 automatically generated diagrams of metabolic pathways covering over 1500 enzymes and more than 1200 metabolites. Clicking on an enzyme or ligand on the diagram triggers a query for that molecule in the database. SoyBase is based on a formal machine-readable data model, as is any AceDB installation, and is available over the web.

SPAD (Signaling Pathways Database)
URL: http://www.grt.kyushu-u.ac.jp/eny-doc/

SPAD provides clickable image maps for a handful of pathways. Clicking on an element of the pathway diagram links to sequence information of the protein or gene. Protein–protein and protein–DNA interactions are covered with regard to signal transduction. The database does not have a formal data model. SPAD has not been updated since 1998 but still gives useful overviews of the pathways it contains.

SPIN-PP (Surface Properties of Interfaces – Protein–Protein Interfaces)
URL: http://trantor.bioc.columbia.edu/cgi-bin/SPIN/

SPIN-PP is a database of all protein–protein interfaces in the PDB. Molecular surfaces are organized in a taxonomy based on surface curvature, electrostatic potential, sequence variability, and hydrophobicity. SPIN-PP contains 855 protein–protein interfaces and is searchable by PDB code and the various surface structural properties listed above. Surfaces of interest can be viewed using the GRASS server [82]. The database does not seem to have been updated regularly since 1999, but is freely available.

STKE (Signal Transduction Knowledge Environment)
URL: http://www.stke.org/
Ref. [83]

STKE is a curated resource for signal-transduction information. It provides a manually created clickable picture of various signal transduction pathways linked to primary database, the Connections Map. The data model is based on an upstream and downstream components view, which is a graph abstraction. Database fields are unstructured and thus are not machine-readable. STKE is available via a paid subscription to Science magazine.

SYFPEITHI
URL: http://www.uni-tuebingen.de/uni/kxi/
Ref. [84]

SYFPEITHI is a database of MHC ligands and peptide motifs. It contains over 3500 peptide sequences known to bind class I and class II MHC molecules. All entries have been compiled from the literature. Although this database is not a typical interaction database, it provides peptide–protein interaction information relevant to immunology.

TRANSFAC
URL: http://www.gene-regulation.com/
Ref. [85]

TRANSFAC is a database of transcription factors containing genomic binding sites and DNA-binding profiles. As such, it is not a typical interaction database, but it does contain protein–DNA interactions. A transcription factor DNA-binding site prediction tool is available. TRANSFAC is available via license.

TRANSPATH
URL: http://transfac.gbf.de/TRANSFAC/
Ref. [85]

TRANSPATH is an effort underway at TRANSFAC to link regulatory pathways to transcription factors. The database is based on a chemical reaction view of interactions and contains a strong graph abstraction. Graph algorithms have been implemented to enable navigation of the data. The database can describe regulatory pathways, their components, and the cellular locations of those components. It can store information about various species. TRANSPATH includes all of the data from the CSNDB and it is obvious that TRANSPATH is using graph theory ideas from the CSNDB. TRANSPATH is available via license.

TRRD (Transcription Regulatory Regions Database)
URL: http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/
Ref. [86]

TRRD contains information about regulatory regions including over 3600 transcription factor binding sites (DNA–protein interactions). This database is very similar to TRANSFAC. It is freely available over the web via an SRS database interface.

WIT (What Is There?)
URL: http://wit.mcs.anl.gov/WIT2
Ref. [87]

WIT is a database project whose purpose is to reconstruct metabolic pathways in newly sequenced genomes by comparing predicted proteins with proteins in known metabolic networks. Predicted metabolic networks are stored in a chemical reaction-based scheme with a graph abstraction. All information in the database can be queried and pathways can be viewed as a computer-generated diagram which is hyper-linked back to the database.

YPD (Yeast Proteome Database – Incyte Genomics)
URL: https://www.incyte.com/proteome/index.html
Ref. [88]

This proprietary commercial curated proteome database from Incyte contains extensive information about all known proteins in yeast. Extensive data about protein interactions, molecular complexes, and subcellular location is present. Most of the database fields are free-form text, but there is enough structure in the data model to make it amenable to machine reading of protein–protein interaction information. Incyte also makes available other proteomes for other model organisms including *Caenorhabditis elegans* and Human, but YPD is the most completely annotated in the Proteome BioKnowledge Library. All of Incyte's proteome databases are proprietary and are available on a subscription basis.

Acknowledgements
GB thanks his supervisor Chris Sander for support during the writing of this chapter update. The original chapter was written in the laboratory of Christopher Hogue for the first edition of this book.

References

1 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

2 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di F, V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, et al. (2001) The sequence of the human genome. Science 291:1304–1351

3 Jones S, Thornton JM (1996) Principles of protein–protein interactions. Proc. Natl. Acad. Sci. USA 93:13–20

4 Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300:445–452

5 Bader GD, Hogue CW (2000) BIND – a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics. 16:465–477

6 Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW (2001) BIND – The Biomolecular Interaction Network Database. Nucleic Acids Res. 29:242–245

7 Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 31:248–250

8 Ostell J, Kans JA (1998) The NCBI Data Model. In: Baxevanis AD, Ouellette BF (eds) Bioinformatics, a Practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons, p 121–144

9 Wang Y, Addess KJ, Geer L, Madej T, Marchler-Bauer A, Zimmerman D, Bryant SH (2000) MMDB: 3D structure data in Entrez. Nucleic Acids Res. 28:243–245

10 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res. 28:235–242

11 Stein L (2002) Creating a bioinformatics nation. Nature 417:119–120

12 Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2003) GenBank. Nucleic Acids Res. 31:23–27

13 Roberts RJ, Vincze T, Posfai J, Macelis D (2003) REBASE: restriction enzymes and methyltransferases. Nucleic Acids Res 31:418–420

14 Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R (2004) The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. Nat Biotechnol 22:177–183

15 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504

16 Pawson T, Scott JD (1997) Signaling through scaffold, anchoring, and adaptor proteins. Science 278:2075–2080

17 Cesareni G, Panni S, Nardelli G, Castagnoli L (2002) Can we infer peptide recognition specificity mediated by SH3 domains? FEBS Lett 513:38–44

18 Fedorov AA, Fedorov E, Gertler F, Almo SC (1999) Structure of EVH1, a novel proline-rich ligand-binding module involved in cytoskeletal dynamics and neural function. Nat Struct Biol 6:661–665

19 Macias MJ, Wiesner S, Sudol M (2002) WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. FEBS Lett 513:30–37

20 de Beer T, Hoofnagle AN, Enmon JL, Bowers RC, Yamabhai M, Kay BK, Overduin M (2000) Molecular mechanism of NPF recognition by EH domains. Nat Struct Biol 7:1018–1022

21 Salcini AE, Confalonieri S, Doria M, Santolini E, Tassi E, Minenkova O, Cesareni G, Pelicci PG, Di Fiore PP (1997) Binding specificity and in vivo targets of the EH domain, a novel protein–protein interaction module. Genes Dev 11:2239–2249

22 Moran MF, Koch CA, Anderson D, Ellis C, England L, Martin GS, Pawson T (1990) Src homology region 2 domains direct protein–protein interactions in signal transduction. Proc Natl Acad Sci U S A 87:8622–8626

23 Durocher D, Jackson SP (2002) The FHA domain. FEBS Lett 513:58–66

24 Paoluzi S, Castagnoli L, Lauro I, Salcini AE, Coda L, Fre S, Confalonieri S, Pelicci PG, Di Fiore PP, Cesareni G (1998) Recognition specificity of individual EH domains of mammals and yeast. Embo J 17:6541–6550

25 Panni S, Dente L, Cesareni G (2002) In Vitro Evolution of Recognition Specificity Mediated by SH3 Domains Reveals Target Recognition Rules. J Biol Chem 277:21666–21674

26 Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, Quondam M, Zucconi A, Hogue CW, Fields S, Boone C, Cesareni G (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295:321–324

27 van Helden J, Naim A, Lemer C, Mancuso R, Eldridge M, Wodak SJ (2001) From molecular activities and processes to biological function. Brief. Bioinform. 2:81–93

28 Lemer C, Antezana E, Couche F, Fays F, Santolaria X, Janky R, Deville Y, Richelle J, Wodak SJ (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. Nucleic Acids Res 32 Database issue:D443–D448

29 Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 30:47–49

30 Szymanski M, Barciszewski J (2000) Aminoacyl-tRNA synthetases database Y2K. Nucleic Acids Res. 28:326–328

31 Thorn KS, Bogan AA (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics. 17:284–285

32 Becker KG, White SL, Muller J, Engel J (2000) BBID: the biological biochemical image database. Bioinformatics 16:745–746

33 Chen X, Lin Y, Gilson MK (2001) The binding database: Overview and user's guide. Biopolymers 61:127–141

34 Chen X, Liu M, Gilson MK (2001) BindingDB: a web-accessible molecular recognition database. Comb. Chem. High Throughput Screen. 4:719–725

35 Chen X, Lin Y, Liu M, Gilson MK (2002) The Binding Database: data management and interface design. Bioinformatics. 18:130–139

36 Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L (2003) Database resources of the National Center for Biotechnology. Nucleic Acids Res. 31:28–33

37 Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370

38 Ellis LB, Hershberger CD, Bryan EM, Wackett LP (2001) The University of Minnesota Biocatalysis/Biodegradation Database: emphasizing enzymes. Nucleic Acids Res. 29:340–343

39 Schomburg I, Chang A, Hofmann O, Ebeling C, Ehrentreich F, Schomburg D (2002) BRENDA: a resource for enzyme data and metabolic information. Trends Biochem. Sci. 27:54–56

40 Kel-Margoulis OV, Romashchenko AG, Kolchanov NA, Wingender E, Kel AE (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. Nucleic Acids Res. 28:311–315

41 Takai-Igarashi T, Nadaoka Y, Kaminuma T (1998) A database for cell signaling networks. J. Comput. Biol. 5:747–754

42 Eeckman FH, Durbin R (1995) ACeDB and macace. Methods Cell Biol. 48:583–605

43 Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G, Sethuraman A, Weng S, Botstein D, Cherry JM (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res. 30:69–72

44 Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 30:303–305

45 Baranov PV, Kubarenko AV, Gurvich OL, Shamolina TA, Brimacombe R (1999) The Database of Ribosomal Cross-links: an update. Nucleic Acids Res. 27:184–185

46 Robison K, McGuire AM, Church GM (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. J. Mol. Biol. 284:241–254

47 Hofmann K, Bucher P, Falquet L, Bairoch A (1999) The PROSITE database, its status in 1999. Nucleic Acids Res. 27:215–219

48 Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S (2002) The EcoCyc Database. Nucleic Acids Res. 30:56–58

49 Selkov E, Basmanova S, Gaasterland T, Goryanin I, Gretchkin Y, Maltsev N, Nenashev V, Overbeek R, Panyushkina E, Pronevitch L, Selkov E, Jr., Yunus I (1996) The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. Nucleic Acids Res. 24:26–28

50 Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res. 28:304–305

51 Schonbach C, Koh JL, Sheng X, Wong L, Brusic V (2000) FIMM, a database of functional molecular immunology. Nucleic Acids Res. 28:222–224

52 Sanchez C, Lachaize C, Janody F, Bellon B, Roder L, Euzenat J, Rechenmann F, Jacq B (1999) Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. Nucleic Acids Res. 27:89–94

53 Kolpakov FA, Ananko EA, Kolesov GB, Kolchanov NA (1998) GeneNet: a gene network database and its automated visualization. Bioinformatics. 14:529–537

54 Kolpakov FA, Ananko EA (1999) Interactive data input into the GeneNet database. Bioinformatics. 15:713–714

55 Serov VN, Spirov AV, Samsonova MG (1998) Graphical interface to the genetic network database GeNet. Bioinformatics. 14:546–547

56 Breitkreutz BJ, Stark C, Tyers M (2003) The GRID: the General Repository for Interaction Datasets. Genome Biol. 4:R23

57 Korber B, Brander C, Haynes B, Koup R, Moore J, Walker B (1998) HIV Molecular Immunology Database 1998. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM

58 Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A (2003) Development of Human Protein Reference Database as an initial platform for approaching systems biology in humans. Genome Res 13:2363–2371

59 Spirov AV, Bowler T, Reinitz J (2000) HOX Pro: a specialized database for clusters and networks of homeobox genes. Nucleic Acids Res. 28:337–340

60 Perler FB (2000) InBase, the Intein Database. Nucleic Acids Res. 28:344–345

61 Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R (2004) IntAct: an open source molecular-interaction database. Nucleic Acids Res 32:D452–D455

62 Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180–183

63 Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147

64 Eilbeck K, Brass A, Paton N, Hodgman C (1999) INTERACT: an object oriented protein–protein interaction database. Ismb.:87–94

65 Baisnee PF, Pollastri G, Pecout Y, Nowick J, Baldi P (2002) ICBS: A Database of Protein–Protein Interactions Mediated by Interchain Beta-Sheet Formation. 10th International Conference on Intelligent Systems for Molecular Biology (ISMB)

66 Blythe MJ, Doytchinova IA, Flower DR (2002) JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics. 18:434–439

67 Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32 Database issue:D277–D280

68 Kohn KW (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. Mol. Biol. Cell 10:2703–2734

69 Castagnetto JM, Hennessy SW, Roberts VA, Getzoff ED, Tainer JA, Pique ME (2002) MDB: the Metalloprotein Database and Browser at The Scripps Research Institute. Nucleic Acids Res. 30:379–382

70 Brusic V, Rudy G, Harrison LC (1998) MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res. 26:368–371

71 Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) MINT: a Molecular-interaction database. FEBS Lett. 513:135–140

72 Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B (2002) MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 30:31–34

73 Wang Y, Anderson JB, Chen J, Geer LY, He S, Hurwitz DI, Liebert CA, Madej T, Marchler GH, Marchler-Bauer A, Panchenko AR, Shoemaker BA, Song JS, Thiessen PA, Yamashita RA, Bryant SH (2002) MMDB: Entrez's 3D-structure database. Nucleic Acids Res. 30:249–252

74 Ghosh D (2000) Object-oriented transcription factors database (ooTFD). Nucleic Acids Res. 28:308–310

75 Crasto C, Marenco L, Miller P, Shepherd G (2002) Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. Nucleic Acids Res. 30:354–360

76 Demir E, Babur O, Dogrusoz U, Gursoy A, Nisanci G, Cetin-Atalay R, Ozturk M (2002) PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. Bioinformatics. 18:996–1003

77 Kreegipuu A, Blom N, Brunak S (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. Nucleic Acids Res. 27:237–239

78 Hendlich M, Bergner A, Gunther J, Klebe G (2003) Relibase: design and development of a database for comprehensive analysis of protein–ligand interactions. J Mol Biol 326:607–620

79 Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. Nucleic Acids Res. 29:72–74

80 Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C, Collado-Vides J (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. Nucleic Acids Res 32 Database issue:D303–D306

81 Ponomarenko JV, Orlova GV, Frolov AS, Gelfand MS, Ponomarenko MP (2002) SELEX_DB: a database on in vitro selected oligomers adapted for recognizing natural sites and for analyzing both SNPs and site-directed mutagenesis data. Nucleic Acids Res. 30:195–199

82 Nayal M, Hitz BC, Honig B (1999) GRASS: a server for the graphical representation and analysis of structures. Protein Sci. 8:676–679

83 Gough NR, Ray LB (2002) Mapping cellular signaling. Sci. STKE. 2002:EG8

84 Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219

85 Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, Pruss M, Schacherer F, Thiele S, Urbach S (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Res. 29:281–283

86 Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res. 30:312–317

87 Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Jr, Kyrpides N, Fonstein M, Maltsev N, Selkov E (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 28:123–125

88 Costanzo MC, Crawford ME, Hirschman JE, Kranz JE, Olsen P, Robertson LS, Skrzypek MS, Braun BR, Hopkins KL, Kondu P, Lengieza C, Lew-Smith JE, Tillberg M, Garrels JI (2001) YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. Nucleic Acids Res. 29:75–79

19
Bioinformatics Approaches for Metabolic Pathways

Ming Chen, Andreas Freier, and Ralf Hofestädt

19.1 Introduction

In living organisms metabolism is necessary to sustain life and for reproduction. Metabolism, defined as the sum of all the chemical transformations taking place in a cell or organism, occurs in a series of enzyme-catalyzed reactions that constitute metabolic pathways [1]. To survive and grow, cells must be able to reproduce the whole cell and replace its constituent parts, and to perform a varying range of chemical functions. To perform these chemical functions the cell uses biomolecules called proteins, which are encoded in the DNA. DNA is a long, threadlike molecule with the shape of a double helix, made up of connected subunits called nucleotides. The sequences of DNA that code for proteins are called genes. To synthesize a protein from a gene the DNA must first be transcribed into RNA. When the RNA has been copied, proteins can be produced from the sequence in a process called translation. A protein is made up of amino acids. For each amino acid there is at least one corresponding triplet of nucleotides, called a codon. During the translation process, the strand of RNA is read in frames of length three, each time adding an amino acid to the growing protein sequence. The proteins which carry out the functions of breaking down and synthesizing new biochemicals are called enzymes, each of which usually catalyzes only one type of reaction. We can then have a series of reactions. Many molecules involved in one reaction can also be found in other reactions in which the molecules act as substrate, activator, or repressor, thus forming a densely connected, intricate and precisely regulated reaction network. Thousands of enzyme-catalyzed reactions occur every second in a living cell. The network of reactions is very large and complex. This metabolic network enables living organisms to synthesize, interconvert, and breakdown molecules required by the cell.

Some genes code for proteins that turn other genes on and off. Groups of these genes constitute networks with complex behavior. These networks control other genes whose protein products catalyze specific biochemical reactions, and the small molecules that are substrates or products of these reactions can in turn activate or deactivate proteins that control transcription or translation. For this reason gene regulation

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

can be said to indirectly control biochemical reactions in cellular metabolism, and cellular metabolism itself exerts control on gene expression. As a result, these connected reactions are normally called gene-regulated metabolic networks. For these reasons, the interdependent biochemical processes of metabolism and gene expression can and should be interpreted and analyzed in terms of complex dynamic networks. Hence modeling and simulation are necessary. Analysis of metabolic pathways is a central topic in understanding the relationship between genotype and phenotype. Molecular biologists and bioinformatists have begun the studies on gene expression [2], gene control [3], and metabolic regulation [4].

As a result of great achievements in the pre-genomic era, more and more experimental data have been systematically collected and stored in specific databases, for example gene sequence (e.g. GenBank [5], EMBL [6], DDBJ [7]), protein (e.g. SwissProt [8], PIR [9], BRENDA [10]), biochemical reactions (e.g. KEGG [11], WIT/EMP [12]), regulatory genes (e.g. GeneNet [13]), transcription factors (e.g. TransFac [14], EPD [3]) and signal-induction reactions (e.g. CSNDB [15], TransPath [16]). URL of these databases are appended at the end of this chapter. This rapid accumulation of biological data provides the possibility of studying metabolic pathways at both genome and metabolic levels. To improve our understanding of cells and organisms as physiological, biochemical, and genetic systems we have to study the whole system, an integrative metabolism system. One major task in the post-genomic era is to implement molecular information systems that will enable integration of different molecular database systems and the design of analysis tools (e.g. simulators of complex biochemical reactions). Effective possibilities for database integration are provided by World Wide Web (WWW) technology. One of the most developed technologies of WWW integration of molecular databases uses the Sequence Retrieval System (SRS) [17]. This is based on local copies of each component database, which must be provided in a text-based format. The results of the query are sets of WWW links. Users can navigate through these links. Within the framework of this approach, however, data fusion is still a task for the user. We also do not find real data fusion; i.e. data for one real world object (e.g. an enzyme) coming from two different databases (e.g. KEGG and BRENDA) is represented twice by different WWW page objects. Therefore research groups try to integrate molecular databases on a higher level than the SRS approach. The first step toward that goal is the integration of databases under a specific biological perspective. The next step will be user-defined molecular information fusion. No special systems are yet available for successful fulfillment of both objectives.

Beyond databases, simulators for metabolic networks that employ most of the currently popular modeling methods are also available via the Internet. In addition to the classical methods of differential equations, discrete methods have become quite important. Examples are graphs [18], knowledge-based simulation [19], Petri nets [20], rule-based systems [21], object-oriented approaches [22], and Boolean nets [23]. Because of the extreme complexity of metabolic networks, there are less prominent tools that can model, simulate, and analyze entire aspects of cell behavior.

Hence, one of the next goals of bioinformatics is implementation of a uniform environment for a homogeneous access to different databases and the presentation and analysis of metabolic networks under specific biological perspectives. With these considerations in mind we developed a soft-

ware system to integrate access to different databases, and exploited the Petri net method to model and simulate metabolic networks in the cell.

19.2 Formal Representation of Metabolic Pathways

Metabolic pathways are normally represented as graphs [24]. This is because biochemical reaction information is rarely collected and handled verbally in scientific texts. Researchers list the reactions and draw their interrelation graph by hand in their original research papers. If, however, one wants to conduct a general survey of biochemical reactions and metabolic interrelationships, it is difficult and time-consuming to extract their overview from numerous lists of single reactions. Computer-based graphics enable rapid drawing and retrieval of information details.

Traditionally, a metabolic pathway is represented as a directed reaction graph with substrates as vertices and directed, labeled edges denoting reactions between substrates catalyzed by enzymes (labels). Several well known databases, for example Swiss-Prot and KEGG have been developed to present diagrams depicting metabolic pathways (networks) which provide online maps of metabolic pathways and the ability to focus on metabolic reactions in specific organisms. Normally, a metabolic pathway is said to be a subset of those reactions that describe the biochemical conversion of a given reactant to its desired end product. It can also be said that a metabolic pathway is a special case of a metabolic network with distinct start and end points, initial and terminal vertices, respectively, and a unique path between them [25].

Let $M = \{m_1,m_2,...,m_n\}$ be a set of metabolites that are involved in the enzymatic reactions by acting as substrates and products; let $E = \{e_1,e_2,...e_m\}$ be a set of enzymes. In metabolic reactions, enzymes act as catalysts in the conversion of some metabolites (substrates) into other metabolites (products). The representation of a metabolic pathway P might be given graphically as a set of related E', i.e. $P = (E',A)$, where $E' \subseteq E$, A is a set of edges of the form (u,v), where $u,v \in E$.

Current metabolic pathway databases have several limitations. They are based on the idea of the static representation of molecular data and knowledge. Some contain only well known pathways and lack automatic construction of dynamic and graphic metabolic pathways. They contain no comprehensive information about metabolic pathways, for example regulation properties of the enzymes that are involved. Also, problems arise if certain reactions are still unknown or new reactions are recently discovered. For this reason, integrative approaches and new methods of pathway drawing/visualization are demanded.

19.3 Database Systems and Integration

19.3.1 Database Systems

Computational analysis of metabolic pathways on the basis of information about genes, enzymes and metabolites requires access to suitable databases. We exploit a number of major biological databases ranging from genomics to metabolism. To name a few, GenBank [5] is a database that contains an annotated collection of all publicly available DNA sequences. Internet access is provided through several interfaces

directly from the American National Center for Biotechnology Information (NCBI) website. Each sequence is linked to other sequences that are similarly based on sequence alignments. Swiss-Prot [8] is a curated protein database that provides a protein retrieval interface that can be searched by AC, ID, description, gene name, organism, etc. Several mirror sites of Swiss-Prot are available in Europe, America, and Asia. BRENDA [10] systematically collects enzyme data. It is essential both for interpretation of kinetic aspects of enzymatic reactions and for retrieval of enzymes by use of various query terms.

Metabolic pathway databases such as KEGG [11], WIT/EMP [12], and EcoCyc/MetaCyc [26] have been developed to present diagrams depicting metabolic pathways. KEGG is composed of three interconnected sections – genes, molecules, and pathways. It represents data about interacting molecules or genes by using the simplest form of representation – binary relationships that correspond to pairwise interactions. It provides both an online static map of metabolic pathways and the ability to focus on metabolic reactions in specific organisms. WIT/EMP includes some 3000 pathway diagrams covering primary and secondary metabolism, membrane transport, signal transduction pathways, intracellular traffic, translation, and transcription. Initially, EcoCyc/MetaCyc described only metabolic pathways. Now it is extended toward an integrative information system that represents the genes (sequences, function), enzymes (amino acids, function, and structure), and metabolic pathways of *E. coli* [27].

Figure 19.1 shows the databases that enable integrative retrieval of information about metabolic pathways.

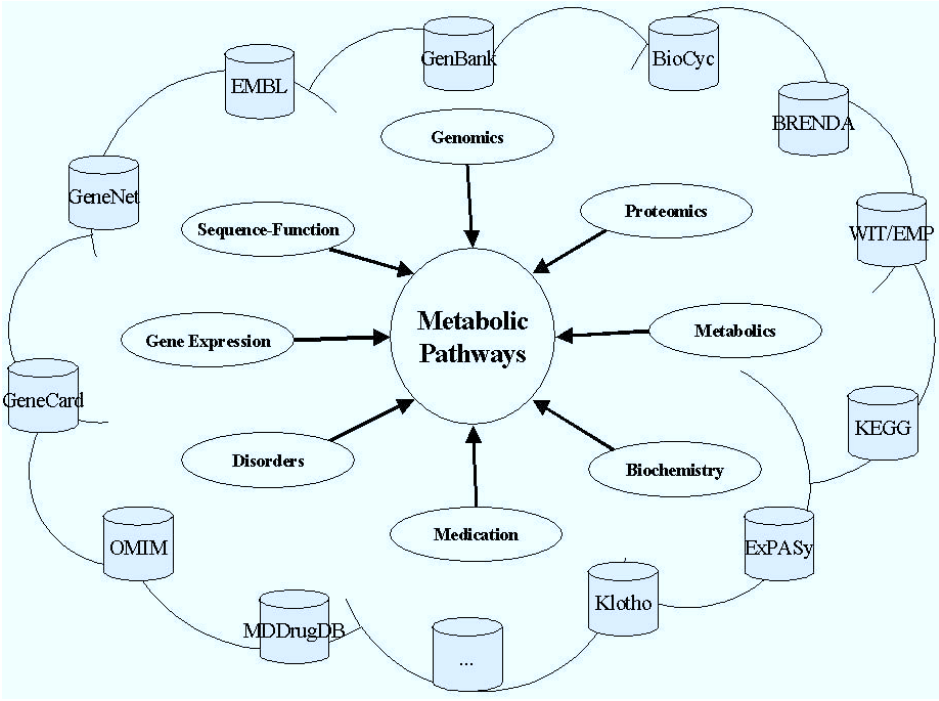


Fig. 19.1 A diagram of biological information sources for metabolic pathway analysis.

19.3.2 Database Integration

The presence of numerous informational and programming resources on gene networks, metabolic processes, gene expression regulation, etc., described above, raises an acute problem of data integration and suitable access. The idea of data integration in molecular biology is not new – several previous and underlying projects focus on the challenging problem of interoperability among biological databases. Karp first addressed biological database integration in early nineties [28]. At the same time the requirements for these integration approaches were formulated [29]. Many integration approaches for molecular biological data sources are currently available. These systems are based on different data-integration techniques, e.g. federated database systems (ISYS [30] and DiscoveryLink [31]), multi database systems (TAMBIS [32]), and data warehouses (SRS [17] and Entrez [33]). Advantages and disadvantages of different approaches are analyzed by comparing the following five properties:

- degree of integration,
- materialization,
- query languages,
- application programming interface (API) standards, and
- various output formats.

ISYS stands for Integrated SYStem and can be characterized as a component-based implementation. The main goal of ISYS is to provide a dynamic and flexible platform for integration of molecular biological data sources. This system is developed as a Java application. Although the system must be installed on a local computer, it has many advantages. Different platforms, for example MS Windows or Solaris, are supported. The locally installed system accesses the distributed data sources on the Internet. One main feature is the global view on to the integrated data sources with the help of a global scheme. Materialization of the integrated sources is not required. ISYS provides a JDBC (Java database connectivity) driver. This feature implements SQL (structured query language) as query language.

The DiscoveryLink system was developed by IBM. It is also based on federated database techniques. A federated system requires the development of a global scheme, thus the extent of integration must be rated as tight. DiscoveryLink accesses its original data sources through views. Read-only SQL is supported as query language. A JDBC and an ODBC (open database connectivity) driver are also provided, and different output formats can also be generated.

TAMBIS integration system is based on multi-database techniques. It can be used through a Java applet. Because of the use of a multi-database query language, it is not necessary to built an integrated global scheme. The degree of integration can therefore be described as loose. As a query language in TAMBIS, a kind of CPL (collection query language) [34] is implemented. CPL is hardwired into the system architecture. This is why it is not so easy to use this query language from outside of the system. Other disadvantages of TAMBIS are the absence of an API, or other public interfaces. The number of input formats, which is limited to one – generated by the Java applet – also proves disadvantageous.

SRS is based on local copies of each integrated data source with a special format that is described in the Icarus language specification. Icarus can help represent the structure of the integrated data source. Through use of these local copies SRS is completely materialized. But during this transfer into the new format no scheme integration is realized. Therefore, the degree of integration

can be characterized as loose. SRS runs on a Web-server and is accessible via any Web-browser. An HTML interface for data queries is provided and the system can be queried by constructing special URLs. But no query languages, for example SQL or OQL (object query language), are supported. SRS also offers a C-API. Different output formats are possible (HTML or ACSII text). One problem with result presentation in SRS is the need to parse the outputs for a further computer-based processing. The absence of any scheme integration is also a disadvantage of the SRS system.

Similar to SRS is the Entrez system. This system integrates only data sources of NCBI. No materialization of the integrated sources is realized. Entrez uses views onto the original sources. Consequently, scheme integration could not be established. Therefore, the degree of integration can be classified as loose. Statements used with SRS to query the system are completely transferable to Entrez. There are no standard query languages, no standardized API, or other interface standards like JDBC. HTML is the only interface provided. Another Entrez feature is the manual construction of special URLs. Various output formats prove to be useful. These include HTML or ASCII text, and XML and ASN.1 files. The greatest disadvantage of the Entrez approach is the restricted number of integrated data sources (only NCBI internal data sources) and the lack of support of query languages. In contrast, the various output formats, primarily XML, are advantageous for the use of this system.

19.3.3 Model-driven Reconstruction of Molecular Networks

Molecular databases provide collected and structured information about molecular objects. Our approach (Fig. 19.2) aims at the integrative and model-driven reconstruction of molecular networks, whereby information from databases is used mostly automatically and interactively to support the process of systems modeling. Data integration is used to overcome the problems of distribution and heterogeneity. In general, information about objects of our molecular model is spread over several databases. Object fusion is used to merge all information about every object into unique objects. Access to different databases and database management systems (DBMS) thus involves structural, syntactic, and semantic problems which must be solved in order to provide homogeneous access.

Biological knowledge is needed in the second step to model networks conceptually. The modeler identifies recent types of object and process appearing in the biological model to create a conceptual model of the system under study. Conceptual models only contain design principles without con-

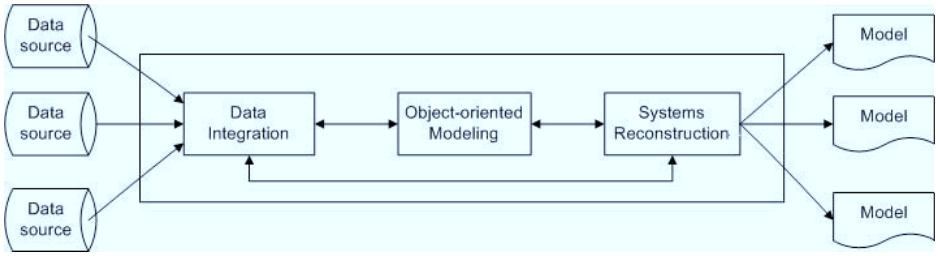


Fig. 19.2 Model-driven reconstruction of molecular networks.

crete objects or processes. All information available from databases will be integrated into the conceptual model and relatively complex networks will be created automatically. Finally, models are extracted and studied in detail afterwards. Extraction applies methods to filter specific information for the modelling of systems, e.g. the flow of material and parameters.

19.3.3.1

Modeling Data Integration

In practice, integrative modeling requires local availability of data. Thus, data integration applied to molecular databases results in *ad-hoc* integrated databases, which are databases merging data collected from several databases called data sources. Mediator-based integration has been implemented and applied in the past [35] to overcome the heterogeneity of data sources. In order to integrate and collect molecular objects into integrated databases, we implement a technique similar to object fusion [36].

Each model contains an integrated database storing information about all elements of the model. Data sources and integrated databases are both represented by database tables. For each type of element in the model a corresponding table in the integrated database has to be defined.

A table is a relation $r(R) \subseteq A_1 \times A_2 \times ... \times A_n$, where the factors A_i define the type of each column. To integrate tuples from data sources, the modeling of mappings between corresponding tables at both sites is required. Each mapping defines the way how to reconstruct tables of the integrated database from tables of a given data source. Let T be a set of tables in the model, D a set of datasources and Q a set of database queries. A mapping $m: T \times D \rightarrow Q$ is valid, if for each element $((t,d),q) \in M$ the query $q(d)$ delivers a relation with the schema R of the corresponding table. The number of mappings that has to be modeled depends on the overlap of information in the data sources. The modeling of separate mappings for each source enables us to flexibly add and remove datasources.

After the specification of mappings we use them as rules to automatically construct the relations of the model. In the first step, the mappings are translated into a query plan. A plan is an ordered sequence of queries, which are then sent to a mediator or directly to the related data sources. Fig. 19.3 illustrates the process of fusing the result sets. Because the schema of the result set equals the schema of the corresponding table in the model, we are able to fuse them using the union operation. Using this operation, identical tuples from different data sources will be removed automatically.

Depending on the content, merging all information about a type of object using one relation will consume a lot of memory. To avoid this, each relation only exists virtually. As explained in the next section, we use an object-oriented model to store the fused information as objects.

To give an example, we briefly explain the definition of mappings for enzymes as drawn in Fig. 19.4. A table "Enzyme" displayed in the left part of the mapping will collect different information about enzymes. On the right side we see tables of a single data source containing related information. What we have to do is to define rules connecting each property of "Enzyme" with corresponding elements of the source. In short, such combinations of different tables are possible, if at least all tables of a mapping can be joined transitively together.

Actually, connected data sources and the biological background change over time. Furthermore, the specification of mappings reflecting the user's point of view cannot be computed automatically. Modeling data integration is required and it consists in:

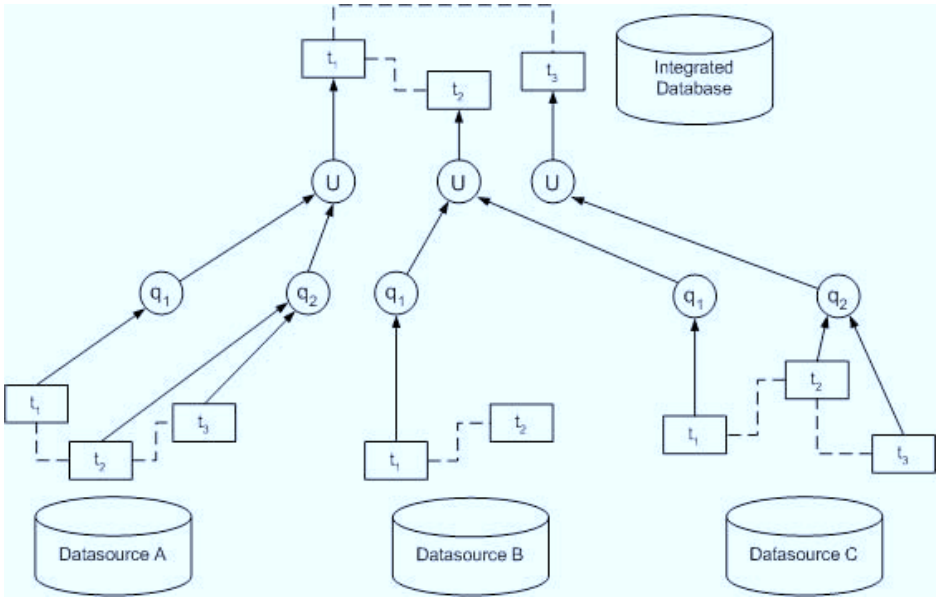


Fig. 19.3 Operations of data integration with object fusion.

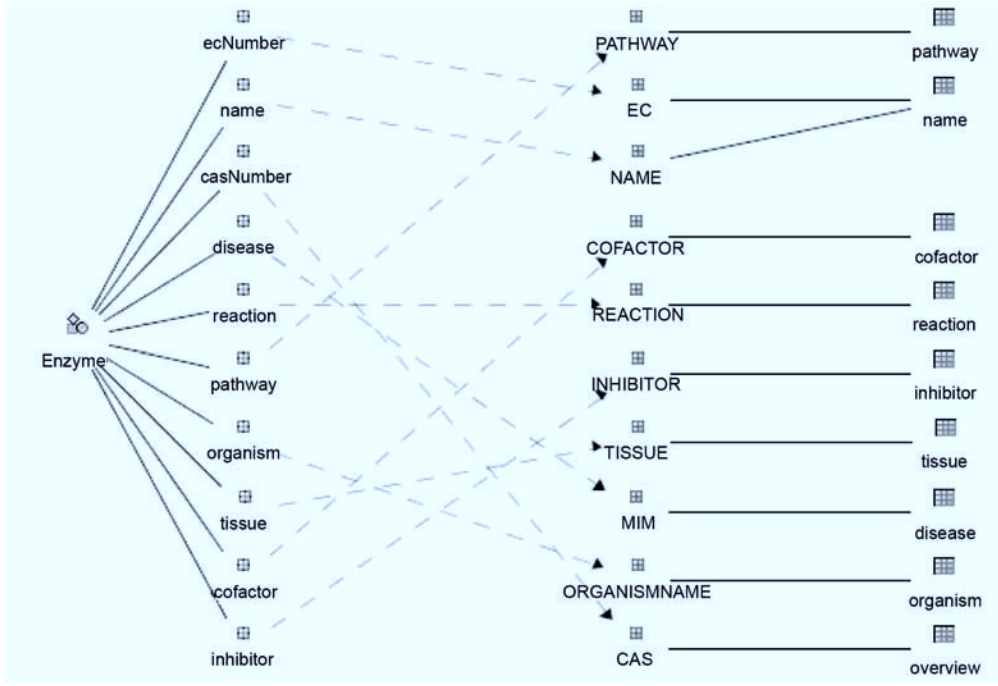


Fig. 19.4 Mapping example for enzymes.

1. selecting and preparing adequate data sources;
2. designing the integrated database depending from the applications background; and
3. defining mappings between integrated databases and data sources.

At the *top-down* approach the structure of integrated databases is already given. The task then is to define mappings to the data sources. The second approach is called *bottom-up*, where the user initially has no idea about how to design the integrated database and he simply copies tables from data sources into identical tables in integrated databases. Both approaches can be combined.

19.3.3.2

Object-oriented Modeling

Obviously, the integration of data is a useful method addressing problems appearing in distributed database environments. Actually, a homogeneous access to recent data sources is provided and new databases with integrated information can be delivered. Respecting the structure of biochemical systems we can see that their complex structure is not discussed with the relational structure of these databases. Evidently, an adequate organization of integrated information is required, which provides a more intuitive exploration of biochemical networks.

The object-oriented approach is a powerful and established method for the organization of complex structured information in general. Most of the object-oriented tools in bioinformatics only provide previously defined and statically implemented data structures. With our system iUDB (individually integrated user databases, http://tunicata.techfak.uni-bielefeld.de/proton) [37] the method is dynamically applicable to the modelling of complex biochemical networks. Moreover, it enables the user to automatically construct models using data integration, as it has been explained in the previous section.

To be able to work with the method, scientists have to be familiar with typical object-oriented concepts. In practice here are a lot of systems using different implementations of the approach, but generally they all include a set of the same basic concepts. Common standards for the conceptual modeling are, e.g. the Unified Modeling Language (UML) and the Interface Definition Language (IDL), which both have been developed by the Object Management Group (OMG). Before we can create molecular objects we have to conceptually model them. Every object has be be associated to a type, e.g. enzymes, genes or metabolites. In the model discussed here all objects of the same type are members of the same class of objects. An object-oriented schema than contains a set of classes. Every class holds a set of attributes, which determine the type of information that should be stored within all objects of the class. Attributes can have a standard datatype (integer, double, ...) or a type of objects that has been defined in the same schema (enzyme, gene, ...). Object-based attributes represent relationships between objects. The structure of all relationships in a schema can be interpreted as a direct graph. Only if there are cycles in the graph of the schema networks of objects can be stored in the model.

For the automated implementation of object-oriented schemata several tools for CASE (Computer Aided Software Engineering) are available. Based on the conceptual model most of the tools produce source code for different object-oriented languages. Our tool iUDB additionally implements an Internet service for each model on the fly, which allows together with a graphical user interface the interactive modeling of objects of the modeled types.

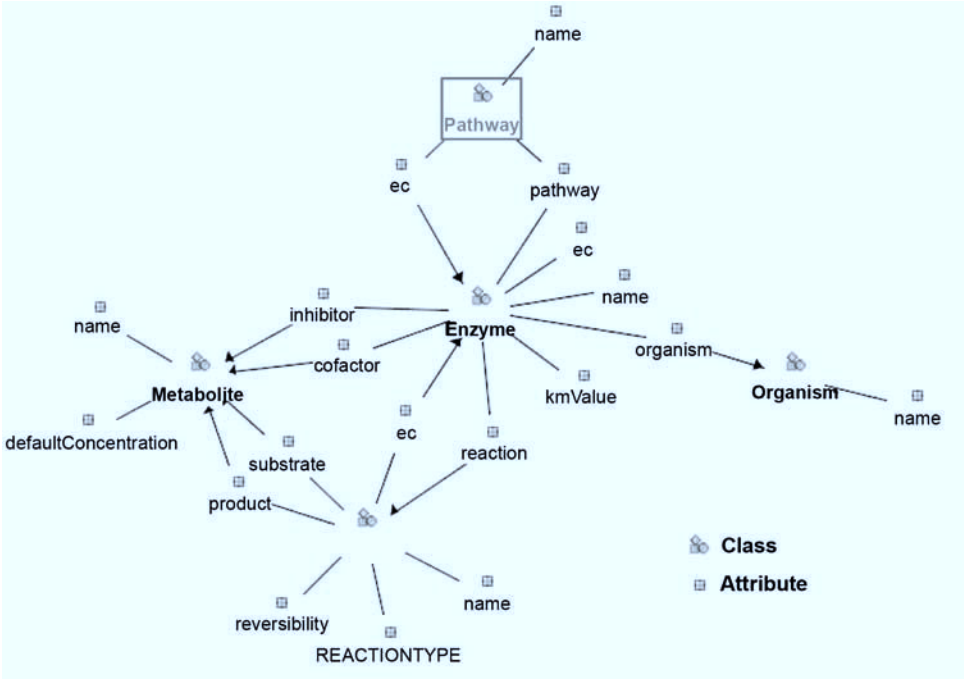


Fig. 19.5 Object-oriented schema of biochemical networks.

To give an example, Fig. 19.5 shows an object-oriented schema displayed as a direct graph. The model contains the five classes "Pathway", "Enzyme", "Organism", "Reaction" and "Metabolite". Each class holds a set of attributes, visualized as additional nodes. References are visualized as directed edges pointing to an object type (e.g. "ec" of "Pathway" points to "Enzyme"). The example contains different types of relationships. While pathways are constructed of enzymes (aggregation), reactions are catalyzed by enzymes (association).

After the conceptual modeling of the biochemical systems static structure we have to enter objects as "facts" into the model. Faced with hundreds of growing molecular databases a systematic as well as automated mechanism is necessary for this task. As mentioned in the previous section, objects can be reconstructed from relations using object fusion [36]. Therefore we have to add specific constraints to our model. For each class the uniqueness of the objects must be defined. In our approach we are using object keys, whereby a key is a subset of the attributes defined in a class. Let C be the set of objects of a given class and the mapping $key\colon C \rightarrow \mathbb{N}$ the keys of all objects in C. The following implication ensures the integrity of the objects:

$$\forall\, a,b\,(a,b \in\ C \wedge key(a) = key(b) \Rightarrow a = b)$$

With key constraints given for each class we can proceed with fusing the information extracted and integrated in the last section into the model. The tuples retrieved from data sources during the integration procedure are processes sequentially. For each appearing key a separate object is created. Does an object for a given key already exist

in the model, the related object will be loaded and the information stored in the tuple added to the attributes. Because a high rank of the requested relations can induce performance problems, this bottleneck can be lowered by a pairwise requesting of the attributes in the result. If a class of objects consist of n attributes where one attribute of them is the objects key, a query has to be split into a total number of n subqueries.

Using a database management system for the persistent storage of the model, we are able to query the model using query languages, e.g. OQL. Result sets of the queries contain all objects of the model, which satisfy the constraints given in the query. To give an example, the query

```
SELECT c
FROM c IN PathwayExtent
WHERE c.ec.reaction.product.
name = "Putrescine"
```

will search the database based on the schema in Fig. 19.5 and compute all objects of the class "Pathway" producing putrescine.

In Fig. 19.6, a part of the environment of the object "Alkaloid biosynthesis II" has been drawn as a direct graph. Note that the objects are associated to different classes. The network visualizes the structure of the integrated model. The object "Alkaloid biosynthesis" of the class "Pathway" points to nine objects of the class "Enzyme". At object "4.3.1.5", which is the enzyme L-tyrosine ammonia-lyase we can see that this object appears in four different pathways, which are Alkaloid biosynthesis, Nitrogen metabolism, Tyrosine metabolism and Phenylalanine metabolism. The enzyme Ornithine decarboxylase labeled with "4.1.1.17" is catalyzing the reaction from L-Ornithine to Putrescine.

19.3.3.3

Systems Reconstruction

Data integration automatically integrated distributed information as objects into the previously defined models. We can now explore integrated databases almost manually starting at known objects. Databases pro-

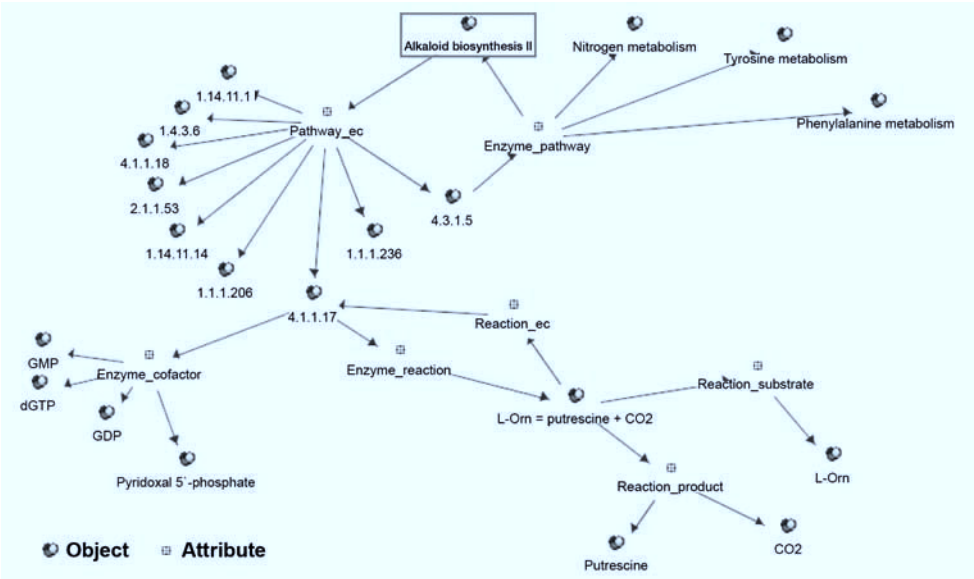


Fig. 19.6 Environment of alkaloid biosynthesis pathway.

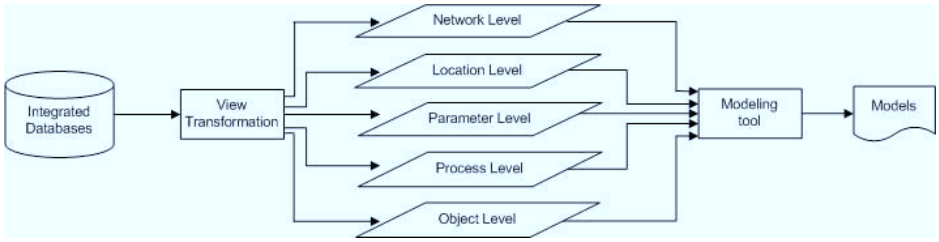


Fig. 19.7 Extraction of models from integrated databases.

vide several mechanisms to retrieve information about the stored objects automatically. Query languages, e.g. OQL enable the user to flexibly query integrated database to select objects of interest.

Approaching molecular networks, it is not sufficient to discuss objects and their relationships only. Moreover, integrated databases should be interpreted at the point of biochemical systems at different levels. Reactions can have different type, they are highly interconnected, placed at different loci and they also transform substances of different classes placed at different locations in the cell. As will become apparent in the next section, different models are applied to the design of biological systems, into which we have to transform integrated data. Actually, we apply view concepts implemented within database systems and methods from graph theory to retrieve, e.g., metabolic, gene regulatory, or signaling networks.

Obviously, this is a task for KDD (knowledge discovery in database), in which we integrate and analyze databases for characteristic patterns. For extraction of biochemical models we get a workflow similar to KDD, as has been drawn in Fig. 19.7. Integrated databases will be translated into different views using view-based transformation. Views contain information at different levels, which are needed for systems modeling. Classes of molecular object hold all objects of the systems. Classes of processes do the same for processes. At parameter level the kinetic parameters of processes are defined. Networks detect and cluster interconnected objects and processes as, e.g., pathways. Finally, the location of objects and processes is essential to reflect the hierarchical structure of biological systems.

19.4 Different Models and Aspects

The availability of rapidly increasing volumes of molecular biology data enhances our capability to study cell behavior. To understand the logic of cells we must be able to analyze metabolic processes in qualitative and quantitative terms. Modeling and simulation are important methods. Mathematical models can be classified as analytical or discrete. Analytical models perform the processes of element functioning as some functional relationships (algebraic, integral-differential, finite-differential, etc.) or logical conditions. They can be studied by qualitative, analytical, or numerical methods. Discrete models are based on state transition diagrams.

The analytical approach to metabolic simulation, for example, typically requires the determination of steady-state rate equations for constituent reactions, followed by numerical integration of a set of differential

equations describing fluxes in the metabolism. The feasibility of the analytical approach is, however, limited by the extent to which the metabolic processes of interest have been characterized. For most metabolic pathways either we are unaware of all the steps involved or we lack rate constants for each step. This lack of information precludes the use of the mathematical approach to describe the process. Even when reaction rates are known, differential equations incur great computational costs. Analytical representations, such as differential equations, lack the robustness required to handle partial and uncertain knowledge. In addition, because analytical simulations model relatively similar structures over relatively similar temporal intervals, interleave simulations are highly constrained.

The discrete-event approach can provide declarative representations for both the structures in the domain and the processes that act on these structures. Most importantly, discrete-event simulations provide natural support for qualitative representation and reasoning techniques, which offer explicit treatment of causality. The graphical model of Kohn and Letzkus [18], which enables discussion of metabolic regulation processes, is representative of the class of graph theoretical approaches. They expanded the graph theory by a specific function that enables modeling of dynamic processes. In this case, the approach of Petri nets is a new method. Reddy et al. [20, 38] presented the first application of Petri nets in molecular biology. In recent years, more applications of Petri net methodology to metabolic pathway modeling and simulation appeared. The following section explores various aspects of modeling biochemical pathways using Petri nets. The interest and potential of Petri nets to help understanding of complex biological processes is reflected.

19.4.1 Petri Net Model

19.4.1.1 Basics

Since the nineteen-sixties, when the Petri net was first introduced and formally defined by Petri [39], the Petri net and its concepts have been extended and developed, and both the theory and the applications of this model have flourished. In contrast to naive graph, the Petri net is a graph oriented design, specification, simulation and verification language. It offers a formal way of representing the structure of a discrete and/or event system, simulating its behavior, and drawing certain types of general conclusions about the properties of the system. Because of their good properties in theoretical analysis, practical modeling, and graphical visualization of concurrent systems, Petri nets, especially high-level Petri nets, are widely used in work-flows, flexible manufacturing, operations research, railway networks, defense systems, telecommunications, the Internet, commerce and trading, and biological systems.

Petri nets are conceptually simple: they consist of places, transitions, and arcs. Each place has a non-negative number of tokens. A transition is enabled if the number of tokens exceeds the weights of the arcs connecting the places. A definition of the ordinary Petri net is given below [40, 41]:

Definition 1 *An ordinary Petri net is a 3-tuple,* $PN = (P,T,F)$ *with:*

$P = \{p_1,p_2,...,p_m\}$ *is a non-empty, finite set of places, drawn as circles;*
$T = \{t_1,t_2,...,t_n\}$ *is a non-empty, finite set of transitions, drawn as bars;*
$P \cap T = \emptyset$ *and* $P \cup T \neq \emptyset$;
$F \subseteq (P \times T) \cup (T \times P)$ *is a non-empty, finite set of arcs, connecting places to transitions or transitions to places but never two places or two transitions.*

The ordinary Petri net contains structural elements only. To define dynamic Petri nets and their firing rules, we need some terminology to identify special sets of places and transitions and the concept of markings.

Definition 2 *Pre- and Post-Sets*

The pre-set ${}^{\circ}t_i$ *of a transition* $t_i \in T$ *contains all places that are connected to* t_i *via a directed arc from the place to the transition:* ${}^{\circ}t_i = \{p \in P : (p, t_i) \in F\}$. *The elements of* ${}^{\circ}t_i$ *are often called input places.*

The post-set t_i° *of a transition* $t_i \in T$ *contains all places that are connected to* t_i *via a directed arc from the transition to the place:* $t_i^{\circ} = \{p \in P : (t_i, p) \in F\}$. *The elements of* t_i° *are often called output places.*
The pre-set p_i *and post-set* p_i *of a place* $p_i \in P$ *are defined in the same way:*
${}^{\circ}p_i = \{t \in T : (t, p_i) \in F\}$
$p_i^{\circ} = \{t \in T : (p_i, t) \in F\}$

Definition 3 *Marking*

A marking of a Petri net is a mapping $M : P \to N$, *that assigns a finite non-negative integer number of tokens to each place of the ordinary Petri net.* $M_0 : P \to \mathbb{N}$ *is the initial marking.*

State changes are carried out by firing enabled transitions. In an ordinary Petri net, a transition is enabled when all its input places have at least one token. When an enabled transition t is fired, a token is removed from each input place of t and a token is added to each output place of t. This gives a new state.

19.4.1.2

Hybrid Petri Nets

Ordinary Petri net models do not have such functions as quantitative aspects, so there are some extension of Petri nets that can support dynamic change, task migration, super imposition of various levels of activities and the notion of mode of operations. Various extensions of PN, such as (stochastic) timed PNs [42, 43], colored PNs [44], predicate/transition nets [45] and hybrid PN [41], enable qualitative and/or quantitative analysis of resource utilization, effect of failures, and throughput rate. Hofestädt [46] also presented an extension formalization, a self-modified Petri net, which enables quantitative modeling of regulatory biochemical networks. We exploit the methodology of hybrid Petri net to model gene regulated metabolic networks in the cell, explain the importance of sustaining core research, and identify promising opportunities for future research.

Herewith a brief description of hybrid Petri nets is presented as the following context.

Definition 4 *A hybrid Petri net is a sextuple* $Q = (P, T, Pre, Post, \mathrm{h}, M)$ *such that:*

$P = \{p_1, p_2, \ldots, p_n\}$ *is a not empty, finite set of places;*
$T = \{t_1, t_2, \ldots, t_m\}$ *is a not empty, finite set of transitions;*
$P \cap T = \varnothing$, *i.e. the sets P and T are disjointed;*
$h : P \cup T \to \{D, C\}$, *called "hybrid function", indicates for every node whether it is a discrete node (sets* P^D *and* T^D*) or a continuous node (sets* P^C *and* T^C*);*
$Pre : P \times T \to \mathbb{R}^+$ *or N, is the input incidence mapping (*$\mathbb{R}^+$ *denotes the set of positive real numbers, including zero, and N denotes the set of natural numbers);*
$Post : P \times T \to \mathbb{R}^+$ *or N is the output incidence mapping;*
$M : P \to {}^+$ *or* $\mathbb{N}$ *is the marking.*

We denote by $M_{(t)} = (m_1^t, m_2^t, \ldots, m_n^t)$ the vector which associates with each place of P its marking at the instant t. $M_0 = M_{(t0)} = (m_1^0, m_2^0, \ldots, m_n^0)$ is the initial marking. At any time the present marking M is the sum of two markings M^r and M^n, where M^r is the

reserved marking and M^n is the non-reserved marking. If $h(P_i) = D$ or C then $m_i(t) = m_i^r(t) + m_i^n(t)$. When a variable dT_j (called the delay time of T_j) is assigned to each discrete transition $T_j(h(T_j) = D)$ and T_j is fired at time $t + \mathrm{d}T_j$, then:

$\forall P_i \in {}^\circ T_j$ (${}^\circ T_j$ denotes the set of input places of transition T_j), $m_i(t) \geq Pre(P_i, T_j)$,
$m_i(t + dT_j) = m_i(t) - Pre(P_i, T_j)$
$\forall P_i \in T_j^\circ$ (T_j° denotes the set of output places of transition T_j),
$m_i(t + dT_j) = m_i(t) + Post(P_i, T_j)$

When a variable vT_j (called the speed of T_j) is assigned to each continuous transition $T_j(h(T_j) = C)$ and T_j is fired at time t during a delay dt, then:

$\forall P_i \in T_j$, $m_i^n(t) \geq Pre(P_i, T_j)$,
$m_i(t + d_t) = m_i(t) - v_j(t) \times Pre(P_i, T_j) \times d_t$
$\forall P_i \in T_j^\circ$,
$m_i(t + d_t) = m_i(t) + v_j(t) \times Post(P_i, T_j) \times d_t$

where $v_j(t)$ is the instantaneous firing flow of T_j at time t.

Given the concept that an inhibitor arc of weight r from a place P_i to a transition T_j allows the firing of T_j only if the marking of P_i is less than r, we can extend the above-defined hybrid Petri net. If the inhibitor arc has its origin at a discrete place and has a weight $r = 1$, the corresponding transition can be fired only if $m_i > 1$, actually, only if $m_i = 0$, because m_i is an integer. If the origin place is continuous, then a conventional value 0^+ is introduced to represent a weight infinitely small but not nil. The new definition of an extended hybrid Petri net is similar to the definition of a hybrid Petri net (Definition 4), except that:

One can have, in addition, inhibitor arcs;
The weight of an arc (inhibitor or ordinary) whose origin is a continuous place takes its value in $\mathbb{R}^+ \cup \{0^+\}$ *instead of* $\mathbb{R}^+$;
The marking of a continuous place takes its value in $\mathbb{R}^+ \cup \{0^+\}$ *instead of* $\mathbb{R}^+$.
Test arcs have no token flow between input places and transitions, i.e. $m_i(t + d_t) = m_i(t)$.

So far, hybrid Petri nets are unified representations of continuous variables represented as continuous place token counts (real numbers) and of discrete ones represented as discrete place token counts (integers). The defined hybrid Petri net turns out to be a flexible modeling process that makes sense to model biological processes, by allowing places using actual concentrations and transitions using functions.

Figure 19.8 shows the basis elements of the hybrid Petri net VON++—discrete place, continuous transition, continuous place, and discrete transition connected with test arc, normal arc, and inhibitor arc, respectively. The tool can be downloaded via its web site at http://www.systemtechnik.tu-ilmenau.de/ ~drath/visual.htm.

The discrete transition is the active element of discrete event Petri nets. A transition can fire if all places connected with input arcs contain equal or more tokens than the input arcs specify. It can be assigned with a delay time. The continuous transition differs from the traditional discrete transition; its activity is not comparable with the abrupt firing of discrete transition.

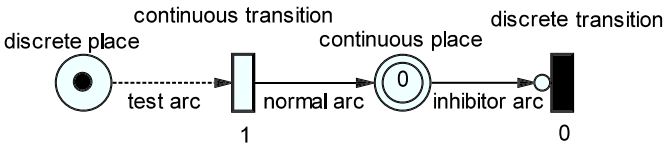


Fig. 19.8 Elements of a hybrid Petri net.

The firing speed assigned to a continuous transition describes its firing behavior and can be constant or a function, i.e. transport of tokens according to $v(t)$, in Fig. 19.8, $v(t) = 1$. However, the rate of bioprocesses is not defined within a Petri net, it should be specified separately. In most chemical systems the rate of processes (transitions) can be defined by the mass action law. The rate of change in the number of tokens (or concentration) is proportional to the number of tokens (or concentration) in all starting places. In biochemistry, the most commonly used expression that relates the rate of enzyme-catalyzed formation of a product to substrate concentration is the Michaelis–Menten equation, which is given as $v = v_{max}S/(K_m + S)$. Such enzyme reactions are characterized by two properties, V_{max} and K_m, and biochemists are interested in determining these experimentally. Fortunately, some public biological databases such as BRENDA are available to provide enzyme-reaction data collected from research literature.

19.4.1.3

Applications

After early application to modeling metabolic pathways [20, 47], Petri nets as new tools and terms for modeling and simulating biological information systems have been investigated more and more. Reddy et al. [48] then presented an example of the combined glycolytic and pentose phosphate pathway of the erythrocyte cell to illustrate the concepts of the methodology. However, the reactions and other biological processes were modeled as discrete events and it was not possible to simulate the kinetic effect. Hofestädt [46] investigated a formalization showing that different classes of conditions can be interpreted as gene, proteins, or enzymes, and cell communication and also presented the formalization of self-modified Petri nets, which enable the quantitative modeling of regulatory biochemical networks. Chen [49] introduced the use of hybrid Petri nets (HPN) for expressing glycolysis metabolic pathways. Using this approach the quantitative modeling of metabolic networks is also possible. Koch et al. [50] extended the Reddy's model by taking into account reversible reactions and time dependencies. Kueffener [51] exploited the knowledge available in current metabolic databases for functional predictions and the interpretation of expression data on the level of complete genomes, and described the compilation of BRENDA, ENZYME, and KEGG into individual Petri nets and unified Petri nets. Goss [52] and Matsuno [53] used Petri nets to model gene regulatory networks by using stochastic Petri nets (SPN) and HPN, respectively. In the DFG workshop "Modeling and Simulation Metabolic Network" 2000, participants also discussed the applications and perspective of Petri nets [54]. Genrich et al. [55] discussed executable Petri net models for the analysis of metabolic pathways. Heiner et al. [56] studied the analysis and simulation of steady states in metabolic pathways with Petri nets. Srivastava et al. [57] also exploited a SPN model to simulate the σ32 stress circuit in *E. coli*. Oliveira et al. [58] developed the mathematical machinery for construction of an algebraic-combinatorial model to construct an oriented matroid representation of biochemical pathways. Peleg [59] combined the best aspects of two models—Workflow/Petri net and a biological concept model. The Petri net model enables verification of formal properties and qualitative simulation of the Workflow model. They tested their model by representing malaria parasites invading host erythrocytes, and composed queries, in five general classes, to discover relationships among processes and structural components. Recently, a spe-

Table 19.1 Summary of Petri net tools used for modeling and simulation of biological systems.

Petri nets type	*Petri net tool*	*Brief description of tool*	*Application*	*Refs.*
High level	Stella	The STELLA software is based on a feedback control framework. The basic self-regulatory, or homeostatic, mechanisms that govern the way living systems operate, are reinforced by the way the software itself operates, it enables users to make their hypotheses explicit using simple iconic building blocks, and then to test these hypotheses by simulation.	Modeling dynamic biological systems, especially ecological systems.	65, 66
Hybrid	VON++	Visual object net++ is an innovative Petri net CAE tool for PC that supports mixed continuous and discrete event Petri nets. Beside the new continuous net elements, the whole well tried concept of the traditional Petri nets is available. The goal of visual object net++ is to study the behavior and characteristics of a class of hybrid Petri nets.	Gene regulatory; metabolic pathways; bioprocesses	49, 53, 67–69
Stochastic	UltraSAN	UltraSAN employs stochastic activity networks (SAN), a variation of Petri nets, to model and analyze the performance and dependability of software, hardware and network system designs. UltraSAN provides analytical solvers and discrete-event simulators.	Protein synthesis from mRNA; plasmid replication; prion propagation	52, 57, 70
Hierarchical	PED	PED supports basically the construction of hierarchical place/transition nets with the specification of different types of places, transitions, and arcs, including their marking.	Pentose phosphate pathway	50
High level	THORNs	THORNs is a general-purpose, graphical, discrete-event simulation tool based on a special class of high-level Petri net called timed hierarchical object-related nets. THORNs enable the specification of individual tokens, they provide delay times and firing durations for transitions, and THORN models can be hierarchically structured with respect to transition refinement and subnet invocation.	Ecological systems	71, 72

Table 19.1 Continued

Petri nets type	*Petri net tool*	*Brief description of tool*	*Application*	*Refs.*
High level	Design/CPN	Design/CPN supports CPN models with complex data types (color sets) and complex data manipulations (arc expressions and guards). The functional programming language Standard ML enables the software package to support hierarchical CP-nets and generate a model from the data extracted from databases.	Glycolysis	51, 55, 73
Functional	GON/Cell Illustrator	Genomic object net is an environment for simulating and representing biological systems. The Commercialized version of GON is named Cell Illustrator (CI).	Biopathways; cell development	64, 74, 75

cial issue on "Petri nets for Metabolic Networks" has appeared in ISB journal [60]. It covers topological analysis of metabolic networks based on Petri net theory [61], qualitative analysis of steady states in metabolic pathways [62], hybrid Petri net modeling and simulation [63], and introduction of the hybrid functional Petri net (HFPN) [64].

Table 19.1 presents a summary of the Petri net tools used to model biological systems. Most publications presented their models based only on a general Petri net tool utilization. These publications are not listed in the table. More Petri net tools can be found at http://www.daimi.au.dk/PetriNets/tools/quick.html.

So far, the intuitively understandable graphical notation and the representation of multiple independent dynamic entities within a system makes Petri nets the model of choice, because they are highly suitable for modeling and simulation of metabolic networks.

19.4.1.4 Petri Net Model Construction

Although many Petri net tools are available, most import and/or export Petri net diagrams in a binary file format which poorly supports the possibility of making diagrams distributed in multiple format files; less mention constructing a net from a text format file. That means it is impossible to extract data from biology databases and construct a Petri net model automatically. Fortunately, several Petri net tools such as PNK (http://www.informatik.hu-berlin.de/top/pnk/), Renew (http://www.renew.de/), and CPN (http://www.daimi.aau.dk/CPnets/) have been equipped with an XML-based file format. The XML-based Petri net interchange format standardization that consists of a Petri net markup language (PNML) and a set of document type definitions (DTD) or XSL schemes has been released and intended to be adopted.

To present a Petri net model automatically from biology databases, we developed an XSLT file to convert the original XML source file from our metabolic pathway data stored in an Oracle system into the desired XML format that can be executed by the Renew XML parser. Figure 19.9 shows the automatic layout of Petri net model with Renew.

Although the above-mentioned Petri net XML standards are available, they have their

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet
href="http://sanfrancisco/xml/db2xml/test.xsl"
type="text/xsl"?>
<database URL="jdbc:oracle:thin:@edradour.cs.uni-
magdeburg.de:1521:orcl">
<table0 QUERY="select * from enzyme where ec =
'3.5.3.1' or ec= '4.3.2.1' or ec='6.3.4.5' or ec='2.1.3.3' or
ec='6.3.4.16'"
>

<record0>
<EC><![CDATA[6.3.4.16]]></EC>
<PRODUCT><![CDATA[ADP]]></PRODUCT>
<SUBSTRATE><![CDATA[NH3]]></SUBSTRATE>
</record0>

<record0>
<EC><![CDATA[6.3.4.16]]></EC>
<PRODUCT><![CDATA[ADP]]></PRODUCT>
<SUBSTRATE><![CDATA[CO2]]></SUBSTRATE>
</record0>

...
</record0>
...
</table0>
</database>
```

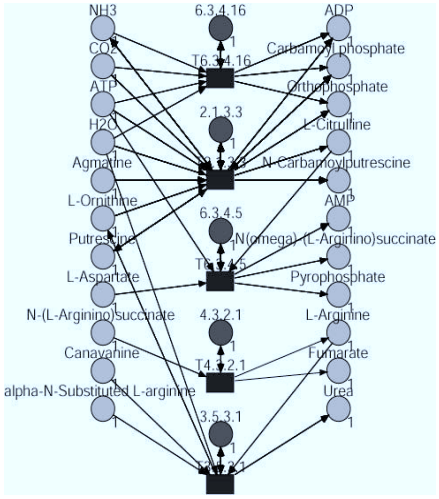


Fig. 19.9 Petri net model layout based on XML. An XSLT file was used to transform the original database XML format into the Renew Petri net XML format.

different definitions and ontologies because of different design destinations. For application of the Petri net methodology to metabolic networks, a new standard should be presented. We proposed a general scheme for biology Petri net markup language (BioPNML) [76], presenting the concepts and terminology of the interchange formats and its syntax, which is based on XML. It should provide a starting point for the development of a standard interchange format for bioinformatics and Petri nets. The purpose of BioPNML is to serve as a common framework for exchanging data about metabolic pathways, and to provide guidance to researchers who are designing databases to store networks and reaction data and modeling them on the basis of Petri nets.

In the iUDB system we also present a means of integrating data and model metabolic networks with Petri nets. After the integrated scheme and applicable rules are designed, integrated data can be transformed into the rule formats by mapping database objects (pathways, reactions, and enzymes) directly into rules. The application of the Petri net rule enables display of these data as a Petri net model.

19.5 Simulation Tools

Many attempts have been made to simulate molecular processes in both cellular and viral systems. Several software packages for quantitative simulation of biochemical metabolic pathways, based on numerical integration of rate equations, have been developed. A list of biological simulators can found at http://www.techfak.uni-bielefeld.de/~mchen/BioSim/BioSim.xml. The most well known metabolic simulation systems are compared in Table 19.2.

Each tool possesses some prominent features which are not present, or are rarely present, in others. After a decade of development, Gepasi is widely applied for both research and education purposes to simu-

Table 19.2 A comparison of metabolic simulators.

Tools	*Gepasi*[a)]	*Jarnac*[b)]	*DBsolve*[c)]	*E-Cell*[d)]	*VON++/GON*[e)]
Stoichiometry matrix presentation	+	+	+	+	–
Core algorithm and method	MCA*	MCA*	MCA*	SRM, MCA*	Petri net
Pathway DB retrievable	–	–	WIT/ EMP	KEGG, EcoCyc	KEGG
Pathways graphic editor	–	++++	++++	–	+++++
Kinetic types	++++	+++	+++	++	++++
Virtual cell model	–	–	–	+	+
Simulation graphic display	++++	+++	++	++	+++
Mathematical model accessible and modifiable	+	+	+	+	+
Data XML import/export	SBML**	SBML**	SBML**	SBML**	Biopathway XML
User interface	++++	+++	++++	+++	++++
Programming language	C++	Delphi 5	C++	C++	Delphi/Java

a) Gepasi [http://www.gepasi.org/]
b) Jarnac [http://members.lycos.co.uk/sauro/biotech.htm]
c) Dbsolve [http://homepage.ntlworld.com/igor.goryanin/]
d) E-Cell [http://www.e-cell.org/]
e) VON++ is further developed to GON. GON has been sold, the commercial name is Cell Illustrator [http://www.gene-networks.com/ci/]
*) MCA (metabolic control analysis) is a phenomenological quantitative sensitivity analysis of fluxes and metabolite concentrations [77]
**) SBML (systems biology markup language) [http://www.cds.caltech.edu/erato/] is a description language for simulations in systems biology. It is oriented toward representing biochemical networks that are common in research on a number of topics, including cell-signaling pathways, metabolic pathways, biochemical reactions, gene regulation, and many others. SBML is the product of close collaboration between the teams developing BioSpice [http://biospice.lbl.gov/], Gepasi, DBSolve, E-Cell, Jarnac, StochSim [http://www.zoo. cam.ac.uk/comp-cell/StochSim.html] and Virtual Cell [http://www.nrcam.uchc.edu/]

late the dynamics and steady state of biochemical systems, because of its powerful simulation engine and user-friendly interface. Jarnac, a replacement of SCAMP, has a good pathway graphic editor, called Jdesigner, which enable users to draw interactively a biochemical network and export the network in XML format. Dbsolve is good for model analysis and optimization. It uses numerical procedures for integration of ODE (ordinary differential equations) or NAE (nonlinear algebraic equations) to describe the dynamics of these models and offers an explicit solver, an implicit solver, and a bifurcation analyzer. The primary focus of E-Cell is to develop a framework for constructing simulatable cell models based on gene sets derived from completed genomes. In contrast with other computer models that are being developed to reproduce individual cellular processes in detail, E-Cell is designed to paint a broad-brush picture of the cell as a whole. Another program, DynaFit [http:// www.biokin.com/dynafit/], is also useful for analysis of complex reaction mechanisms, for which traditional (algebraic) kinetic equations cannot be derived.

In predicting cell behavior, the simulation of a single pathway or a few intercon-

nected pathways can be useful when the pathways being studied are relatively isolated from other biochemical processes. In reality, however, even the simplest and most well studied pathways, for example glycolysis, can exhibit complex behavior because of connectivity. In fact, the more interconnections between different parts of a system, the harder it becomes to predict how the system will react. Moreover, simulations of metabolic pathways alone cannot account for the longer time-scale effects of processes such as gene regulation, cell division cycle, and signal transduction. When systems reach a certain size they will become unmanageable and non-understandable without decomposition into modules (hierarchical models) or presentation of graphs. Moreover, modeling and simulation of large-scale metabolic networks requires intensive data integration. In this sense, tools mentioned above seem to be weak.

In comparison, Petri nets capture the basic aspects of concurrent systems of metabolism both conceptually and mathematically. The major advantages of Petri nets are graphical modeling representation and sound mathematical background; these make it possible to analyze and validate the qualitative and quantitative behavior of a Petri net system, enable clear description of concurrency and long experience in both specification and analysis of parallel systems, and enable description of a Petri net model on different levels of abstraction (hierarchical models). In addition, the development of computer technology enables Petri net tools to have more friendly interfaces and the possibility of standard data import/export support.

The requirements of a biology-specific Petri net tool are discussed in the following paragraphs.

19.5.1 Metabolic Data Integration

At present a Petri net tool that can process data integration is still missing. Coupled with the data integration techniques discussed above it is possible to construct an integrative model out of various data sources. Another possibility is to use standard data format. XML is already a standard for storing and transferring data. Many biological databases such as Transpath and MPW have already considered using the technique or are doing so now. The intended software should support XML import and export, link to internal and external related databases of genes, enzymes, reactions, kinetics, organisms, compartments, and initial values of (ODE-NAE) systems. BioPNML is intended to be a standard.

With the complete annotated sequence of a genome we can, moreover, generate drafts of the organism's metabolic networks. For the next generation of metabolic models, which will probably be integrated with genome databases, it should be possible to include fields containing information on the evidence for particular values, e.g. evidence codes in gene ontology [http://www.geneontology.org/doc/GO.evidence.html].

19.5.2 Metabolic Pathway Layout

Structural knowledge of a physical system is the foundation of a simulation. Petri net representation is a type of object-oriented representation. It can present a structure of metabolic pathways naturally. The metabolic pathway editor tends to be based on Petri net methodology. In addition, Petri nets can be executed and the dynamic behavior observed graphically. When the model becomes very large, however, e.g. for a whole

cell model, computation time and space increase exponentially.

19.5.3 Dynamics Representation

Dynamics representation is critical to the success of a simulation. To build a quantitative model, kinetic properties of enzyme-catalyzed reactions involved in pathways should be outlined. Traditionally, ODE and NAE models, for example Michaelis–Menten kinetic models, are used to simulate metabolic reactions. Different types of enzyme kinetics (Michaelis–Menten equation, reversible mass action kinetics equation, allosteric inhibition equation, etc.) and initial parameter values can, in part, be obtained from databases and the literature. Otherwise, a user-defined model should be built as E-Cell and Gepasi do. Nevertheless, a Petri net-based simulation system still can deal with bioprocesses as discrete-event or semi-quantitative models when the required kinetic data are unavailable.

19.5.4 Hierarchical Concept

Hierarchical biochemical systems are biochemical systems that consist of multiple modules that are not connected by a common mass flux, but communicate only through regulatory interactions. The models of a virtual cell should contain metabolic pathways and the levels of transcription and translation, and so on. Reactions in different compartments require hierarchical model representation. E-Cell models perform this technique very well. With mature mathematical support, Petri nets also can handle it and at the same time it enables structural reduction of the Petri net, because otherwise the state space of Petri net structure will be very large in graphs.

19.5.5 Prediction Capability

Because metabolism is far less well understood than a manufactured system, biological simulations often yield highly uncertain results. For this reason a bifurcation analyzer or a fuzzy analyzer should be included in the software, as Dbsolve does. Also, because concentration values of metabolites within a cell fluctuate within a normal range; prearrangement of such data in the software is necessary.

The pathway simulator is also able to predict pathways and find alternative pathways from several known biochemical reactions. Each reaction is thermodynamically feasible, i.e. ΔG is equal to or less than zero. It can calculate thermodynamic characteristics. Otherwise, the requirements for coupling of reactions (combined with ATP utilization) should be checked and any two coupled reactions must proceed via a common intermediate. The reversibility of one reaction is determined and displayed in case abnormal situations occur, even though metabolite flow tends to be unidirectional.

19.5.6 Parallel Treatment and Development

After many years of development, quantitative modeling can now be handled by Petri nets. They have a mature mathematical algorithm and can solve NAE and ODE and stoichiometric matrices. But biochemical systems are also rich in time scales and thus require sophisticated methods for the numerical solution of the differential equations that describe them. Especially in virtual cell modeling and simulation, parallel treatment of these equations during simulation is of importance, yet difficult to achieve. When, moreover, we consider oth-

er functions of the metabolism, for example MCA methodology and bifurcation analysis, it is necessary for the tool to be powered by a more efficient algorithm. MatLab is one of the most popular software systems in applied mathematics, so integrating MatLab [http://www.mathworks.com/] in Petri net models is probably a good solution. In addition, MatLab can itself be applied as an attractive Petri net tool builder. It is now possible to analyze and visualize Petri net models by transferring them to convenient graphical design tools. Export to matrix representation in MatLab is possible, and Svádová [78] has reported an approach using the MatLab standard to build a Petri nets toolbox that enabled Petri net modeling, analysis, and visualization of simulation results.

19.6 Examples and Discussion

Model-driven approaches enable conceptual and integrative modeling of large-scale molecular networks. Without directly knowing single objects, processes, and properties, biochemical classification, mechanisms and resulting networks are modeled conceptually. Data integration enables the modeler to automatically fill the conceptual model with data loaded from molecular databases. Although data integration is based on common database models only, we introduced advanced concepts for the modeling of systems based on data integration.

At the objects level, views enable conceptual modeling of different types of network, e.g. metabolic or regulatory gene networks. Using searches based on graphs the com-

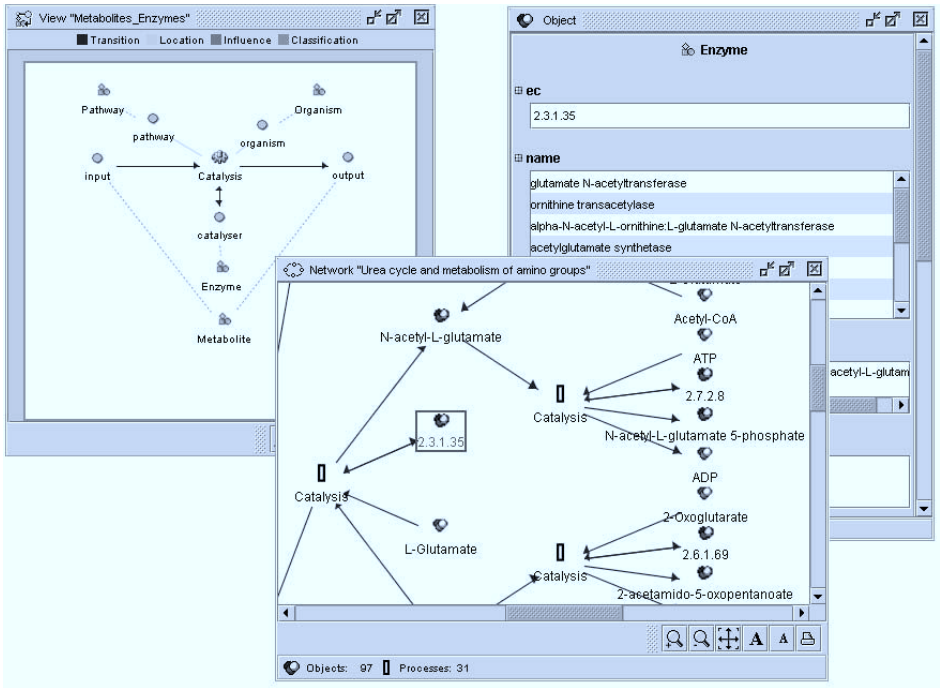


Fig. 19.10 Conceptual model and integrated pathway.

putation of bound networks, e.g. alternative pathways, is possible. The conceptual specification of loci enhances the model with information about the placement of each object and process; this enables filtering of networks by, e.g., organisms, tissues, and pathways.

At the left of Fig. 19.10 we show an example of a view which retrieves pathway-related information from an integrated database. Objects of the classes "Metabolite" and "Enzyme" participate in reactions of the type "Catalysis". Pathways and organisms specify the locations of processes. By filtering the view using the parameters *pathway* = "Urea cycle*" we obtain all the processes related to the urea cycle pathway. The network automatically retrieved is shown in middle of the figure. By selecting objects from the network found we obtain additional information merged from the data sources, which is here the enzyme ornithine transacetylase.

Object-oriented approaches and Petri nets are new methods in the domain of metabolic modeling and simulation research. Both concepts are suitable because they represent the natural behavior of these systems. A Petri net model of the urea cycle is presented and simulated by VON++ in this section. A rough model and its data is obtained from the iUDB. The urea cycle hybrid Petri net model is shown in Fig. 19.11. This model of the intracellular urea cycle is made of the composition of the gene regulatory network and metabolic pathways. It comprises 153 Petri net elements, 15 kinetic blocks, 42 dynamic variables, and 23 reaction constants.

The dynamic behavior of the model system, such as metabolite fluxes, NH_4^+ input, and urea output are well described by continuous elements whereas control of gene expression is outlined with discrete elements. The initial values of the variables

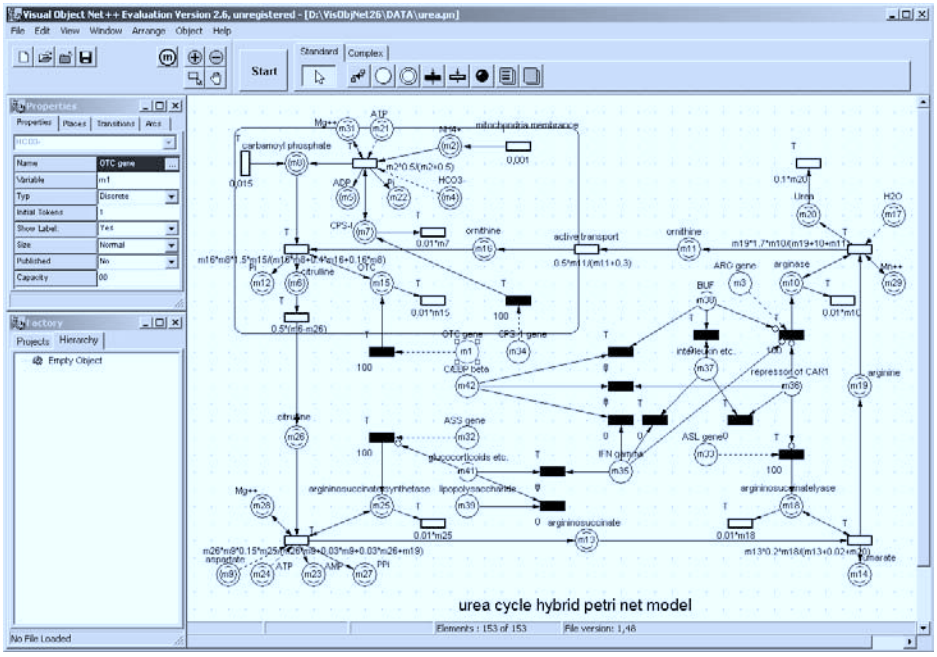


Fig. 19.11 A screenshot of VON++.

were assigned and tuned so that the behavior of the model system would comply maximally with available experimental data on the dynamic characteristics of the system's behavior. The places of the main metabolites are directly linked to their transitions and the reaction rates are obtained by use differential equations. In the model, inhibitor arcs are used to represent negative effects of repressors and/or inhibitors of gene expression. On the biochemical reaction level, negative effects of metabolites are expressed as enzyme inhibition; these include competitive inhibition, noncompetitive inhibition, irreversible inhibition, and feedback inhibition. Sequentially, regulation of urea cycle enzyme activity can be achieved in two ways. First, gene expression regulated by activators and inhibitors controls enzyme synthesis whereas enzyme synthesis and degradation determine the amounts of the enzymes. Second, the activities of these enzymes can be altered during metabolic catalysis. The Petri nets model is executed; the dynamic behavior is observed graphically in Fig. 19.12.

Hybrid Petri nets enable easy incorporation of qualitative insights into a pure mathematical model and adaptive identification and optimization of key parameters to fit system behavior observed in gene regulated metabolic pathways. The advantages of ap-

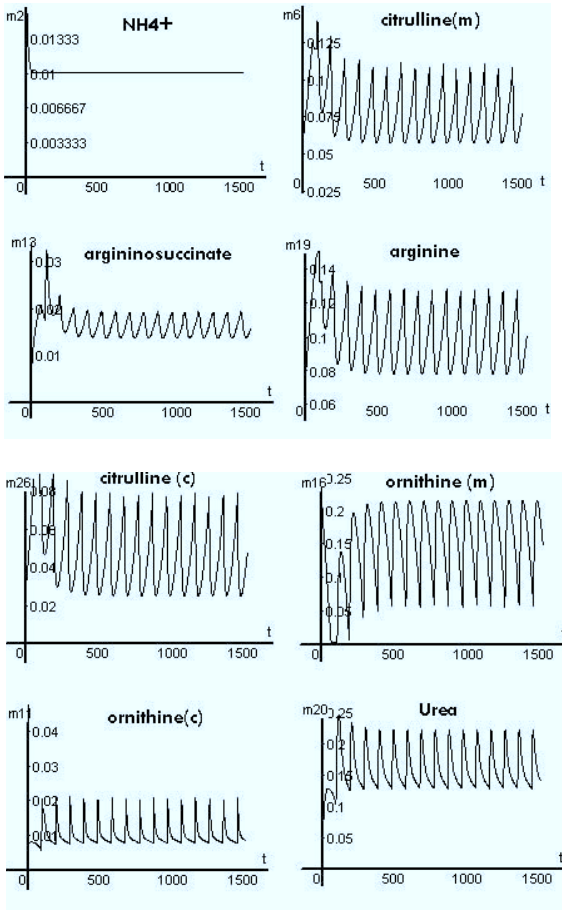


Fig. 19.12 Results from simulation of the urea cycle. Concentrations of NH_4^+, urea and other metabolites remain constant within a certain range. The oscillations arise from production of enzymes encoded discretely by the corresponding genes.

plying hybrid Petri nets to metabolic modeling and simulation are:

- with discrete and continuous events, the HPN can easily handle gene regulatory and metabolic reactions;
- the HPN model has a user-friendly graphical interface which enables easy design, simulation and visualization; and
- powered with mathematical equations, simulation is executed and dynamic results are visualized.

Using powerful Petri nets and computer techniques, data for metabolic pathways, gene regulation, and signaling pathways can be converted for Petri net destination application. Thus, a virtual cell Petri net model can be implemented, enabling attempts to understand cell activity

Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft (DFG).

URL List

BioCyc	http://biocyc.org/
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
Boehringer Mannheim	http://us.expasy.org/tools/pathways
BRENDA	http://www.brenda.uni-koeln.de/
DDBJ	http://www.ddbj.nig.ac.jp/
EcoCyc	http://ecocyc.org/
EMBL	http://www1.embl-heidelberg.de/
Entrez	http://www.ncbi.nlm.nih.gov/Entrez/
EPD	http://www.epd.isb-sib.ch/
ExPASy	http://www.expasy.ch/
GeneCard	http://bioinfo.weizmann.ac.il/cards/
GeneNet	http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/
ISYS	http://www.ncgr.org/isys/
KEGG	http://www.genome.ad.jp/kegg/
Klotho	http://www.biocheminfo.org/klotho/
MetaCyc	http://metacyc.org/
NCBI	http://www.ncbi.nlm.nih.gov/
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/
OMIM	http://www.ncbi.nlm.nih.gov/omim/
PathAligner	http://bibiserv.techfak.uni-bielefeld.de/pathaligner/
PDB	http://www.rcsb.org/pdb/
PIR	http://pir.georgetown.edu
SRS	http://srs.ebi.ac.uk/
Swiss-Prot	http://us.expasy.org/sprot/
TAMBIS	http://imgproj.cs.man.ac.uk/tambis/
TRANSFAC and TRANSPATH	http://www.biobase.de/
WIT/EMP	http://wit.mcs.anl.gov/WIT2/

References

1 Lehninger AL, Nelson DL, Cox MM (1993) Principles of biochemistry. 2nd ed., New York, pp 360

2 Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, Brazma A (2003) From Gene Networks to Gene Function. Genome Research 13:2577–2587

3 Schmid CD, Praz V, Delorenzi M, Périer R, Bucher P (2004) The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. Nucleic Acids Res. 32:D82–D85

4 Gibson DM, Harris RA (2002) Metabolic regulation in mammals. Taylor and Francis, London

5 Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2004) GenBank: update. Nucleic Acids Res. 32(1):D23–D26

6 Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, Faruque N, Garcia-Pastor M, Harte N, Kanz C, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Stoehr P, Stoesser G, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R (2004) The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 32(1):D27–D30

7 Miyazaki S, Sugawara H, Ikeo K, Gojobori T, Tateno Y (2004) DDBJ in the stream of various biological data. Nucleic Acids Res. 32(1):D31–D34

8 Boeckmann B, Bairoch A, Apweiler R (2003) The SWISSPROTprotein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31(1):365–370

9 McGarvey PB, Huang H, Barker WC, Orcutt BC, Garavelli JS, Srinivasarao GY, Yeh LS, Xiao C, Wu CH (2000) PIR: a new resource for bioinformatics. Bioinformatics16(3):290–291

10 Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 30(1):47–49

11 Kanehisa M, Goto S, Kawashima S (2002) The KEGG databases at GenomeNet. Nucleic Acids Res. 30(1):42–46

12 Overbeek R, Larsen N, Pusch GD (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. 28(1):123–125

13 Kolchanov NA, Nedosekina EA, Ananko EA, Likhoshvai VA, Podkolodny NL, Ratushny AV, Stepanenko IL, Podkolodnaya OA, Ignatieva EV, Matushkin YG (2002) GeneNet database: description and modeling of gene networks. In Silico Biol. 2(2):97–110

14 Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. 28(1):316–319

15 Takai-Igarashi T, Kaminuma T (1999) A Pathway Finding System for the Cell Signaling Networks Database. In Silico Biology 1:129–146

16 Schacherer F, Choi C, Gotze U, Krull M, Pistor S, Wingender E (2001) The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. Bioinformatics 17(11):1053–1057

17 Etzold T, Ulyanow A, Argos P (1996) SRS: Information Retrieval System for Molecular Biology Data Banks. Methods Enzymol. 266:114–128

18 Kohn MC, Letzkus W (1983) A Graph-theoretical Analysis of Metabolic Regulation. Journal of Theoretical Biology 100(2):293–304

19 Brutlag DL, Galper AR, Millis DH (1992) Knowledge-Based Simulation of DNA Metabolism: Prediction of Action and Envisionment of Pathways, in: Artificial Intelligence and Molecular Biology (Hunter L. Ed). MIT Press, Cambridge, MA.

20 Reddy VN, Mavrovouniotis ML, Liebman MN (1993) Petri Net Representation in Metabolic Pathways, in: Proceedings First International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, pp 328–336

21 Hofestädt R, Meinecke F (1995) Interactive modelling and simulation of biochemical networks. Computers in Biology and Medicine 25:321–334

22 Ellis LB, Speedie SM, McLeish R (1998) Representing metabolic pathway information: an object-oriented approach. Bioinformatics 14(9):803–806

23 Akutsu T, Miyano S, Kuhara S (2000) Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics 16(8):727–734

24 Michal G (1993) Biochemical Pathways. Boehringer, Mannheim, Penzberg

25 Forst VC, Schulten K (199) Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. J. Comput. Biol. 6:343–360

26 Karp DP, Riley M, Paley SM (2002) The MetaCyc Database. Nucleic Acids Res. 30(1):59–61

27 Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M (1999) EcoCyc: Encyclopedia of Escherichia coli Genes and Metabolism. Nucl. Acids Res. 27(1):50–53

28 Karp PD (1995) A Strategy for Database Interoperation. J. Comput. Biol. 2:573–586

29 Davidson SB, Overton C, Buneman P (1995) Challenges in Integrating Biological Data Sources. J. Comput. Biol. 2:557–572

30 Siepel A, Farmer A, Tolopko A, Zhuang M, Mendes P, Beavis W, Sobral B (2001) ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. Bioinformatics 17:83–94

31 Haas LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC (2001) DiscoveryLink: A system for integrated access to life sciences data sources. IBM Systems Journal 40:489–511

32 Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, Brass A (2000). TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. Bioinformatics 16:184–185

33 Tatusova TA, Karsch-Mizrachi L, Ostell JA (1999) Complete genomes in WWW Entrez: data representation and analysis. Bioinformatics 15:536–543

34 Davidson SB, Overton C, Tannen V, Wong L (1997) BioKleisli: a digital library for biomedical researchers. International Journal on Digital Libraries 1:36–53

35 Freier A, Hofestädt R, Lange M, Scholz U, Stephanik A (2002) BioDataServer: A SQL-based service for the online integration of life science data. In Silico Biology 2(2):37–57

36 Papakonstantinou Y, Abiteboul S, Garcia-Molina H (1996) Object fusion in mediator systems, in: 22nd Conference on Very Large Data Bases (Vijayaraman TM, Buchmann AP, Mohan C, Sarda NL Ed) Bombay, India, Morgan Kaufmann

37 Freier A, Hofestädt R, Lange M (2003) iUDB: An object-oriented system for modelling, integration and analysis of gene controlled metabolic networks. In Silico Biology 3(1/2):215–227

38 Reddy VN, Mavrovouniotis ML, Liebman MN (1994) Modeling Biological Pathways: A Discrete-Event Systems Approach in Molecular Modeling. in: ACS Symposium Series, Vol. 576 (Kumosinski TF, Liebman MN, Ed), Washington DC, pp 221–234

39 Petri CA (1962) Kommunikation mit Automaten, Dissertation, Institut für Instrumentelle Mathematik, Schriften des IIM Nr. 2, Bonn

40 Reisig W (1986) Petri Netze. Springer, Heidelberg

41 David R, Alla H (1992) Petri Nets and Grafcet – Tools for Modeling Discrete Event Systems. Prentice Hall

42 Wang J (1998) Timed Petri Nets: Theory and Application. Kluwer Academic Publishers, Boston

43 Wang J, Jin C, Deng Y (1999) Performance analysis of traffic networks based on stochastic timed Petri net models, in: Proc. 5th Int. Conf. on Engineering of Complex Computer Systems, Las Vegas, pp 77–85

44 Kurt J (1997) Coloured Petri Nets – Basic Concepts, Analysis Methods and Practical Use. 2nd edition, Springer, Berlin

45 Genrich HJ (1987) Predicate/Transition Nets, in: Lecture Notes in Computer Science, vol 254: Petri Nets: Central Models and Their Properties, Advances in Petri Nets (Brauer W, Reisig W, Rozenberg G Ed), Bad Honnef, Springer, pp 207–247
46 Hofestädt R, Thelen S (1998) Quantitative modeling of biochemical networks. In Silico Biology 1(1):39–53
47 Hofestädt R (1994) A petri net application of metabolic processes. Journal of System Analysis Modelling and Simulation 16:113–122
48 Reddy VN, Liebman MN, Mavrovouniotis ML (1996) Qualitative analysis of biochemical reaction systems. Comput Biol. Med. 26(1):9–24
49 Chen M (2000) Modelling the glycolysis metabolism using hybrid petri nets. in: DFG-Workshop – Modellierung und simulation metabolischer netzwerke, Mai 19–20, Magdeburg, Germany, pp 25–26
50 Koch I, Schuster S, Heiner M (1999) Simulation and analysis of metabolic networks by time_dependent Petri nets, in: Proceedings of the German Conference on Bioinformatics GCB '99, Oct 4–6, Hanover, Germany
51 Kueffner R, Zimmer R, Lengauer T (2000) Pathway Analysis in Metabolic Databases via Differential Metabolic Display (DMD). Bioinformatics 16(9):825–836
52 Goss PJE, Peccoud J (1999) Analysis of the Stabilizing Effect of Rom on the Genetic Network Controlling ColE1 plasmid replication. Pacific Symposium on Biocomputing, pp 65–76
53 Matsuno H, Doi A, Nagasaki M, Miyano S (2000) Hybrid Petri net Representation of Gene Regulatory Network. Pacific Symposium on Biocomputing, pp 341–352
54 Hofestädt R, Lautenbach K, Lange M (2000) Proc. Modellierung und Simulation Metabolischer Netzwerke, DFG-Workshop, Magdeburg, Germany
55 Genrich H, Kueffner R, Voss K (2001) Executable Petri Net Models for the Analysis of Metabolic Pathways. International Journal on Software Tools for Technology Transfer 3(4):394–404
56 Heiner M, Koch I, Voss K (2001) Analysis and simulation of steady states in metabolic pathways with Petri nets, in: CPN '01 – Third Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools, University of Aarhus, Denmark, pp 15–34
57 Srivastava R, Peterson MS, Bentley WE (2001) Stochastic kinetic analysis of the Escherichia coli stress circuit using σ32-targeted antisense. Biotechnology Bioengineering 75(1):120–129
58 Oliveira JS, Bailey CG, Jones-Oliveira JB, Dixon DA (2001) An Algebraic-Combinatorial Model for the Identification and Mapping of Biochemical Pathways. Bulletin of Mathematical Biology 63(6):1163–1196
59 Peleg M, Yeh I, Altman RB (2002) Modelling biological processes using workflow and Petri Net models. Bioinformatics 18(6):825–837
60 Hofestädt R (2003) Petri nets and the simulation of metabolic networks. In Silico Biol. 3(3):321–322
61 Zevedei-Oancea I, Schuster S (2003) Topological analysis of metabolic networks based on Petri net theory. In Silico Biol. 3(3):323–345
62 Voss K, Heiner M, Koch I (2003) Steady state analysis of metabolic pathways using Petri nets. In Silico Biol. 3(3):367–387
63 Chen M, Hofestadt R (2003) Quantitative Petri net model of gene regulated metabolic networks in the cell. In Silico Biol. 3(3):347–365
64 Matsuno H, Tanaka Y, Aoshima H, Doi A, Matsui M, Miyano S (2003) Biopathways representation and simulation on hybrid functional Petri net. In Silico Biol. 3(3):389–404
65 Ruth M, Hannon B (1997) Modeling dynamic biological systems. Springer, New York
66 Goldberg Elaine, http://mvhs1.mbhs.edu/mvhsproj/CellResp/cell_table.html
67 Doi A, Drath R, Nagasaki M, Matsuno H, Yamino S (1999) Protein Dynamics Observations of Lambda Phage by Hybrid Petri Net. Genome Informatics 10:217–218
68 Matsuno H, Doi A, Drath R, Miyano S (2001) Genomic Object Net: Basic Architecture for Representing and Simulating Biopathways, in: RECOMB '01
69 Chen M (2002) Modelling and Simulation of Metabolic Networks: Petri Nets Approach and Perspective, in: Proc. ESM 2002, 16th European Simulation Multiconference, June 3–5, Darmstadt, Germany, pp 441–444
70 Goss PJE, Peccoud J (1998) Quantitative Modeling of Stochastic Systems in Molecular

Biology by Using Stochastic Petri Nets. Proc. Natl. Acad. Sci. USA, pp 6750–6755

71 Gronewold A, Sonnenschein M (1997) Asynchronous Layered Cellular Automata for the Structured Modeling of Ecological Systems, in: 9th European Simulation Symposium (ESS '97), SCS, pp 286–290

72 Gronewold A, Sonnenschein M (1998) Event–based Modelling of Ecological Systems with Asynchronous Cellular Automata. Ecological Modeling 108:37–52

73 Voss K (2000) Ausführbare Petrinetz Modelle zur Simulation Metabolischer Pfade, in: DFG-Workshop, Modellierung und Simulation Metabolischer Netzwerke, 2000, Mai 19–20, Magdeburg, Germany, pp 11–13

74 Matsuno H, Murakami R, Yamane R, Yamasaki N, Fujita S, Yoshimori H, Miyano S (2003) Boundary formation by notch signaling in Drosophila multicellular systems: experimental observations and a gene network modeling by Genomic Object Net. Pacific Symposium on Biocomputing, pp 152–163

75 Nagasaki M, Doi A, Matsuno H, Miyano S (2003) Genomic Object Net: a platform for modeling and simulating biopathways. Applied Bioinformatics 2(3):181–184

76 Chen M, Freier A, Köhler J, Rüegg A (2002) The Biology Petri Net Markup Language. In: Lecture Notes in Informatics (Desel J, Weske M, Ed), Proc. Promise 2002, Oct. 9–11, Potsdam, Germany, Vol. 21:150–161

77 Fell D (1992) Metabolic control analysis: a survey of its theoretical and experimental development. Biochem. J. 286:313–330

78 Svádová M, Hanzalek Z (2000) Matlab Toolbox for Petri nets, in: INOVACE 2000, Praha: Asociace inova_ního podnikání _R, pp. 15–20

20
Systems Biology

Nathan Goodman

20.1
Introduction

Numerous definitions of systems biology are offered in the literature and on the web. TheFreeDictionary.com has a good general definition: "Systems biology is an academic field that seeks to integrate biological data as an attempt to understand how biological systems function. By studying the relationships and interactions between various parts of a biological system ... it is hoped that an understandable model of the whole system can be developed." Although this definition could be applied to many areas of biology, the term as commonly used is limited to work in molecular biology.

Systems biology is a young, interdisciplinary field with fluid boundaries. The literature is vast, duplicative, and unwieldy. Fortunately, excellent reviews have been published from various points of view which provide good introduction to the primary literature. This chapter includes a guide to these reviews in lieu of inline citations. Systems biology rests on analyses of large-scale datasets – gene expression, protein–protein interactions, gene–phenotype relationships, and many others – individually or in combination. These datasets provide a global view of the system under study, albeit incomplete and imperfect. As a first step, investigators use these data to study the general properties of the system. A more ambitious goal is to construct mathematical models that explain how the system works and predict its response to experimental or natural perturbations. Several next steps are possible. Some investigators study the models themselves, looking for general patterns and properties that reveal biological insights. Others work on better modeling methods. Yet others proceed to test the models' predictions using classical laboratory methods. A more far-reaching strategy is to refine and expand the models through additional rounds of large-scale data generation and data analysis. Taken to this level, the practice of systems biology becomes an iterative process in which researchers analyze large-scale datasets to devise models, analyze the models to make predictions, and conduct further large-scale experiments to test the predictions and refine the models. Few studies have reached this level, but this is the vision expounded by many leaders in the field.

Because of the emphasis on large-scale data sets and experimentation, the methods of systems biology cannot be applied in all

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

situations. A prerequisite is having an experimental system that is amenable to large-scale experimentation. This generally means working with a system that can be readily perturbed in numerous ways and interrogated using global methods. To get started without spending a lot of time and money, it is desirable to work with experimental systems for which baseline datasets have already been collected.

Most systems biology research to date has studied simple model organisms. Yeast is far and away the most widely used model. There is also considerable work on prokaryotes – including *E. coli* and an archeon *Halobacterium* – and a smattering of work on worm, fly, and sea urchin. Research on human and mammalian models is limited, and is generally small scale, preliminary in nature, or peripheral to studies of simpler organisms.

Systems biology emerged from the confluence of three trends. First is the continued success of molecular biology in charactering the genes and proteins involved in biological processes, and the organization of these elements into pathways of increasing scope and complexity. Second is the growing success of mathematical modelers in devising pathway models that can be analyzed and simulated. Third, and most pressing, is the explosion in methods for large-scale data production, starting with the ramp-up of DNA sequencing in the early 1990s and continuing with technologies for measuring gene expression, protein–protein interactions, gene–phenotype relationships, and many others. The traditional models of molecular biology have been unable to cope effectively with the flood of data. Many investigators simply cherry pick the best bits for further study, leaving most of the data untouched on the tree. "Best" in this setting usually meant players – genes or proteins – that can be readily connected to existing pathways, or else there would be no way to interpret the findings – to "tell a story" as the saying goes.

Systems biology is an attempt to rescue the data left behind by traditional methods, to let the data tell the story with the help of sophisticated data analysis. It is simultaneously an evolutionary extension of established approaches and a radical new way of doing biology.

20.2 Data

20.2.1 Available Data Types

Systems biology has a voracious appetite for data and works with many kinds of data produced in many different ways.

The foundation is data generated by large-scale data-production methods and includes:

- genome and gene sequences;
- gene expression profiles generally from microarrays and, to a lesser extent, SAGE;
- protein–protein interactions from yeast two-hybrid and affinity purification methods;
- gene–phenotype relationships and gene–gene interactions from large-scale gene-deletion projects and, increasingly, from RNA interference;
- protein–DNA interactions from ChIP-chip;
- protein identification and abundance from mass spectroscopy; and
- sub-cellular protein localization data from systematic imaging studies.

As new data productions methods come online, additional data types will come into use.

These laboratory methods are augmented by data produced by large-scale computational analysis. This includes:

- predictions of transcription factor binding sites;
- identification of binding domains in protein sequences;
- predictions of binding domains from empirically determined or computed three dimensional protein structures
- functional clustering of genes and proteins through text mining of the literature,

and many others.

Biological interpretation of large-scale datasets requires connection of novel data to known biological "truth". The connection comes from manually curated datasets produced by experts. Examples include gene annotation databases, lists of known transcription factor binding sites, curated pathways, and data from small scale protein–protein interaction experiments that are manually extracted from the literature. These curated datasets are occasionally quite large and greatly expand the data available from large-scale experiments.

20.2.2 Data Quality and Data Fusion

Many large-scale datasets suffer from high error rates. Error rates have been studied extensively for protein–protein interaction data from yeast. Large-scale datasets produced by different laboratories using similar methods have been compared, as have datasets produced by the same or different laboratories using different methods, specifically yeast two-hybrid vs. affinity purification. Other studies have compared large-scale datasets with interactions reported in the literature or with known protein complexes. The conclusion is that 50% or more of reported interactions are false, and that many known interactions are not detected even under conditions that favor their detection. There are fewer studies of other data types, but the general consensus is that all high-throughput data are suspect. No single data point can be believed.

To draw valid inferences from such error-prone data, it is necessary to combine multiple data points, preferably from different sources. Combining data in this way is a central theme of systems biology.

Another line of study is to assess the concordance of different data types, for example protein–protein interaction vs. gene expression data. Of course, one would not expect perfect agreement in such instances, because each data type measures a different aspect of the system. But, if the concordance were too low, it would call into question the basic premise of combining multiple data types to overcome the errors in each.

Some research in this area is descriptive, and seeks to characterize data quality and concordance, whereas other work is prescriptive and aims to devise ways of improving the situation. The latter includes work on statistical methods for assessing the quality of individual data points, making it possible to filter the dataset and select data which are most reliable. It also includes work on methods for combining multiple data types to create a reliable merged dataset.

Merging of multiple data types to produce a more reliable dataset is a major theme in systems biology. Remarkably, there is no standard term for this in the field. Most often, it goes by the banal label of *data integration*, a very generic term that refers to any activity in which data from multiple sources is combined. A better term, I think, is *data fusion*. This is a term borrowed from the field of remote sensing

where it refers to the process of combining multiple incomplete data streams to yield a more comprehensive view of remote phenomena.

20.3 Basic Concepts

20.3.1 Systems and Models

The term *system* is very general. Two standard definitions are, "an assemblage of inter-related elements comprising a unified whole" (from TheFreeDictionary.com), and "a group of interacting, interrelated, or interdependent elements forming a complex whole" (from Dictionary.com). These definitions are hardly limiting, because everything in the universe, with the exception of elementary particles, is composed of interacting or inter-related parts. This includes everything in biology. The real meaning of the term comes from usage: when we call something a system, the connotation is that we intend to study the whole in terms of its parts and their inter-relationships, and the properties we intend to study arise from non-trivial interactions among the parts.

A *model* is "an abstract or theoretical representation of a phenomenon" (from TheFreeDictionary.com) that represents some aspects of reality and ignores others. A good model represents those aspects that are significant for the problem at hand and ignores all others.

A simple model of a biological system might represent the proteins encoded by the genome of an organism, potential physical interactions among those proteins determined by means of large-scale yeast two-hybrid experiments, and estimates of protein abundance measured by gene expression microarray experiments. Obviously, this model is a gross simplification of reality. Such simplification is the essence of modeling.

Systems have *static* properties that do not change and *dynamic* properties that do. Dynamic properties can vary over any dimension of interest, including time, space, experimental condition, genetic background, disease state, etc. In our simple example, static properties include the proteins themselves (because we are working with a single organism) and the protein–protein interactions (because yeast two-hybrid experiments are usually performed under generic baseline conditions). The dynamic properties are the gene-expression levels, assuming these experiments are conducted over a range of time points or experimental conditions.

20.3.2 States

The *state* of a system (more precisely, the state of a system model) is the ensemble of its dynamic properties at one point in the multidimensional space over which these properties vary. In the example of the preceding section, the state is simply a gene expression profile for one time point and experimental condition. The totality of all possible states for a system is called its *state-space*.

Most biological systems are *homeostatic*, meaning that if the system is unperturbed, it will remain in a single state or, more typically, will travel among a set of related states. The state or set of states is a *stable* region of the state space. When the system is perturbed, it might deviate temporarily from the stable region and then return, or travel to a new stable region, or become unstable and fluctuate between widely separated states.

An important goal in analyzing biological systems is to understand the *control mecha-*

nisms that keep the system in the appropriate stable region or that guide the system from one stable region to another.

20.3.3 Informal and Formal Models

In biology as traditionally practiced, most models are informal and are intended to communicate ideas among fellow scientists. Ideker and Hood explain it well [1]: "Conventionally, a biological model begins in the mind of the individual researcher, as a proposed mechanism to account for some experimental observations. Often, the researchers represent their ideas by sketching a diagram using pen and paper. This diagram is a tremendous aid in thinking clearly about the model, in predicting possible experimental outcomes, and in conveying the model to others. Not surprisingly, diagrammatic representations form the basis of the vast majority of models discussed in journal articles, textbooks, and lectures."

In systems biology, the models are formal (i.e. mathematical) and can be analyzed and simulated rigorously. It is important that these models be intelligible to the systems biologists who work with them, but there is no pretense that the models will make sense to traditionally trained scientists.

The construction of a formal model is a difficult balancing act which must accommodate the conflicting demands of realism and tractability. As a model represents more aspects of the real phenomenon, it becomes more complex and more difficult to study. On the other hand, if the model is not sufficiently realistic, the conclusions drawn from it might be biologically meaningless. In words attributed to Albert Einstein, "Make everything as simple as possible, but not simpler." A further constraint is the availability of data: there is little point in modeling an aspect of biological reality for which no data are available except, perhaps, for theoretical purposes.

20.3.4 Modularity

A *module* is a portion of a model that is reasonably self-contained. The elements within a module can be highly inter-related, but there should be relatively few connections between elements of different modules. A model that is *modular* can be decomposed into smaller units – the modules – which can be understood more or less independently. Modules can be formed from static or dynamic aspects of a model. Modules in biological systems include structures, for example protein complexes, organelles, and membranes, and processes and pathways.

A modular model can be *hierarchical* with low-level modules serving as building blocks for higher-level modules. Thus, the ribosome (a protein complex) is a component of translation (a process), and translation is a component of numerous cellular activities, e.g. proliferation. Hierarchical modularity is an essential design principle for human-engineered systems without which large systems would be unintelligible even to their designers.

It is self evident that biological systems are modular and hierarchical at least in some respects. At the bottom of the hierarchy are genes, proteins, and other molecules. Next are direct interactions among these elements, for example, physical protein–protein interactions or the binding of a protein (transcription factor) to DNA for the purpose of regulating the expression of a gene. These direct interactions are organized into less direct, but still local, phenomena such as protein complexes and groups of co-regulated genes (sometimes called *transcription modules* or *regulons*).

These local interactions are organized into higher-level modules that implement particular processes, e.g. transcription or transport, and these processes are combined to achieve cellular functions, e.g. proliferation.

There is growing evidence that hierarchical modularity is pervasive in biological systems and is a fundamental property of evolved systems. A key goal is to discover the hierarchical modularity that underlies phenomena of interest. Whether or not nature is hierarchical, models of large biological systems must almost certainly be hierarchical. This is because models are human-engineered and must be intelligible to the researchers who develop them.

20.4 Static Models

Static system models address structural aspects of the system – the parts and how they are related. Static models can also consider states in isolation, for example to derive relationships between gene-expression levels for different genes at one time point and experimental condition, or to identify groups of genes whose expression levels are co-regulated across conditions. Static models are much simpler than dynamic ones, and much of the work in the field to date has focused on these.

20.4.1 Graphs

Graphs have become the mathematical formalism of choice for static models in systems biology (Fig. 20.1). A *graph* can be visualized as a diagram consisting of dots, and lines connecting the dots. The dots are called *nodes* or *vertices*, and lines are called *edges* or *arcs*. Edges can be *undirected* (usually drawn as an arrowless line) or *directed* (usually drawn as a line with an arrowhead). Graphs in which all edges are directed are sometimes called *networks*. It is possible to attach additional information (called *labels*) to the nodes and edges. The *density* of a graph is the ratio of edges to nodes.

In a model of protein–protein interaction data the nodes would represent proteins and the edges would indicate which proteins interact. In the undirected case the two endpoints are symmetric – an edge from protein A to protein B would indicate that A interacts with B and that B interacts with A. Directed edges are able to represent asymmetric relationships, for example to distinguish between the bait and prey proteins in a yeast two-hybrid experiment. Edge labels can be used to indicate the kind or strength of evidence that supports the interaction.

Graphs can naturally represent all kinds of pairwise interaction data or relationships. For example, data telling which proteins regulate which genes can be represented by a graph in which each node represents both a gene and its protein product (this is biologically sloppy, but remember it is just a model), and a edge from node A to node B means that protein A regulates gene B. Gene expression data from knockout experiments can be represented by a graph whose nodes represent genes, and an edge from gene A to gene B means that knocking out A affects the expression of B. Genes with correlated expression patterns can be represented by a graph whose nodes represent genes, and whose edges indicate which genes have highly correlated profiles.

It is possible to combine multiple graphs to analyze multiple data types simultaneously. Merging all the graphs from the example above yields a graph indicating which proteins interact, which proteins regulate which genes, which genes affect the expression of other genes, and which genes have similar

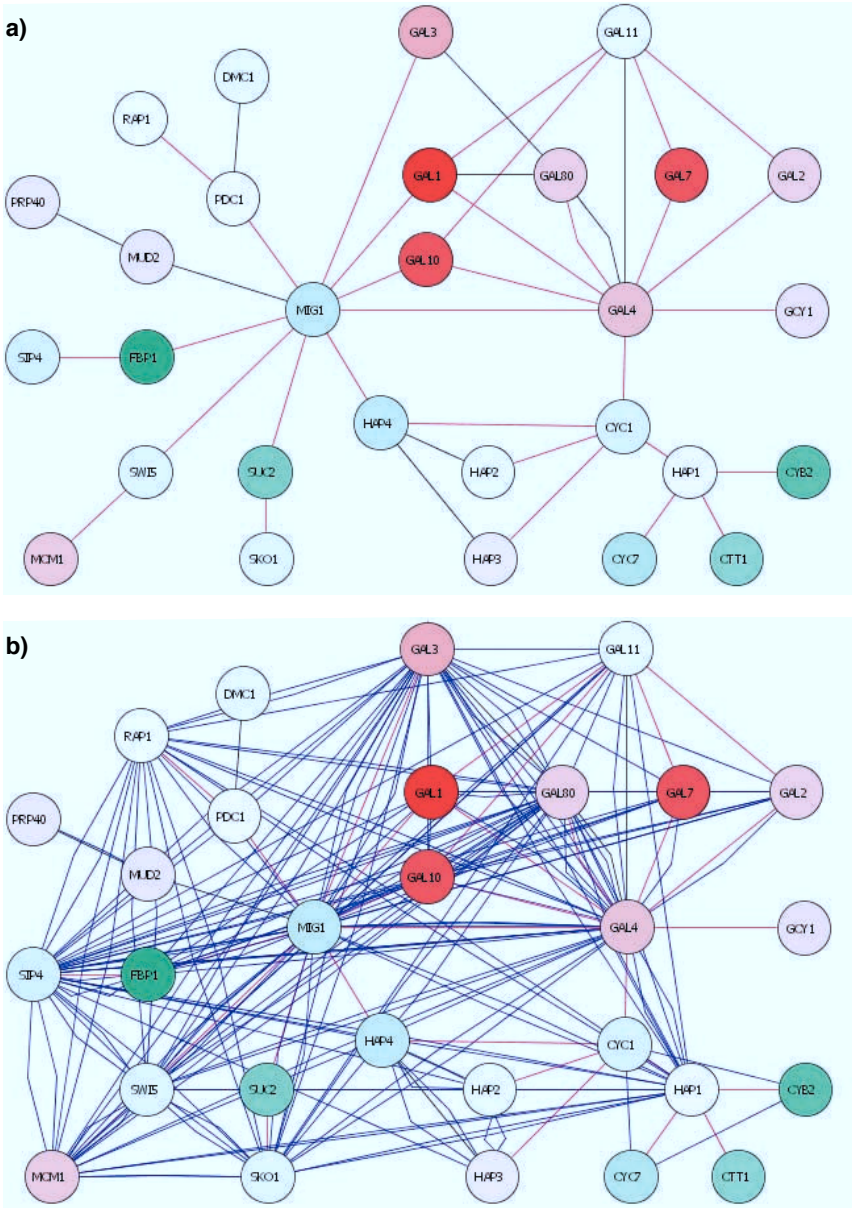


Fig. 20.1 Static model depicted as a graph (a) This graph is a model of protein–protein and protein–DNA interactions centered on the *GAL* genes in yeast. The data are from Ref. [2]. The graph was drawn using Cytoscape [3] and is based on the small network tutorial on the Cytoscape website (http://www.cytoscape.org). Black edges are protein–protein interactions, and red edges are protein–DNA interactions. Node colors reflect gene-expression levels under one of twenty experimental conditions. (b) The same graph with additional edges connecting nodes that are annotated in GO as participating in the same biological process (at level 6). This illustrates the ability of graphs to represent multiple kinds of relationship and also the difficulty of working with dense graphs.

expression patterns. It is easy to imagine analyzing the combined graph to look for patterns that are biologically meaningful. Although easy to imagine, it is quite challenging to do this in a mathematically rigorous fashion. Integrated analysis of this sort is a major research area in systems biology.

The study of graphs is a major topic in mathematics and computer science. Mathematicians have cataloged a vast number of interesting graph properties, and computer scientists have devised effective algorithms for solving many graph problems. The advantages of building on top of this existing knowledge are obvious.

Graphs also have some disadvantages. One is that they are only good at representing pairwise relationships. Although it is possible to represent higher-order relationships, the resulting graphs are often clumsy and difficult to manipulate. A second problems is that even with pairwise relationships, graphs are only effective over a limited range of densities. If there are too many edges or too few, the graph will not have enough structure to be useful.

20.4.2 Analysis of Static Models

A large fraction of the field is devoted to analysis of static models represented as graphs. These graphs may embody just a single data type or multiple data types that have been fused into a single dataset.

One general thrust is to use these graphs, perhaps in conjunction with other data, to identify functionally related groups of genes. This has been achieved by using graph-clustering methods which divide the graph into highly connected sub-graphs. Another approach is to look for sub-graphs that are highly correlated with gene expression changes. Having found a group of putatively related genes, the natural next step is to assign a function to the group. This is done by connecting the model to curated datasets of functional annotations. It is also possible to include functional annotations from the beginning and use them as part of the grouping process; if sufficient weight is given to the annotations, the effect is to extend existing known functional groups rather than finding groups *de novo*. A related, less ambitious problem is to assign functions to individual novel genes by connecting them to known groups.

Similar methods have been used to find groups of genes apparently regulated by the same transcription factors; such groups have been dubbed *transcriptional modules* or *regulons* by some investigators. This usually requires that the model include data on transcription factor binding from empirical protein–DNA binding studies (e.g. ChIP-chip), curated datasets, or computational prediction of transcription factor binding sites. Some authors have proposed feeding the results back to refine binding-site predictions.

Another surprisingly fruitful avenue has been the study of general, mathematical properties of these graphs. One profound conclusion is that biological systems (or, at least, static models of such systems) are *scale-free*. This implies that most nodes have a small number of neighbors whereas a few nodes have a large number of neighbors. The latter have come to be called *hubs*. Further research has explored the biological significance of hubs – relationships have been found between hubs and protein complexes, and between hubs and essential or synthetically lethal genes. Other studies of graph properties have found evidence of hierarchical modularity. Another direction examines relationships between graph properties and evolution to determine, for example, whether genes with many neighbors evolve more slowly, or whether genes

and their neighbors co-evolve. Some authors have explored evolutionary processes that can give rise to systems with the observed properties.

20.5 Dynamic Models

Dynamic models are concerned with how systems change over time, space, experimental condition, disease state, or other dimension of interest. Like a static model, a dynamic model starts by defining the parts and how they are inter-related. It then adds a definition of the allowable states the model can occupy (in other words, the state-space), and a set of *transition rules* that prescribe how the state at one point is transformed into the state at another point. Derivation of these rules is a difficult problem that is central to systems biology.

Research in dynamic modeling follows a logical progression. The first step is to develop mathematical formalisms for expressing models. Next these formalisms are used to create models for specific systems of interest. Finally the models are studied (by mathematical analysis or simulation) to gain biological insight.

In systems biology at present these steps tend to be tightly linked. The same research groups often do all three activities, and frequently a single paper will introduce a new formalism, describe a new model, and report some biological conclusions. There has been little systematic or rigorous comparison of formalisms or models. The net effect is that the field is littered with a large number of modeling formalisms and models, each used by one or a few groups, with little rationale for deciding which are best. The situation should be rectified as the field matures, but it is hard to say how quickly this will happen.

20.5.1 Types of Model

Dynamic models are classified into four broad biological categories. *Metabolic pathways* produce the substances required for cellular functioning. *Signal transduction pathways* transmit and transform information, for example, to communicate a signal from the cell surface to the nucleus. *Gene regulatory networks* control the pattern of gene expression to keep the cell in a stable region of the state space or guide the system from one stable region to another. General *regulatory networks* combine all of these.

Metabolic pathways are the mainstay of classical research on mathematical modeling of biological systems. In these models, the parts are enzymes and metabolites. The inter-relationships are enzymatic reactions that consume input metabolites and produce output metabolites. The states are concentrations of metabolites. The transition rules are rate equations or other means that specify the effect of each reaction as a function of the concentrations of its input and output metabolites.

In systems biology, gene regulatory networks are a major focus. In these models, the parts are genes and their encoded proteins which most models treat as one and the same thing. The inter-relationships indicate which proteins (transcription factors) regulate the expression of which genes and might include protein–protein relationships that can be used to deduce connections that close the loop from genes back to transcription factors. The states are gene-expression levels which also serve as surrogates for protein abundance. The transition rules specify how expression levels at one point in time affect expression levels at a later time.

20.5.2 Modeling Formalisms

Dynamic modeling formalisms are diverse. Some are mechanistic and directly encode reactions or regulatory events, for example "protein A binds B activating C which travels to the nucleus and enables the expression of gene D". Others express functional relationships and say things like "proteins A and B are required to activate expression of gene D". Yet others are quantitative and describe how the abundance of molecules at one point in time affects the abundance at the next point in time. Many are graph-based. Most formalisms are designed to enable automatic inference of models from data, but some are intended for manual use, to enable an expert to hand-craft a model. Most are designed to be studied using simulation, but some also enable mathematical analysis.

Modeling formalisms differ in the nature of the states that are allowed. The most general models allow states to be continuous; in such models, expression levels would be represented as real numbers. The other extreme is Boolean models which permit on/off values only. In between are models with discrete, but multi-valued, states. This includes qualitative models in which states are categories like high, medium, and low, and stochastic models in which states are integers. Stochastic models can directly represent the number of molecules available for various reactions; this can be important for modeling activities that involve small numbers of molecules.

Another aspect is the nature of the transition rules. These can be deterministic and define unequivocally how one state is transformed into the next, or probabilistic and define distributions of next states. The mathematical framework for expressing the transition rules can also vary. Some examples are differential equations, difference equations, Boolean logic, general mathematical or computational logic, and Bayesian or other probabilistic networks.

20.6 Summary

Systems biology, like most science, is driven by data. The foundation is data generated by large-scale data-production methods augmented by large computationally derived datasets. Biological interpretation comes from manually curated datasets produced by experts, some of which are quite large and greatly expand the data available from large-scale experiments. The field can be expected to embrace new data sources and data types as they become available.

Large-scale datasets are often plagued by high error rates, and a major theme in systems biology is coping with these errors. The usual solution is to combine multiple data sources and data types to produce a more reliable dataset, a process we term data fusion.

The main activity in systems biology is modeling of biological systems. A model is an abstraction of reality that represents aspects that are significant for the problem at hand and ignores all others. Models in biology have traditionally been informal and serve to communicate ideas among fellow scientists. In systems biology, the models are formal (i.e. mathematical) and can be analyzed and simulated rigorously.

Static models represent the structure of a system – the parts of the system and their inter-relationships. Graphs have emerged as the mathematical formalism of choice for static models in systems biology. Graphs can naturally represent all kinds of pairwise relationships, and a huge body of existing mathematical and computer science research on graphs can be exploited directly.

A major research area in systems biology is analysis of static models represented as graphs. One general topic is identification of groups of functionally related genes, for example, by means of graph clustering. An important case is groups of genes that are regulated by the same transcription factors. A second theme is study of general, mathematical properties of these graphs. A key finding is that these graphs are *scale-free*, implying the presence of a small number of highly connected hubs. Research on the biological significance of hubs has found connections between hubs and protein complexes, and between hubs and essential or synthetically lethal genes. Other studies of graph properties have found evidence of hierarchical modularity.

Dynamic models describe how the system state changes over any dimension of interest, including time, space, experimental condition, or disease state. The core of a dynamic model is a set of *transition rules* that prescribe how the state at one point is transformed into the state at another point. Dynamic modeling formalisms are diverse and range from highly mechanistic to highly abstract. They vary as to the kinds of states that are allowed, the nature of the transition rules, and the mathematical framework for expressing these rules. Most enable automatic inference of models from data, but some are intended for manual hand crafting of models.

Dynamic modeling of complex biological systems is a central activity in systems biology. Much current work is focused on gene regulatory networks which are important control mechanisms. A large number of modeling formalisms and models have been published, but there has been scant work comparing these results in a systematic or rigorous fashion. At present, no methods have gained widespread acceptance, and there is little evidence favoring one formalism or model over the others.

Systems biology is a vibrant field. Investigators are aggressively pursuing many different problems with different approaches. The field will probably mature over the next few years, but for now it remains a frontier with few rules to guide newcomers.

20.7 Guide to the Literature

Because of the interdisciplinary nature of the field, the systems biology literature is vast and diverse. A good way to get started with the literature is to work from the many excellent reviews that have been published. Here we provide a list of recommended reviews.

20.7.1 Highly Recommended Reviews

Comments	*Refs.*
General perspectives. Note that reference [4] avoids the term systems biology, but covers the same ground	4–7
General technical overview. An excellent introduction to many topics in systems biology: data production methods for protein–protein and protein–DNA interactions; computational methods for prediction of protein–protein interactions; integration (fusion) of protein–protein interaction datasets; inference of Boolean and Bayesian networks; and graph properties	8
Data production. Comprehensive overview of the major large-scale data types used in systems biology	9
Data quality. Excellent discussion of data quality issues in large-scale datasets – microarray vs. SAGE; gene expression vs. protein abundance; ChIP-chip vs. gene expression; protein–protein interactions; gene deletions	10

Comments	Refs.
Data fusion. High level, but insightful, review of data fusion issues and what can be learned by combining data types	11
Graph properties of static models. Detailed introduction to analysis of graph properties of static models. Includes graph motifs, evidence for hierarchical modularity, and evolutionary considerations	12
Interaction inference. Review of protein interaction domains, and their relevance for inferring regulatory networks	13
Dynamic modeling of gene regulatory networks	
Very detailed and comprehensive review of dynamic modeling formalisms	14
Comprehensive review of modeling methods for gene regulatory networks. Includes a classification of methods	15
Focus on Boolean networks, clustering of gene-expression data, and inference of models from gene expression data. Other modeling formalisms are also covered	16
Detailed description of the model and modeling method developed by the authors to study regulation of the sea urchin endo16 gene.	17
Mammalian example. Three models drawn from the study of immune receptor signaling. Focus on the biology and biological insights from the models	18

20.7.2 Recommended Detailed Reviews

Comments	Refs.
Broad perspective of many issues in systems biology with a philosophical bent	19
General discussion of protein–protein interactions focused on data-production methods and data quality. Some discussion of gene expression, gene-deletion studies, and data analysis	20
Overview of protein–protein interactions. Covers the basic biology, data-production methods, data-quality issues, and data analysis	21
Data production methods for protein–protein interactions with some discussion of computational methods for predicting interactions	22
Modeling methods for static and dynamic models. Includes a discussion of graph properties. Some description of data integration, but not data fusion	23
Introduction to gene (transcriptional) regulatory networks. Combines discussion of modeling methods and evolutionary considerations	24
Formalisms for continuous models. Proposes evaluation criteria for comparing formalisms and evaluates five methods using simulated data	25
Detailed explanation of the authors' approach to inference of gene regulatory networks from microarray data	26
Case study of modeling a gene regulatory network for specification of endomesoderm in sea urchin	27
Methods for predicting gene function from sequence	28
Experimental design considerations for inference of gene regulatory networks	29
Discussion of graph properties and possible biological significance of these properties.	30
Interactions among signaling domains in yeast	31

20.7.3 Recommended High-level Reviews

Comments	Refs.
Perspectives	32

Comments	*Refs.*
Perspectives with some detailed examples	1, 33
Modeling goals and issues. Compares high level functional models with low level mechanistic models	34
Overview of protein–protein interactions touching on data production, data quality, and data analysis. Discusses computational methods for predicting interactions, and for extrapolating interactions from one species to another	35
Very high-level overview of protein–protein interactions	36
Analysis of protein–protein interaction data, including computational prediction of interactions. Contains a list of protein–protein interaction databases	37
Computational methods for predicting interactions	38
Data fusion	39, 40
Modeling methods for gene regulatory networks	41
Modeling methods for gene regulatory networks. Use of ChIP-chip data and computational prediction of transcription factor binding sites	42
Modeling methods for gene regulatory networks. Inference of models from time series microarray data. Inference of gene function from models	43
Computational methods for predicting transcription factor binding sites. Also discusses ChIP-chip	44
Review of ChIP-chip data production method with some examples of how data can be used	45
Many topics pertaining to gene regulatory networks. Focus on empirical and computational approaches to transcription factor binding sites and integration with protein–protein interaction data	46
Review of graph properties and dynamic modeling	47

References

1 Ideker, T., T. Galitski, and L. Hood, A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet, 2001. 2:343–372

2 Ideker, T., et al., Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science, 2001. 292(5518):929–934

3 Shannon, P., et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 2003. 13(11):2498–2504

4 Vidal, M., A biological atlas of functional maps. Cell, 2001. 104(3):333–339

5 Kitano, H., Systems biology: a brief overview. Science, 2002. 295(5560):1662–1664

6 Hood, L., Systems biology: integrating technology, biology, and computation. Mech Ageing Dev, 2003. 124(1):9–16

7 Aitchison, J.D. and T. Galitski, Inventories to insights. J Cell Biol, 2003. 161(3):465–469

8 Xia, Y., et al., Analyzing cellular biochemistry in terms of molecular networks. Annu Rev Biochem, 2004. 73:1051–1087

9 Bader, G.D., et al., Functional genomics and proteomics: charting a multidimensional map of the yeast cell. Trends Cell Biol, 2003. 13(7):344–356

10 Grunenfelder, B. and E.A. Winzeler, Treasures and traps in genome-wide data sets: case examples from yeast. Nat Rev Genet, 2002. 3(9):653–661

11 Ge, H., A.J. Walhout, and M. Vidal, Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet, 2003. 19(10):551–560

12 Barabasi, A.L. and Z.N. Oltvai, Network biology: understanding the cell's functional organization. Nat Rev Genet, 2004. 5(2):101–113

13 Pawson, T. and P. Nash, Assembly of cell regulatory systems through protein interaction domains. Science, 2003. 300(5618):445–452

14 de Jong, H., Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol, 2002. 9(1):67–103

15 van Someren, E.P., et al., Genetic network modeling. Pharmacogenomics, 2002. 3(4):507–525

16 D'Haeseleer, P., S. Liang, and R. Somogyi, Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics, 2000. 16(8):707–726

17 Bolouri, H. and E.H. Davidson, Modeling DNA sequence-based cis-regulatory gene networks. Dev Biol, 2002. 246(1):2–13

18 Goldstein, B., J.R. Faeder, and W.S. Hlavacek, Mathematical and computational models of immune-receptor signalling. Nat Rev Immunol, 2004. 4(6):445–456

19 Huang, S., Back to the biology in systems biology: what can we learn from biomolecular networks? Brief Funct Genomic Proteomic, 2004. 2(4):279–297

20 Legrain, P., J. Wojcik, and J.M. Gauthier, Protein–protein interaction maps: a lead towards cellular functions. Trends Genet, 2001. 17(6):346–352

21 Chen, Y. and D. Xu, Computational analyses of high-throughput protein–protein interaction data. Curr Protein Pept Sci, 2003. 4(3):159–181

22 Cho, S., et al., Protein–protein interaction networks: from interactions to networks. J Biochem Mol Biol, 2004. 37(1):45–52

23 Sun, N. and H. Zhao, Genomic approaches in dissecting complex biological pathways. Pharmacogenomics, 2004. 5(2):163–179

24 Babu, M.M., et al., Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol, 2004. 14(3):283–291

25 Wessels, L.F., E.P. van Someren, and M.J. Reinders, A comparison of genetic network models. Pac Symp Biocomput, 2001:508–519

26 van Someren, E.P., L.F. Wessels, and M.J. Reinders, Linear modeling of genetic networks from experimental data. Proc Int Conf Intell Syst Mol Biol, 2000. 8:355–366

27 Davidson, E.H., D.R. McClay, and L. Hood, Regulatory gene networks and the properties of the developmental process. Proc Natl Acad Sci U S A, 2003. 100(4):1475–1480

28 Gabaldon, T. and M.A. Huynen, Prediction of protein function and pathways in the genome era. Cell Mol Life Sci, 2004. 61(7-8):930–944

29 Stark, J., et al., Reconstructing gene networks: what are the limits? Biochem Soc Trans, 2003. 31(Pt 6):1519–1525

30 Alm, E. and A.P. Arkin, Biological networks. Curr Opin Struct Biol, 2003. 13(2):193–202

31 Yu, J.W. and M.A. Lemmon, Genome-wide analysis of signaling domain function. Curr Opin Chem Biol, 2003. 7(1):103–109

32 Hood, L., Leroy Hood expounds the principles, practice and future of systems biology. Drug Discov Today, 2003. 8(10):436–438

33 Herrgard, M.J., M.W. Covert, and B.O. Palsson, Reconstruction of microbial transcriptional regulatory networks. Curr Opin Biotechnol, 2004. 15(1):70–77

34 Ideker, T. and D. Lauffenburger, Building with a scaffold: emerging strategies for high- to low-level cellular modeling. Trends Biotechnol, 2003. 21(6):255–262

35 Bork, P., et al., Protein interaction networks from yeast to human. Curr Opin Struct Biol, 2004. 14(3):292–299

36 Tucker, C.L., J.F. Gera, and P. Uetz, Towards an understanding of complex protein networks. Trends Cell Biol, 2001. 11(3):102–106

37 Salwinski, L. and D. Eisenberg, Computational methods of analysis of protein–protein interactions. Curr Opin Struct Biol, 2003. 13(3):377–382

38 Valencia, A. and F. Pazos, Computational methods for the prediction of protein interactions. Curr Opin Struct Biol, 2002. 12(3):368–373

39 Gerstein, M., N. Lan, and R. Jansen, Proteomics. Integrating interactomes. Science, 2002. 295(5553):284–287

40 Wei, G.H., D.P. Liu, and C.C. Liang, Charting gene regulatory networks: strategies, challenges and perspectives. Biochem J, 2004. 381(Pt 1):1–12

41 Brazhnik, P., A. de la Fuente, and P. Mendes, Gene networks: how to put the function in genomics. Trends Biotechnol, 2002. 20(11):467–472

42 Futcher, B., Transcriptional regulatory networks and the yeast cell cycle. Curr Opin Cell Biol, 2002. 14(6):676–683

43 Dewey, T.G., From microarrays to networks: mining expression time series. Drug Discov Today, 2002. 7(20 Suppl):S170–S175

44 Li, H. and W. Wang, Dissecting the transcription networks of a cell using computational genomics. Curr Opin Genet Dev, 2003. 13(6):611–616

45 Wyrick, J.J. and R.A. Young, Deciphering gene expression regulatory networks. Curr Opin Genet Dev, 2002. 12(2):130–136

46 Pennacchio, L.A. and E.M. Rubin, Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet, 2001. 2(2):100–109

47 Monk, N.A., Unravelling Nature's networks. Biochem Soc Trans, 2003. 31(Pt 6):1457–1461

Part IV
Ethical, Legal and Social Issues

21 Ethical Aspects of Genome Research and Banking

Bartha Maria Knoppers and Clémentine Sallée

21.1 Introduction

Genetic research is increasingly used to cover a wide range of research activity. This activity extends from classical research into diseases following Mendelian patterns of inheritance, to the search for genetic risk factors in common diseases, to the more recent interest in pharmacogenomics and, finally, to the actual need for studies of normal genetic variation across entire populations. This all-encompassing nature of the term genetic research would not be so problematic were it not for the fact that corresponding distinctions (if necessary) cannot be found in ethical norms applied to evaluate such research. To address this issue we need to understand the ethical concerns raised by the different types of genetic research. Beginning with a overview of the types of genetic research (Sect. 21.2), we then proceed to an introduction to the ethics norms of research in general (Sect. 21.3), before analyzing their further elaboration in genetic research (Sect. 21.4). Particular attention will then be paid to the problems raised by DNA banking (Sect. 21.5), with the conclusion focusing on the issue of ownership of the samples in an increasingly commercial environment (Sect. 21.6). Finally, the term "human genetic research databases" will be used to cover all systematic collections of stored tissue samples used in genetic research whether obtained during medical care or specifically for research.

21.2 Types of Genetic Research

The Human Genome Project, which identified 3 billion DNA base pairs, offered scientists an invaluable source of genetic information – the nearly complete sequence map of the human genome. Further projects such as the SNP consortium, aiming to identify single-nucleotide polymorphisms and the International HapMap consortium, which seeks to draw a haplotype map of the human genome (ancestral blocks of SNPs), are furthering researchers' understanding of genetics and genomics. The parallel development of biotechnology and bio-informatics has enabled access to advanced tools for data storage and analysis. As a result, the focus of medicine and genetic research has shifted from diagnosis (symptomatic medicine with the study of

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

single-gene disorders and their manifestation), to prevention and individualized medicine (asymptomatic medicine and pharmacogenomics). This change of emphasis has been accompanied by increased reliance on human genetic research databases, diverse not only in their nature but also in their scale and duration (e.g. Refs. [1, 2]).

Incredible progress has been achieved in our ability to discover and develop diagnostic tests for hereditary, single-gene disorders, and the possibility of predicting morbidity and mortality. The same progress has not been made in the treatment of these conditions. These conditions are, however, prime candidates for gene therapy research. Often inherited not only through families but also through racial and ethnic lines, these latter features, and the quasi-certainty of expression have led to the development of ethical guidelines and legislation sensitive to both potential discrimination and to the possibility of stigmatization by association Ref. [3] (Sect. 21.4).

Understanding the role of genetic factors in common conditions such as hypertension, cancer, and diabetes is more complex. Other than perhaps rare forms of these conditions that follow familial patterns, their expression is often determined by the interplay of environmental, socio-economic, cultural, and other influences. This poses interdisciplinary challenges to reviews of ethics, to say nothing of determining the appropriateness of legislation in this area.

Pharmacogenomics is seeking to understand the role of genetic variation (polymorphisms) in individual response (e.g., toxicity, efficacy, dosage, etc.) and requires the expansion of epidemiology studies to entire populations (whether ill, at-risk, or not) to enable understanding of genetic diversity (e.g. Ref. [4]).

Population genetics research holds much promise. It poses new ethical challenges and requires frameworks adapted to its unique scale and purpose. Interestingly, most population studies of genetic variation do not require identifying medical information but rather seek to use coded, double-coded, or anonymized DNA samples and associated data (Sect. 21.5).

Across this spectrum then, from certainty, to probabilistic percentages in common diseases, to individual susceptibility, to the coded or anonymized sample, the possibility of applying uniform ethical criteria is unlikely. The same difficulties might not be present, however, in the application of the larger ethical framework governing biomedical research generally.

21.3 Research Ethics

The last few years have seen the adoption of numerous international, regional, national, and professional codes which govern, make recommendations, or draw guidelines for biomedical research generally, or for more specific areas such as health or genetic research and genomic databases. Reports by national ethics committees questioning the need to establish, regulate and govern national population human genetic research databases have also been published – see, e.g., Refs. [5–8] (International), Ref. [9] (Regional), and Refs. [10–12] (National). This proliferation and specialization of laws and policies have posed international, regional and national challenges to possible harmonization.

While internationally, some consensus has been achieved on the broad general ethical principles pertaining to biomedical and genetic research, the multiplicity of standards and rules and the variability of the vocabulary used to describe the same reality is detrimental to research workers. This is even more worrisome in an era of height-

ened data and knowledge sharing and in which there is a desperate need for increased international collaboration. Harmonization is a priority for the international circulation of data (access and transfer) and cooperation between countries, researchers and/or clinicians, as well as international organizations. The Council for International Organizations of Medical Sciences states that "the challenge to international research ethics is to apply universal ethical principles to biomedical research in a multicultural world with a multiplicity of health-care systems and considerable variation in standards of heath care" (Ref. [5], Introduction). Likewise, the Secretary-General of the United Nations Economic and Social Council affirms that:

> "... despite the existence of numerous declarations, guiding principles and codes dealing with the issue of genetic data, the changing conditions of genetic research call for the establishment of an international instrument that would enable States to agree on ethical principles, which they would then have to transpose into their legislation (Ref. [13], para. 44, p. 11)."

Nationally, countries are submerged in a myriad of rules with various systems and approaches existing in parallel. This creates confusion, overlap, conflicts, or even areas lacking regulation. There is no comprehensive or coherent legal and/or ethical framework regulating tissue or genetic research (e.g. Refs. [2, 14]). For instance, in France the National Consultative Ethics Committee for Health and Life Science notes that:

> "... neither the collection of elements, tissues, cells, etc. of human origin, nor the study of the genetic characteristics of an individual, nor the establishment of computerized files, nor the processing of the resulting information, are unregulated activities, on the contrary, several systems co-exist so that the problems are approached from different angles which ignore each other (Ref. [1] (France))."

The US National Bioethics Advisory Commission has also emphasized this state of disorganization [15]. Indeed, in the US, federal regulations (privacy legislation, regulation by the Food and Drug Administration, Common Law rules), and state laws all find application (e.g. Ref. [14]). Harmonization on both the national and international arena is thus considered a fundamental issue for national ethics committees (see, e.g., The 3rd Global summit of National Bioethics Commissions, Imperial College, London, UK, Sept. 20–21, 2000; Ref. [1] (France)). They are, furthermore, calling for the implementation of a "new regulatory framework" addressing more specifically the issue of tissue banking or population genetic research databases – see, e.g., Refs. [11] (Israel) and [16] (Singapore).

Globalization, the explosion of new technologies, the North–South divide, sensitivity to differing cultural and religious worldviews and to the lessons learned from the biodiversity debate make such harmonization difficult but not impossible. The real test might well be that of ensuring that not only the public sector but also the private sector (which is the largest source of funding), abide by a future international approach. The other challenge relates to an endemic problem, that of proper on-going oversight.

It is interesting to note that there is a growing recognition of the need to involve the public and professionals in the drafting of harmonized ethical frameworks for research and banking, the latter having shown their concern and interest with the adoption of policies (e.g. Ref. [14]). For example, the Singapore Bioethics Advisory Committee recommends that:

> " ... given the background of a rapidly shifting and evolving body of ethics, legal rules and opinion governing human tissue research and banking in the leading scientific

> jurisdictions, we recommend that a continuing professional and public dialog be initiated towards: setting the principles which should guide the conduct of tissue banking against the background of evolving consensus on these principles in the leading scientific jurisdictions ... This dialog should be undertaken with a view towards ensuring the harmonization of our laws with accepted international best practice and consensus on relevant legal doctrines and principles such as being developed in the leading jurisdiction around the world." (Ref. [16], rec. 13.6 and 13.7; see, also, Ref. [17], art. 28, Ref. [18], rec. 9, Ref. [19], sect. 3.5, and Ref. [10], rec. 1)."

Following the adoption of the Nuremberg Code (1946–1949) [20], of the Helsinki Declaration [21], and more recently of the CIOMS guidelines [5], the main tenets of research ethics are both integrated into the biomedical world and yet evolving. The most common elements include respect for privacy and autonomy through the process of informed consent and choice, the right to withdraw, and the protection of the vulnerable.

The last decade has seen the emergence of new issues and additional elements such as: community consent, right not to know, commercialization, statutory regulation of clinical trials, benefit-sharing, inclusiveness, equitable access to research trials and benefits, and free flow of information. There is also a much greater specificity in that particular areas or groups of persons are singled out, for example those suffering from mental disorders, HIV/AIDS, fetuses, deceased individuals, persons deprived of liberty, and the disabled. Moreover, as already mentioned, frameworks are being developed for particular areas such as organ transplantation, reproductive technologies, stem cells, tissue banking, genetic or genomic databases, etc. (e.g. Refs. [14, 22]).

The adoption of the European Convention on Biomedicine and Human Rights [17] illustrates the difficulty of finding common principles and positions when technologies are already well-entrenched and different countries have adopted legislation. For example, no agreement could initially be reached in the Convention on embryo research, an area where guidance is required now stem cell and therapeutic cloning techniques are promising new breakthroughs. Only in 2001 was the Additional Protocol on the Prohibition of Cloning Human Beings [23] finally added to the Convention. On September 7, 2000, the European Parliament narrowly passed a resolution (237 vs. 230 votes with 43 abstentions) condemning the deliberate creation of embryos for therapeutic cloning [24]. The same difficulty finding consensus will no doubt hold true in the actual elaboration of specific protocols pursuant to the Convention or European resolution on, for example, stem-cell research, or the attempt to adopt a council decision amending decision 2002/834/EC on the specific program for research, technological, development and demonstration: "Integrating and strengthening the European research area" [25].

The Convention [17] is, however, notable in its broadening of the inclusion criteria governing incompetent adults and children. Rather than excluding them from biomedical research in the absence of direct benefit, the Convention would permit inclusion with the consent of the legal representative even if the benefits were only indirect, that is, for persons of the same age and condition (art.17). The same principle was upheld by the Council for International Organizations of Medical Sciences provided the research is of minimal risk and is scientifically and ethically justified; such justification must be overriding when the research presents a risk that is more than minimal (Ref. [5], guideline 9). This was further upheld by the World Health Organ-

ization when addressing the inclusion of vulnerable individuals in genetic databases (Ref. [19], para. 4.3). In short, the evolution of biomedical ethics in human research in general has received wider acceptance. Is the same true for the field of bioethics and genetics [22]?

21.4 "Genethics"

At the international level, UNESCO adopted the Universal Declaration on the Human Genome and Human Rights in 1997 [26]. The Declaration is prospective in nature. It embraces the concepts of human dignity and the diversity of the genome as the common heritage of humanity, of noncommodification, of the need for international solidarity, and of concern over technologies such as germ-line interventions that could affect future generations (art. 24). It specifically prohibits human reproductive cloning (art. 11).

The proliferation of genetic/genomic databases over the last decade, and the development of genetic research using such resources have prompted international organizations to implement a set of guiding principles with regard to the collection, processing, use, and storage of genetic information [7], and/or the management of genetic [19] or genomic databases [8]. These texts reaffirm the principles embodied in the Universal Declaration on the Human Genome and Human Rights. They assert the communal value of genetic data or databases and insist on the fundamental nature of the principles of human dignity, solidarity, equality, justice and on the need for research to be carried out transparently and responsibly. UNESCO's International Declaration on Human Genetic Data aims to "... ensure the respect of human dignity and protection of human rights and fundamental freedoms in the collection, processing, use and storage of human genetic data, human proteomic data and of the biological samples from which they are derived ..." (art. 1a).

The WHO embraces "... the pursuit of human well-being, the quality of human dignity, including fundamental human rights and the principle of nondiscrimination, the principle of respect for persons, including the imperatives of beneficence and nonmaleficence and the principle of respect for individual autonomy ..." (Sect. 2.2) as primary guiding values; it also specifies that "... accepted principles regarding personal information and human rights apply also to genetic information." It acknowledges, however, the western philosophical undertone of the report and the principles endorsed, and recommends the examination of other perspectives.

Both these documents, while warning against the risk of discrimination and stigmatization, insist on the need to avoid genetic determinism or exceptionalism which could ensue from the current scientific and public focus on genetics and genomics (reducing individuals to their genetic characteristics and distinguishing genetic information from other health information even when not justified, thus creating "genetic" specificity) (Ref. [7], arts. 3 and 7(a) and Ref. [19], Preface; see also Ref. [4], art. 33). Finally, the Human Genome Organization, promoting the free flow of data and fair and equal sharing of research benefits, holds human genomic databases to be global public goods, i.e. goods "... whose scope extends worldwide, are enjoyable by all with no groups excluded, and, when consumed by one individual are not depleted for others ..." [8]. These international instruments then, come at the beginning of a technology and hopefully will serve to guide national

approaches, thus ensuring a minimum of harmonization.

Also anticipatory in nature and 10 years in the making, the 1998 European Directive on the Legal Protection of Biotechnological Inventions [27], not only a clarifies (if not ratifies) existing trends but also innovates. The Directive reaffirms the nonpatentability of human genes in their natural state and, under the umbrella of public policy (an ethical filter also found in the European Patent Convention), prohibits techniques such as human cloning and germ-line intervention (art. 5(2)). The preamble ("recital"), while not having legal force, requires that a patent application for an invention using human biological material must be "... from a person who has had the opportunity of expressing free and informed consent thereto, in accordance with national law ..." (para. 26). This means that, at a minimum, participants in genetic research and banking must be notified of the possibility of eventual commercialization. In the absence of the Directive's implementation into "national law", however, its impact will be weakened. (see, also, Ref. [18]).

It is interesting to note that both international and regional instruments are strengthening barriers to access by third parties (e.g. insurers, employers, educational institutions). This is achieved either implicitly by limiting the purposes for which genetic information and/or tests can be used or with the adoption of express provisions circumscribing access to information (biological material, data derived, and results). Notable in this regard is the European Convention on Human Rights and Biomedicine [17] mentioned earlier, which, in confining genetic testing to health purposes (art. 12), effectively limits requests for testing by insurers and employers (see also, Data Protection Working Party, Ref. [28]. p.7, which would enable genetic tests to be used by insurers and employers only in the most exceptional situations). Internationally, the collection, processing, use and storage of human genetic and proteomic data can only be undertaken for purposes of diagnosis and health care, research (medical and scientific), forensic medicine, and investigations subject to certain exceptions, and "... any other purpose consistent with the Universal Declaration on the Human Genome and Human Rights and the international law of human rights ..." (Ref. [7], arts. 1 and 5). Furthermore, access by third parties to identifiable genetic, proteomic information and samples is proscribed except when consent of the individual has been secured or where provided for by national laws within the limits of important public interests (art. 14b). Similarly the World Health Organization recommends that parties "... concerned with their own financial gain ..." not be allowed access to genetic data or databases (Ref. [19] Sect. 6.1).

A significant development is the creation of a right not to know alongside the more traditional right to know one's own genetic information. Adopted by the Convention on Human Rights and Biomedicine under article 10(2), it has been ascertained in later international and regional documents (see, e.g., internationally, Ref. [7] art. 10 and Ref. [19] rec. 16; and, regionally, Ref. [9] art. 18(iii) and Ref. [28] arts. 12(3) and 12(4)). This right to decide whether or not to be informed is not absolute. It is subject to the data being identifiable, to the limitations laid down by law in the interests of public health, and to the results being individual and not just general (on return of results see, e.g., Ref. [4] art. 34, Ref. [19] rec. 8 (Quebec), and Ref. [10] rec. 6) The right not to know, an expression of the individual right to autonomy, remains uncertain in scope because it implies the providing of information in order for consent to be in-

formed as well as limits in such information not to impact on the decision of the individual not to be informed.

As emphasized by the World Health Organization "... in circumstances where there is no evidence of what an individual would want to know, it is not possible to seek to advance their autonomy by asking them if they would wish to know, for to do so is to indicate that there is, indeed, something to know, and thereby any possible harm will have been caused." It thus proposes a list of factors to be taken into account and weighed before an individual is approached, one of which is the vague" ... question of how the individual might be affected if subjected to unwarranted information, and whether the individual has expressed any views on receiving information of this kind" (Ref. [19], rec. 16).

The right not to be informed becomes more controversial when applicable to genetic relatives, as established by UNESCO's Declaration on Genetic Data (art. 10), as they have not voluntarily and in an informed way decided to be subjected to testing or to participate in research. The manner in which this right will be implemented bears examination.

Correlative to the recognition of a right not to know is the development of a new exception for professional disclosure to at-risk family members of serious preventable or treatable conditions where the patient refuses to do so. This emerging duty to warn, subject, in critical situations, to ethical approval, has been adopted by many international instruments. This is the position of the 1998 Proposed International Guidelines of WHO [29], of the 1998 HUGO Statement on DNA Sampling: Control and Access [30] of the ESHG [31], of the 2003 UNESCO Declaration on Genetic Data [7], and of the 2003 WHO Genetic Databases report and recommendations [19]. Thus, this ethical duty is not only an option for clinicians but is also applicable in the research or banking context, and thus creates an ongoing ethical obligation for the researcher/banker as new tests become available. The existence of an ethical duty to warn is, however, sometimes opposed nationally by proponents of a strict notion of confidentiality in the patient/subject–clinician/researcher relationship (see infra. 5.3 (France) [1]).

Finally, another change in international "genethics" is the attempt to move away from traditional, categorical, wholesale prohibitions in the area of cloning and germline therapy. Although the International Bioethics Committee of UNESCO in its penultimate draft had agreed to keep the Universal Declaration on the Human Genome and Human Rights [17] free from mention of any specific technology, the aim being to guarantee its viability over time and its universality and to strengthen the impact of concepts such as human diversity and dignity, prohibitions were added by representatives of governments as a condition of approval. Indeed, governmental representatives who were convened to approve the Committee's final draft sought (political?) refuge in inserting "technique-specific" prohibitions in the Declaration with regard to human cloning and germ-line therapy as mentioned earlier. Currently, an international declaration on human cloning is being discussed at the United Nations. The main point of contention is its scope, that is, whether it is to be limited to reproductive cloning or include therapeutic cloning. It bears noting that the WHO in both its 1998 Proposed International Guidelines [29], its 1999 Draft Guidelines on Bioethics [32], and its resolution on Ethical, Scientific and Social Implications of Cloning in Human Health [33] distinguishes between the different types of cloning. Both WHO and

HUGO [30] prohibit human reproductive cloning but encourage relevant research in the field of therapeutic cloning and stem-cell research. This is instructive in the banking context where, as we shall see, formerly, wholesale proclamations about DNA as "person" or as "property" have ultimately proved secondary to the need to ensure personal control, irrespective of the legal qualification and without impact as regards commercialization.

21.5 DNA Banking

The last fifteen years have seen tremendous upheaval and uncertainty in the world of DNA banking and research. Indeed, 1995 saw the hitherto unfettered access by researchers to archived samples come to a halt with the report of an NIH study group on informed consent for genetic research on stored tissue samples suggesting that proof of consent to research was required even for samples already stored during routine medical care [34]. Although, in general, the ethical and legal norms governing banking had been moving toward a more informed choice approach with options in the case of samples provided in the research context *per se*, the implementation of this approach would effectively have halted the largest "source" of DNA samples for genetic research to say nothing of epidemiological or public health research (even if the latter wished to use only anonymized samples) (e.g. Refs. [35, 36]). This conservative position was followed by a myriad of contradictory positions around the world.

Five years later, in May 2000, the UK's Royal College of Physicians Committee on Ethical Issues in Medicine published its recommendations on Research based on Archived Information and Samples [37] and the circle seemed closed. Indeed, the Committee did not consider unethical the secondary "unconsented-to" use for research of biological samples obtained during clinical care (e.g. "left-over"), post-mortem examination, or research, provided certain conditions were met and anonymization occurred at the earliest stage possible; the minimum level of anonymization being that which precludes identification of the individual from the research [37]. This report, even though to be welcomed for adopting a more liberal approach to research, could, however, be regarded today as inadequate when applied to human genetic research databases or "biobanks", because anonymization might not always be welcome. Biobanks differ from traditional hypothesis-driven genetic research using stored samples and data, notably, in their fundamental nature, that of "resources" for researchers. As a consequence, it is impossible to encompass all possible uses when they are established, because they aim to store samples and data indefinitely.

In 2004, the German National Ethics Council published its opinion on "Biobanks on Research" [38], closing the discussion on the ethical principles that should govern the establishment of biobanks that started with the adoption of a joint statement by the French and German National Ethics Committees in 2003 [12]. In that statement, both Committees concluded that "... despite some differences, there is a need in both France and Germany, to elaborate a new regulatory framework covering collection, conservation, processing, and utilization of the elements and data assembled in biobanks, and the development of research protection of individuals ... "

Balancing patients rights, through respect for the most fundamental ethical principles, and freedom of research, through the adoption of a practical and sensible ap-

proach, the opinion of the German National Ethics Council goes further than existing norms pertaining to biobanking. Recognizing the need for archived samples (obtained for diagnosis and treatment) to remain available for further use while respecting consent requirements, they hold that a "form-based" broad consent should be obtained at the time of collection (regulatory proposal 2). They further conclude that consent requirements can be waived when samples and data are anonymized, or even coded ("pseudonymized") provided researchers do not have access to the code, the data protection officer being responsible for ensuring respect for privacy requirements (regulatory proposal 3). According to German data-protection legislation, even where no "precautionary consent" was secured, consent can be waived for research on data and samples in a personalized form when donors' interests are outweighed by the scientific importance of the research and the research cannot proceed otherwise or can proceed only at too high a cost, and disproportionate efforts. However, the consent of the individual should be obtained if possible (regulatory proposal 4). Finally, for biobanks created for research purposes, consent can be general as to the type of research and length of storage, the information provided to the donors being limited to "... personal risks to the donor arising directly in connection with the use of samples and data in biobanks ..." and not extending to more general risks such as those of discrimination or stigmatization (regulatory proposals 5, 6, 9, and 12). All research projects intending to use a biobank ought, however, to receive prior ethical approval (regulatory proposal 17).

Where does this position stand relative to international norms or that of other countries? To answer that question we will examine the varying responses in the period 1995–2004 with respect to samples already archived which were obtained during medical care, and then samples obtained for research but where other research is now proposed.

It should be mentioned at the outset that perhaps more confusing than the plethora of contradictory positions is that of the terminology used. Only terms such as "identified", "nominative", or "personally identified" are understandable by all. In contrast, "identifiable", "traceable", or "pseudonymized", "proportional", or "reasonable" anonymity are used interchangeably with the terms "coded" or "double-coded", and the term "anonymous" (i.e. never had any identifiers, such as specimens found on archeological sites) is often confused with "anonymized."

For the purpose of clarity, we use the term "anonymized" (e.g., originally identified or coded/identifiable/traceable/pseudonymized to include clinical or demographic data but now stripped except for clinical or demographic data) and the term "coded" or "double-coded" (e.g., identifiable only by breaking the unique code or the two unique codes given the sample). We will examine international (Sect. 21.5.1) and regional (Sect. 21.5.2) positions on "medical care" samples and then on research samples before turning to the positions of particular countries (Sect. 21.5.3).

21.5.1 International

Prior to the beginning of this century, with the notable exception of the Human Genome Organization's 1998 Statement on DNA Sampling: Control and Access [30], international statements and guidelines on the ethics of genetic research failed to address the specific issue of archived samples originating from medical care, the context

of medical care being largely left to individual countries. This was regrettable for many reasons, the major one being the extreme difficulty, if not impossibility, of fulfilling the ethical obligation of international collaboration due to the lack of international guidance and harmonization. Acknowledging the value of such archived material and information, the impracticability in many cases of obtaining renewed consent, and the harm that could be caused to individuals who could be "... faced with an unwarranted approach with information that they might not wish to know ..." (Ref. [19], sect. 4.4), international instruments are beginning to include limited exceptions to the traditional need for renewed informed consent. At a minimum, ethical committees can dispense researchers with consent requirements for studies on anonymized samples and data.

While in its 1998 Proposed International Guidelines the World Health Organization did not take a position on left-over or "abandoned" samples except to say that "... specimens that could be useful to families in the future should be saved and should be available ..." (Ref. [29], Tab. 10, guideline 10), it specifically addressed the issue of archived material in its 2003 report on genetic databases. The latter enables the use of such material (e.g. "... pre-existing health records, specific health disorder databases or physical samples that have been retained ..."), when anonymized and provided no future identification is possible of the "sample's source" notably through research results (Ref. [19], rec. 10).

HUGO [30], CIOMS [5], and UNESCO [7] have adopted an even less restrictive approach, enabling stored samples and data to be used not only in an anonymized but also in a coded form, provided certain conditions are met. For example, the Human Genome Organization Ethics Committee held in 1998 that:

> "Routine samples, obtained during medical care and stored may be used for research if: there is a general notification of such a policy, the patient has not objected, and the samples obtained during medical care and stored before notification of such a policy may be used for research if the sample has been anonymized prior to use (Ref. [30], rec. 2)."

In the same way, the Council for International Organizations of Medical Sciences, in its 2002 ethical guidelines for biomedical research, states that consent requirements can be waived if individuals are notified and their confidentiality or anonymity is protected. However, it holds that "... waiver of informed consent is to be regarded as uncommon and exceptional, and must in all cases be approved by an ethical review committee ..." (Ref. [5], guideline 4). It further decides that such a committee can waive some or all the requirements of informed consent only for studies that pose minimal risks, are expected to yield significant benefits, and could not realistically or reasonably take place were consent requirements to be imposed (impracticability). The guidelines consider that prior opting out must be respected, except in situations of public emergencies, and that reluctance or refusal to participate is not sufficient to establish impracticability (Commentary under guideline 4).

Finally, UNESCO's 2003 International Declaration on Human Genetic Data leaves to national legislation the task of implementing the specific conditions pertaining to the secondary use of samples and data collected in the course of medical care (Ref. [7], art.16). Nevertheless, it holds that research can be undertaken without consent if the information is anonymized ("... irretrievably unlinked to an identifiable individual ...") or when consent cannot be obtained, provided proper ethical oversight occurs (art. 16b). Waiver of informed consent

might also be warranted when an important public interest is at stake, assuming fundamental human rights are respected (art. 16a).

With regard to samples and data collected for a specific research project that are to be used subsequently for other research purposes, similar waivers are contemplated (e.g. Ref. [30] rec. 3, and Ref. [7] art. 16). The Council for International Organizations of Medical Sciences specifies that subsequent research be circumscribed by the original consent, and that any conditions specified in that initial assent apply equally to secondary uses (Ref. [5] Commentary on guideline 4). It affirms the critical need for anticipation of future uses when samples and data are first collected. Thus, investigators should, during the original consent process, inform potential participants about any foreseen secondary uses, privacy protection, or destruction procedures that will be implemented, and of their rights to request destruction of any material or information they deem sensitive, or to opt out (Commentary on guideline 4; see also, e.g., for genetic databases, Ref. [19] rec. 6). As for biological samples and data collected as part of clinical care, however, elements of informed consent can be waived by an ethical review committee in exceptional circumstances (Guideline 4 and its attached commentaries). UNESCO's 2003 Declaration on Human Genetic Data holds that, health emergencies excepted, secondary uses incompatible with the conditions set out in the initial consent form cannot proceed without a renewed consent (Ref. [7] art. 16).

Turning to consent procedures pertaining to the collection of new research samples, it is worth outlining the emergence of the notion of well-informed generalized or broad consent to future research, particularly adapted to large-scale longitudinal genetic studies on variation using human genetic research databases. Broad consent, sometimes called "authorization", has not, however, become the principle, but rather an acceptable exception for certain types of research, if it is justified both scientifically and ethically and certain criteria are met.

As early as 1998, and in stark opposition to the more conservative positions of the time, the World Health Organization's Proposed International Guidelines (Ref. [29] Tab. 10, guideline 10) maintained that "... a blanket informed consent that would allow use of a sample in future projects is the most efficient approach ..." (Tab.10). This was somewhat tempered by the assertion that "... genetic samples from individuals must be handled with respect, should be taken only after the consent is obtained, and, should be used only as stated in the document ..." (p. 4). In its 2003 report on genetic databases [19], the World Health Organization reaffirms the possibility for a limited "... blanket consent for future research ..." when data and samples are anonymized. Referring to large-scale population human genetic databases, it insists on the need for strong ethical justification to deviate from traditional consent requirements. A certain number of criteria must be satisfied with regard to the expected benefits, privacy protection, and education of the public (Ref. [19] rec. 14).

The Human Genome Organization defines informed consent for research as including "... notification of uses (actual or future), or opting out, or, in some cases, blanket consent ..." (Ref. [8] rec. 4). Finally, the Consortium on Pharmacogenetics is of the opinion that some flexibility should be allowed for in consent forms, consent to a "... range of related studies over time ..." being a reasonable policy under certain circumstances (Ref. [4] Research B).

In conclusion, the international efforts undertaken to differentiate between data

and samples that are anonymized, coded, or identified should be emphasized and welcomed. These distinctions affect the rules which apply to secondary uses of data and samples, publication of results, transfer and trans-border exchange of information, and rights to withdraw, to name but a few. Furthermore, UNESCO's 2003 Declaration on Human Genetic Data holds that "... human genetic data and human proteomic data should not be kept in a form which allows the data subject to be identified for any longer than is necessary for achieving the purposes for which they were collected or subsequently processed ..." (Ref. [7] art. 14e). The World Health Organization, acknowledging the value of anonymization with regard to participants' privacy, requires the anonymization process to be scrutinized by an ethics committee, a necessary intermediary to ensure its legitimacy and maintain adequate standards (Ref. [19] sect. 4.2. and rec. 7). Finally, the Consortium on Pharmacogenetics assesses the advantages and drawbacks of anonymization compared with coding or double-coding, stating its preference for the latter (Ref. [4] Research F).

Regrettably, however, as mentioned earlier, the terminology used to describe samples and data is rather confusing. The resulting complexity undermines possible harmonization, especially when new concepts such as reasonable or proportional anonymity are introduced by the WHO (Ref. [19] para. 4.2).

21.1.2 Regional

At the regional level, other than upholding the need for informed consent for all medical interventions including research, the Council of Europe's Convention on Human Rights and Biomedicine [17] provides little guidance on genetic research with regard to either archived samples left over after clinical care or research samples. Article 22 maintains that: "... when in the course of an intervention any part of a human body is removed, it might be stored and used for purposes other than that for which it was removed only if this is done in conformity with appropriate information and consent procedures." No definition of what is considered appropriate can be found in the Convention, because it is left to the discretion of national states.

Guidance can, however, be found in the European Society of Human Genetics (ESHG) 2001 Recommendations on Data Storage and DNA Banking [39] and in the Council of Europe Proposal for an Instrument on the Use of Archived Human Biological Materials in Biomedical Research [9].

These texts do not specifically refer to samples left-over from clinical care but rather to archived samples or existing collections in general. The ESHG recommendations are notable in that they distinguish existing collections based on the degree of identifiability of the samples and data and the length of storage [39]. The Society considers that consent requirement can be waived when samples are anonymized (rec. 9), and, provided it is approved by an ethics committee, in situations where the collection can be considered as abandoned ("old collections") (rec. 14). For collections of coded information, although, in principle, re-consent of participants for new studies is necessary, ethics review committees can waive the requirement for such consent when re-contact is impracticable and the study poses minimal risks (rec. 12). Finally, post-mortem uses of samples are subject to the donor's advance wishes (rec.13). In the absence of any known wishes, use of those samples should be regulated, a policy of unfettered use not being ethically justified (rec.13).

In contrast, the Council of Europe does not allow re-consent to be waived when stored biological clinical material and data are to be used subsequently for research (Ref. [9] art. 14). Consent can, however, be either implicit or explicit depending on the intrusiveness of the study and the previous directives of the donor (art. 16). Individuals enjoy the right to withhold their sample from certain future research uses of their sample and the right to consent to subsequent procedures (art. 15.1). Finally, post-mortem uses or uses of data from individuals unable to consent have to meet satisfactory information and consent measures (art. 17). However, the Steering Committee on Bioethics (CIOB) working party on Research on stored Human Biological Material held that this draft proposal did not apply to biobanks for resarch in the future but only to specific research projects (Ref. [64]). The Council of Europe and the ESHG both insist on the need to differentiate data according to their degree of identifiability. On the one hand, the Council of Europe proposal holds that the decision to use coded, identified or linked anonymized human biological material or personal data "... shall be justified by the needs of the research ..." (Ref. [9] art. 9). On the other hand, the European Society of Human Genetics explains that while anonymization is acceptable for "... sample and information sharing for research purposes with minimum risk ..." maintaining identifiability as protected by coding is valuable and advisable because it "... will permit more effective biomedical research and the possibility of re-contacting the subject when a therapeutic option becomes available ..." (Ref. [39] recs.10 and 11).

Only the ESHG document explicitly addresses the issue of consent for new collections. It retains the fundamental principles of ethics committee' oversight and informed consent for "... all types of DNA banking ..." (Ref. [39] rec. 3). While not mentioning blanket consent, it contemplates the possibility of obtaining consent for broad research uses, the consequence being that individuals need not be re-contacted although they are to be kept informed and are to be given the possibility to express their desire to withdraw (rec. 8). Although international documents urge groups, communities, and populations to become involved in the discussion surrounding the establishment of large-scale population human genetic research databases, the ESHG goes one step further requiring additional consent "... at a group level through its cultural appropriate authorities ..." for population studies (rec. 15). This latter consent bears examination. How it will be implemented in practice remains to be seen. Indeed, it might prove difficult to define what would amount to an adequate community or group consent, what amount of opposition would result in the study not being undertaken, and the consequences as far as other rights, notably that of withdrawal, are concerned.

21.5.3 National

Before 2000, most countries did not distinguish between clinical or archived research samples nor had positions on the issue of other uses. As a consequence, in the absence of a new and explicit consent, valuable archived material collected in the course of clinical care or research was not exploitable for purposes other than those outlined in the original protocol. Although today this distinction is not always made, several documents have addressed the issue – e.g. Refs. [40] (Australia) and [16] (Singapore).

We are, furthermore, witnessing a specialization of laws and ethical guidelines with the drafting of legal instruments ad-

dressing the specific issue of human genetic research databases or biobanks, and the growing recognition of the ethical validity of broad or general consent to genetic research or banking.

As a preliminary comment, as previously mentioned, it should be noted that national jurisdictions are insisting on the need to involve the public in any debate related to the establishment of populational or community-based human genetic research databases to ensure both transparency and acceptance of the project, and to address the concerns thus voiced – e.g. Refs. [41] (para. 2) and [42] (USA), Ref. [1] (France); Ref. [10] rec.1 (Quebec), and Ref. [16] rec.13.17 (Singapore). It is recommended that public involvement take the form of public consultation or engagement mechanisms. While not constituting "consent" *per se*, it is advocated that strong opposition could jeopardize a project's lifespan. What remains to be seen, however, is the amount of public objection that will prevent the establishment of a human genetic research database.

With regard to tissues primarily collected for therapeutic, diagnosis purposes or so called "left-over" tissue, the need for a new consent for further use in research is strongly emphasized, waiver of consent, as a principle, not being ethically justified. As outlined by the Singapore Bioethics Advisory Ethics Committee "... clinicians should not assume that tissues left over after diagnosis or research can be used for research without consent ..." (Ref. [16] rec. 13.13). The Australian Law Reform Commission further states that "... it is easier to argue that consent should be waived for research purposes than for research on tissue originally collected for other purposes (therapeutic or diagnostic). In the latter case ... research is an unrelated secondary purpose [not] within expectations of the individual concerned ..." (Ref. [40] para. 15.5). In stark contradiction with the principles mentioned, the Icelandic Act on Biobanks allows samples taken for clinical purposes to be stored in the absence of any objections from the donor – Ref. [43] art 3 (Iceland). This controversial presumed consent is dependent upon information being posted in medical institutions. However, similar provisions of the Health Sector Database Act have been held unconstitutional by the Icelandic Supreme Court and the existing decode biobank has obtained explicit consent from all the participants. Ref. [65]. Despite the established principle of informed consent, some documents do allow for the unconsented-to further use of left-over tissues for ethically approved research procedures if the information is anonymized (e.g. Ref. [44] rec. 5.9 (UK)) or when the intended use is incidental to the original purpose (e.g. Ref. [16] para. 8.8 (Singapore)). To avoid confusion when faced with archived tissue collected primarily for therapeutic or diagnostic purposes, it is also recommended that individuals be required to consent in two separate forms to the diagnostic or therapeutic act on the one hand, and to the storage of residual tissues and further use for research at the time of collection on the other hand – e.g. Ref. [44] rec. 6.2 (UK) and Ref. [16] para. 8.9 (Singapore).

In general, national jurisdictions are becoming less reluctant to allow for the possibility of deviating from the traditional principle of explicit re-consent for future use of samples collected for a specific research purpose. For instance, Australia's 1999 National Statement on Ethical Conduct in Research Involving Humans [45] "normally" calls for informed consent from donors of archived samples (princ. 15.7). Yet, the possibility of waiver by an ethics committee for the obtaining of another consent is foreseen in the context of research samples (princ. 15.6). Indeed:

" ... an HREC (Human Research Ethics Committee) may sometimes waive, with or without conditions, the requirement of consent. In determining whether consent may be waived or waived subject to conditions, an HREC may take into account:

- the nature of any existing consent relating to the collection and storage of the sample;
- the justification presented for seeking waiver of consent including the extent to which it is impossible or difficult or intrusive to obtain specific consent;
- the proposed arrangements to protect privacy including the extent to which it is possible to de-identify the sample;
- the extent to which the proposed research poses a risk to the privacy or well-being of the individual;
- whether the research proposal is an extension of, or closely related to, a previously approved research project;
- the possibility of commercial exploitation of derivatives of the samples; and
- the relevant statutory provisions." (princ. 15.8)

In 2003, the Australian Law Reform Commission approved of such an exception to the principle of informed consent, further recommending the HREC "... report annually to the Australian Health Ethics Committee (AHEC) with respect to human genetic research proposals for which waiver of consent has been granted under the National Statement ..." to ensure transparency and accountability (Ref. [40] rec. 15-1). It must be remembered that traditionally, waiver of informed consent for secondary use of archived samples and data for research were dependent on identifiability (anonymization), the impracticability of obtaining contemporary consent, the minimum risk of the contemplated study, and the approval of a research ethics committee (e.g. Ref. [46] p. 88 (The Netherlands)), or for public safety reasons. As underlined by the Australian Law Reform Commission and the Australian Health Ethics Committee [47], what amounts to "impracticability" needs further definition and explanation (Australian Law Reform Commission, Australian Health Ethics Committee, 2003, prop. 12-2).

Turning to consent to new collections for research and more particularly consent for inclusion in human genetic research databases, national jurisdictions seems to be moving toward recognition of broad consent procedures. As stated by the French National Consultative Ethics Committee:

> " ... these alterations to the notion of individual consent rests on the notion that the mass of information and connected data have in fact only acquired value for those participating because they are assembled and cross-matched for a great many people. They gradually constitute an asset which is detached from the person who has supplied an element of his/her body, the only value of which is the common use that progress has made possible ... "(Ref. [1] p. 5 (France)).

To take but a few examples, the Japanese Bioethics Committee of the Council for Science and Technology in its Fundamental Principles of Research on the Human Genome [48] mentions that:

> " ... if a participant consents to provide a research sample for a genome analysis in a particular research project and, at the same time, anticipates and consents to the use of the same sample in other genome analyses or related medical research, the research sample may be use for 'studies aimed at other purposes' ..." (princ. 8.1.a)

(see, also, (U.S.A.) Council of Regional Networks for Genetic Services, [49], which did not exclude blanket consent for samples obtained by the medical profession). The German Senate Commission on Genetic Research goes even further [50]. Supporting the need for less stringent consent procedures for epidemiological and genetic research, it considers that:

> " ... in principle ... decisions taken in a deliberate state of ignorance and uncertainty can be an expression of the donor's right to self-determination ..." (p. 46).

They conclude that:

> " ... consequently, no overriding objections can be raised to a consent phrased in general terms which does not specify all possible uses of specimens and data, or even to an all encompassing blanket authorization ..." (p. 46).

However, some jurisdictions remain faithful to traditional requirements and specific consent and re-consent must be obtained for any further use not contemplated in the original consent form (Ref. [51] Chapt. 3, Sect. 5 (Sweden)).

In short, on the national level, three positions typify the move away from the strict rule of requiring a new consent for other uses of samples. The first is that of requiring ethics review when foreseeing the possibility of either anonymizing or coding the sample without going back to the source provided there is only minimum risk and confidentiality is ensured (e.g. Ref. [15] rec. 9f (USA); Refs. [45] and [47] proposal 15-1 (Australia); and Ref. [52], 10.2 (UK)). The second requires ethics review but samples must always be anonymized (an option found in older documents: e.g. Ref. [46] (The Netherlands); Ref. [53] (USA); and Ref. [54] (Canada)). Finally, the third eschews the automatic exclusion of "blanket consents" to future research (e.g. Ref. [16] para. 8.11 (Singapore) and Ref. [50] pp. 46–47 (Germany)). Indeed, members of the National Bioethics Commission of the United States would allow the use of "coded materials" for any kind of future study without further specification as to what kind of research, or the need for further consent, or even anonymization (Ref. [15] rec.9). Coding raises, however, other issues such as that of recontact should subsequent findings become clinically significant.

The advantages and drawbacks of anonymization and coding are outlined by the French and German National Ethics Committees [12]. They both:

> " ... are in any case aware of the difficulties arising out of two antagonistic viewpoints as regard free informed consent: on the one hand, the best interest of patients and the protection of their personal data, in the name of which one might be tempted to erase as quickly as possible the link between biological material, corresponding information, and an identified individual; on the other hand there is scientific interest which justifies the possibility of being able to locate the person concerned so as to correlate his/her particular circumstances to new results. The person concerned may also wish to access these new results" (p. 38).

Indeed, the advantage of coded samples is that clinical data can be added over time and so remain scientifically viable. The disadvantage for researchers over time is that at a certain point in time the combination of research and clinical knowledge will become significant enough to have medical importance in the situation where prevention or treatment is available. This in turn calls into question the right to know/not to know and in failing to discharge it, the duty to warn.

The National Bioethics Advisory Commission of the United States has recommended that in this "exceptional" circumstance (of clinical significance) recontact and disclosure should occur (Ref. [15] Rec. 14). Similarly, other jurisdictions have recognized a right to know (and not to know) to research participants. Although the right to know is unconditional in some countries (e.g. Ref. [55] paras. 11, 12 (Estonia)), most ethics committees advocate a much more circumscribed or cautious approach. The right to know is thus not automatically applicable, but depends on the seriousness of

the condition (significant clinical value), the existence of an available cure and the availability of genetic counseling (e.g. Ref. [52] sect. 8.1 and 8.2 (UK) and Ref. [50] sect. 5.3, p. 36 (Germany)).

While adapted to a clinical setting or a single-purpose or specific research, the right to know (and not to know) might nevertheless be ill-suited when applied to large-scale population human genetic research databases, in light of the nature of the research (genetic variation), the limited amount of personal information collected and the absence of a clinical setting to convey the results. As outlined by the UK Biobank Ethics and Governance Framework:

> " ... it is questionable whether telling participants the results of measurements would be useful to them, as the data would be communicated outside of a clinical setting and would not have been evaluated in the context of the full medical record or knowledge of medication or other treatment. The significance of the observations might not be clear and the research staff will not be in a position to interpret the implications. Further it would not be constructive and might be harmful to provide information but not interpretation, counseling or support ..." (Ref. [56] comment I.B.3).

This absence of a personal right to know is balanced by the fact that "... researchers will regularly share data out of respect for the participation of the population ... "(rec. 6) in accordance with the principle of reciprocity (Ref. [10] (Canada)). This commonly takes the form of periodic publication of general aggregated results that will not enable participants' identification (see, also, Ref. [16] para. 13.1.8 (Singapore)).

In the same vein as the right to know, on the issue of access by relatives to information arising out of research (once again not applicable to large-scale human genetic research databases), Japan's Bioethics Committee of the Council for Science and Technology holds that:

> " ... in case the genetic information obtained by research may lead to an interpretation that a portion of the genetic characteristics of the participant is or, is supposed to be, connected to the etiology of a disease, this interpretation may be disclosed to his/her blood relatives following authorization by the Ethics Committee only if a preventive measure or cure has already been established for the disease in question" (princ. 15.2) (Ref. [48] (Japan)).

The conditions under which such a duty to warn arises, and its scope, are still uncertain. The Australian Law Reform Commission [47] would uphold it in situations where the threat to the life or health of the relative is not imminent but where the information disclosed would be relevant in light of the high risk to develop the disease and provided the family member's right not to know is given full consideration. Furthermore, it argues that "... in some circumstances ... a health professional has a positive 'duty to warn' third parties – even if doing so would infringe the patient's confidentiality" (21.45) (Ref. [47] Chapt. 21, and recommendations). The existence of a duty to warn in violation of a patient's right to confidentiality remains controversial, however. For instance, the French National Consultative Ethics Committee for Health and Life Sciences holds that "... medical confidentiality must be observed as regards to (*sic*) third parties including family members" (Ref. [57] (France)). The Council is of the opinion that the duty to reveal any information lies with the individual, the non-communication of critical findings not constituting grounds for finding physician criminally liable (Ref. [58] (France)).

The scientific advantage of coded samples has to be weighed against the emergence of such ongoing obligations that may be eventually adopted in national legisla-

tion. Even if such obligations could be foreseen in part by asking research participants in advance whether they would want to be recontacted or not in the event of medically significant findings, what is the longevity or validity of an anticipatory "yes" or "no"? No doubt, the courts will settle this latter question and in the meantime the option should be presented. If not, automatic communication of at-risk information to participants might run foul of the emerging right not to know and yet, on failure to do so, of an emerging duty to warn!

To conclude this section on banking, the following comments can be made with regard to the issue of other uses without obtaining another explicit consent:

1. the wholesale prohibition against both general/blanket consent to future unspecified uses of research samples and against the use of left-over samples from medical care without a specific consent is increasingly nuanced;
2. the automatic anonymization of samples as the expedient solution to ethical and legal quandaries is being re-examined and coding is emerging as the preferred option in many situations;
3. a distinction is more clearly drawn between refusal of access to third parties such as insurers or employers and the legitimate need (in prescribed circumstances) for communication to blood relatives;
4. discussion is required on the issue of recontact and communication of results in the situation of other research that yields medically relevant information; and
5. the specificity of population human genetic research databases should be clearly established and a harmonized set of principles developed; in relation to this particular type of genetic research the role of the public should be clearly defined.

Underlying these difficult choices is the ultimate question – to whom does the DNA belong in this commercialized research environment ?

21.6 Ownership

Intimately linked to the issue of ownership is that of the legal status of human genetic material. Even though this issue is one of principle, surprisingly a different legal status – person or property – has not had a concomitant impact on the ultimate issue, that of control of access by the individual over its use by others.

Internationally, regionally, and nationally, all the legal and ethical instruments acknowledge the sensitivity and special status of genetic material. This results from its individual but also familial and communal nature, and foreseen and potential value and significance (e.g. Ref. [7] art. 4a (International), Ref. [59] sect. 7 (Canada)). Traditional legal categories (property or person) cannot account for the allegedly unique nature of genetic information. A *sui generis* approach has been suggested as more appropriate (e.g. Ref. [60], and Ref. [59] sect. 7.4 (Canada)).

Internationally, there is increasing recognition and confirmation that, at the level of species, the human genome is the common heritage of humanity (e.g. Refs. [61] and [26] art.1 internationally) and that human genetic research databases are global public resources [8]. In contrast with common misunderstanding, the notion of "common heritage of humanity" means that, at the collective level, like outer space and the sea, no individual exclusive appropriation is possible by nation states. Other characteristics of this approach include peaceful and responsible international stewardship or cus-

todianship with a view to future generations and equitable access. In the absence of a binding international treaty (UNESCO's Declaration and HUGO's positions being only proclamatory in nature), it remains to be seen if this concept will come to legally binding fruition.

This position, however, is particularly important in that it serves to place new sequences and information that fail to meet the strict conditions of patenting or copyright into the public domain. While patenting is not the subject of this chapter, both a personal or property approach to DNA samples would theoretically require a specific personal consent to eventual patenting. Indeed, whether the DNA sample has the status of "person" or property, consent must be obtained, or at a minimum notification of patenting, as mentioned under the 1998 European Directive (Ref. [27] para. 26). At the international level, then, this position in favor of both the common heritage level of the collective human genome and that of personal control over individual samples and information has slowly been consolidated.

The last few years have seen the emergence of a new concept in the international arena, that of benefit-sharing. This approach largely initiated by HUGO and adopted by UNESCO and WHO, is gradually taking hold in industry. It mandates recognition of the participation and contribution of participating individuals, populations, and communities. Founded on notions of justice, solidarity and equity, it upholds the common heritage approach and so encourages the "giving-back" by researchers or commercial entities (Human Genome Organization [8]). This return of benefit "... to the society as a whole and the international community ..." (UNESCO [7] art. 19) accrues to the participating individuals or group to which they belong (for instance WHO [19] rec. 19), to healthcare services notably through the resulting availability of new prevention and diagnostic tools and treatments proceeding from research, the research community, and also to developing countries that do not possess the means and techniques to adequately collect and process data in their own territory (for instance UNESCO [7] art. 19; see also more specifically with regards to developing countries, CIOMS [5] guidelines 10 and 20).

Turning to the regional level, the "gift" language of a decade ago, was replaced with "source," and "owner", only to return in the following years (e.g., ESHG [31]). Although mentioning the traditional gift relationship as worthy of consideration, however, the European Society of Human Genetics in its Data Storage and DNA Banking of Biomedical Research [39] adopts the position that ownership and access agreements should be private and not regulated by legislation. It upholds the principle according to which, unless anonymized and thus abandoned, samples and data should remain under the control of the individual, with the investigator and processor acting as custodians. Interestingly, it affirms that in determining intellectual property rights due consideration should be paid to the notion of benefit sharing (art. 27).

In general, the language of gift is not found in international instruments, the former emphasizing the common heritage concept (Human Genome Organization [61]) or the notion of "general property" or "public domain" (Human Genome Organization [8]), thus obviating the issue of status, excluding private ownership by the "donor", and concentrating on "shared goods." One exception to this trend is, however, worth mentioning – the World Health Organization position on Genetic Databases: Assessing the Benefits and the Impact on Human and Patient Rights [19]. Promoting

the protection of individual interests and rights, the World Health Organization maintains, similarly to the European Society of Human Genetics, that individuals should remain the primary controllers of their genetic information. Such control is not, however, unfettered, but subject to the information bring identifiable. In contrast to the general return to "gift" language, it recommends that

> " ... serious consideration should be given to recognizing property rights for individuals in their own body samples and genetic information derived from those samples ..."

while ensuring that

> " ... some kind of benefit will ultimately be returned, either to the individual from whom the materials were taken, or to the general class of person to which that individual belongs ..." (rec.19).

The European Convention [17] mirroring both UNESCO and WHO, limits itself to prohibiting financial gain by stating "... the human body and its parts shall not, as such, give rise to financial gain ..." (art. 21), thus, indirectly, eschewing a property approach. This principled proscription of financial gain is reiterated specifically in the context of archived human biological material and personal data used in biomedical research in its 2002 Proposal on Archived Biological Materials [9]. The Proposal further maintains that "... research on human biological materials and personal data shall be undertaken if this is done in conformity with appropriate consent procedures ..." (art. 14). This specific consent to further use of stored samples and data for research includes consent to eventual commercialization as the participant must be informed of "... any foreseeable commercial uses of materials and data, including the research results." Finally, the Council of Europe in its 2001, Recommendation 1512 – Protection of the Human Genome by the Council of Europe, adopts the notion of the human genome as the "common heritage of mankind" in order to limit patenting rights, urging member states to change "... the basis of patent law in the international fora ..." (Ref. [18] rec. II (vi)).

At the national level, it should be stated at the outset that payment to a research participant for time and inconvenience or cost recovery by the researcher or institution (both being minimal in the case of DNA sampling) neither affords the status of property to a sample nor undermines the notion of gift. Furthermore, the notion of gift, while obviously involving transfer, might not necessarily create immediate property rights of the researcher. Indeed, in the absence of intellectual property which could be afforded to any invention, increasingly we will see that the researcher–banker is described as a custodian. This is both a real and symbolic statement. Real, in that the current complex, private-public funding of research involves multiple economic partners in any eventual profits from patenting. Symbolic, in that the researchers involved can be bench scientists or clinician-researchers and so may be both simple guardians and fiduciaries of the samples for the research participants or patients and their families.

The "nonownership" language of most national legal and ethical documents is subject to some exceptions, notably with regard to population human genetic research databases. In Estonia, by virtue of paragraph 15 of the Human Gene Research Act [55] ownership rights of samples and uncoded information are vested with the bank's chief processor. In the United Kingdom, UK Biobank Limited, the biobank legal entity, will be the legal owner of both the database and the sample collection (Ref. [56] part. II. sect. A). However, in both cases, these ownership rights are circumscribed, and it appears that the so-called "owners" of the banks act as traditional custodians enjoying certain rights over the samples and data. The chief processor in Estonia is prohibited

from transferring his rights or the written consent of gene donors. In the United Kingdom, the Ethics and Governance Framework specifies that UK Biobank Limited will act as the steward of the resource; it will not exercise all the rights akin to ownership (e.g. selling the samples), and will ensure that the public good is being served. This position is in accordance with the Medical Research Council Operational and Ethical Guidelines on Human Tissue and Biological Samples for Research, which has placed the onus on the custodian of a tissue collection to manage access to samples which are seen as gifts (Sect. 21.4.1) [52]. Its definition of custodianship "... implies some rights to decide how the samples are used and by whom, and also responsibility for safeguarding the interests of the donor ..." (Ref. [52] glossary).

In contrast and surprisingly, Iceland, with its controversial presumed consent to the storage and use of health data, extends the notion of "nonownership" to any company licensed by the Government to do research on accompanying samples:

> "The licensee shall not be counted as the owner of the biological samples, but has rights over them, with the limitations laid down by law, and is responsible for their handling being consistent with the provisions of the Act and of government directives based on it. The licensee may thus not pass the biological samples to another party, nor use them as collateral for financial liabilities, and they are not subject to attachment for debt." (Ref. [43] art. 10).

Interestingly, at the national level, the notion of benefit-sharing is acknowledged as worthy of consideration in relation to human genetic research databases, this being the result of their large commercial potential, and possible abuse by commercial entities and biased researchers. The French Conseil Consultatif National d'Éthique pour les Sciences de la Vie et de la Santé and the German Nationaler Ethikrat [12], in a joint statement on the necessity to promote a public debate on the establishment of national biobanks or "biolibraries", argue that the concept of benefit-sharing that has arisen on the international forum should be studied in depth. They support the sharing of benefits (rather than profits or advantages) with the population as a whole, and the nonmarketability of the human body and its parts, proscribing any right to financial return to individual participants (Ref. [12] art. 39–40).

The German National Ethics Council, in its recent opinion on "Biobanks for Research," confirmed this earlier position by holding that although individual donors should not benefit from the research, "... benefit sharing at a level higher than that of the individual, in the form of voluntary contributions to welfare funds, is conceivable and desirable ..." (Ref. [38] regulatory proposal 30).

Similarly in Canada, "... population research should promote the attribution of benefits to the population ..." and not be limited to those who participated (Ref. [10] rec.7). In Israel, the sharing of benefits is seen as primordial both to ensure the participation of private individuals in the research and to attract investment (Ref. [11] para. 15).

In contrast, the German Research Foundation, although considering voluntary agreements on benefit-sharing as "... welcome policy decisions ..." does not believe that researchers have an ethical or moral obligation to do so, the question of benefits and their allocation not bearing any particular significance and importance in the field of genetics. They argue that:

> " ... such a line of argumentation (the absence of any benefit-sharing policy, more specifically to donors) would equally question the fairness of national and international public policy which governs private-sector industry, fiscal policy, national and global

> health policies. There is no objective justification for describing, in the public debate, the distribution of potential benefits derived from research based on gene data as a specific problem ... Provided adequate donor privacy is ensured, investors need not have a '-guilty conscience,' nor do they have any special 'redistribution obligations' ..." (Ref. [50] art. 48-49).

In short, with regard to human genetic material (in general), the language of "donation" prevalent in the countries of civilian tradition (e.g. Ref. [62] (Quebec)), has also been adopted in common law jurisdictions (e.g. Ref. [16] (Singapore)). In the United States, even those American states that have adopted the Genetic Privacy Act [36] have not included the original articles on the property rights of the "source". Theoretically, the implementation of this approach would have given every "source-owner" an opportunity to sell and/or bargain for a percentage of eventual profits (if any).

The result of all this debate and of increased commercialization of genetic research, is that most, if not all, consent forms now inform research participants of eventual commercialization and possible profits (and the policy concerning their sharing). Ultimately, with the exception of existing and future population human genetic research databases, for which sharing of benefits with the population as a whole is contemplated, it is usually universities, research institutes, and/or commercial entities that maintain human genetic research databases and share in any profits that might ensue (e.g. Ref. [63]).

21.7 Conclusion

Genetic research has moved to the forefront of the bioethics debate in the last few years. This is, in part, the consequence of the growing public interest in understanding the role of genetic factors in common diseases and in benefiting from drugs tailored to individual genetic susceptibility. Ethical frameworks have started to make the corresponding shift from an emphasis on monogenic diseases and the stigma they carry to the "normalization" of genetic factors in common diseases. This is especially important as the study of normal genetic variation (diversity) requires large population human genetic research databases. Paralleled "normalization" of the treatment of DNA samples and genetic information with increased protection when necessary will also have to be made.

Two issues have been the primary focus of this chapter as an example of the ethical issues surrounding genetic research and DNA sampling – consent (with its related questions) and commercialization. We have seen that the issue of consent is increasingly characterized and stratified by the origin of the sample (medical or research), the degree of identifiability of the information accompanying the sample, and the issue of secondary uses. A more sensible and realistic approach is being adopted with regard to consent procedures, with the freedom of researchers on the one hand and the protection of participants on the other being better balanced. Furthermore the need to distinguish between coded and anonymized samples has been recognized; the scientific and personal value of the former being favored over the low-risk of possible socioeconomic discrimination of the latter. Increasingly, participants want to be "coded" and followed-up over time and be offered that choice over time. This will also be required when the emerging rights to know, not to know, and correlative duty to warn are fully shaped. From the standpoint of researchers, anonymized samples might lose their

scientific utility over time considering the absence of ongoing clinical information. Researchers might also come to favor coding, because the issue of recontact is being clarified.

On the issue of commercialization of research, some clarification has been forthcoming in that raw sequences with no specific or substantial utility are seen as being in the public domain and not patentable *per se*. The issue of benefit-sharing raises the possibility of balancing the legitimate return on investment (profit-making) with concerns with equity, justice and reciprocity for participating individuals, families, communities, and populations. The prospect of personal benefit by research subjects is being largely limited to ensuring clear renunciation of any interest in potential monetary profit including intellectual property. The next step will consist in clarification of the role and responsibility to be played by the researcher, the private or public institution, the governance and ethics structure (if applicable), and the industry. The possibility of conflicts of interest is real and actual where the researcher is not only a clinician but also the custodian of the sample and has a financial interest in the research.

As we move from gene mapping to gene or protein function, there is a need to understand normal genetic variation and diversity. This requires the participation of large populations. The lessons learned in the last fifteen years, specifically the need to not only respect personal values and choices in the control of and access to DNA samples in genetic research but also to clearly communicate its goals without necessarily emphasizing general vague risks, should serve to direct the next decade. Transparency, ongoing communication and public engagement strategies will do much to ensure public trust in the noble goals of genetic research and hopefully, reduce the "stigma" of the term "genetic".

Acknowledgments

The authors would like to acknowledge the contributions of Madelaine Saginur and Kurt Wong, Research Assistants at the Centre de recherche en droit public, Université de Montréal. Research completed to May 5, 2004.

References

1 (France) National Consultative Ethics Committee for Health and Life Sciences (2003), Ethical Issues Raised by Collections of Biological Material and Associated Information Data: "Biobanks," "Biolibraries," Opinion 77, Paris, March 20, 2003, Official site of the CCNE, http://www.ccne-ethique.fr/english/pdf/avis077.pdf (date accessed: April 13, 2004)

2 Robertson, J.A. (2003), Ethical and Legal Issues in Genetic Biobanking, in Knoppers, B.M.K., ed., Populations and Genetics: Legal and Socio-Ethical Perspectives (Martinus Nijhoff, Leiden, 2003)

3 Collins, F.S. (1999), Shattuck Lecture – Medical and Societal Consequences of the Human Genome Project (July 1, 1999), N. Engl. J. Med. 341, 28

4 Consortium on Pharmacogenetics (2002), Pharmacogenetics : Ethical and Regulatory Issues in Research and Clinical Practice, April 2002, http://www.firstgenetic.net/latest_news/article_16.pdf (date accessed: April 15, 2004)

5 Council for International Organizations of Medical Sciences (2002), International Ethical Guidelines for Biomedical Research Involving Human Subjects, Geneva, November 2002, Official site of the CIOMS, http://www.cioms,ch/frame_guidelines_nov_2002.htm (date accessed: April 14, 2004)

6 World Medical Association (2002), Declaration on Ethical Considerations regarding Health Databases, Washington, October 6, 2002, Official site of the WMA, http://www.wma.net/e/policy/d1.htm (date accessed: April 13, 2004)

7 United Nations Educational, Scientific and Cultural Organization (2003), International Declaration on Human Genetic Data, Geneva, October 16, 2003, Official site of UNESCO, http://portal.unesco.org/shs/en/file_download.php/6016a4bea4c293a23e913de638045ea9Declaration_en.pdf (date accessed: April 14, 2003)

8 Human Genome Organization (2002), Statement on Human Genomic Databases, London, December 2002, Official site of HUGO, http://www.hugo-international.org/hugo/HEC_Dec02.html (date accessed: April 13, 2004)

9 Council of Europe (Steering Committee on Bioethics) (2002), Proposal for an Instrument on the Use of Archived Human Biological Materials in Biomedical Research, Strasbourg, October 17, 2002, http://www.doh.gov.uk/tissue/instrbiomatproposal.pdf (date accessed: April 13, 2004)

10 (Quebec/Canada) Network of Applied Genetic Medicine (2003), Statement of Principles on the Ethical Conduct of Human Genetic Research Involving Populations, Montreal, January 1, 2003, Official site of the RMGA, http://www.rmga.qc.ca/en/index.htm (date accessed: April 14 2004)

11 (Israel) Bioethics Advisory Committee of the Israel Academy of Sciences and Humanities (2002), Population-Based Large-Scale Collections of DNA Samples and Databases of genetic Information, Jerusalem, December 2002, http://www.weizmann.ac.il/bioethics/PDF/Finalized_DNA_Bank_Full.pdf (date accessed: April 14, 2004)

12 (France and Germany) Comité Consultatif National d'Éthique pour les Sciences de la Vie et de la Santé, Nationaler Ethikrat (2003), Ethical Issues raised by Collections of Biological Material and Associated Information

Data: "biobanks," "Biolibraries," Joint Document, in National Consultative Ethics Committee for Health and Life Sciences (2003), Ethical Issues raised by collections of biological material and associated information data: "biobanks," "biolibraries," Opinion 77, March 20, 2003, Official site of the CCNE, http://www.ccne-ethique.fr/english/pdf/avis077.pdf (accessed: April 13, 2004)

13 United Nations Economic and Social Council (2003), Report of the Secretary-General on Information and Comments Received from Governments and Relevant International Organizations and Functional Commissions pursuant to Council Resolution 2001/39, Geneva, June 11, 2003, Official site of the UN, http://www.un.org/esa/coordination/ecosoc/Genetic.privacy.pdf (date accessed: April 14, 2004)

14 Baeyens, A.J., Hakiman, R., Aamodt, R., Spatz, A., (2003) The Use of Human Biological Samples in Research: A Comparison of the Laws in the United Sates and Europe, Bio-Science Law Review (September 24, 2003), Bio-Law Science Review, http://atlas.pharmalicensing.com/features/disp/1064164853_3f6dddf5630a1 (date accessed: April 13, 2004)

15 (U.S.A.) National Bioethics Advisory Commission (NBAC) (1999), Research Involving Human Biological Materials: Ethical Issues and Policy Guidance, Vol. I., Report and Recommendations of the National Bioethics Advisory Commission, Rockville, Maryland, August 1999, http://bioethics.georgetwon.edu/nbac/hbm.pdf (date accessed: April 13, 2004)

16 (Singapore) Bioethics Advisory Committee (2002), Human Tissue Research, Singapore, November 2002, Official site of the Bioethics Committee, http://www.bioethics-singapore.org/resources/reports2.html (date accessed: May 4, 2004)

17 Council of Europe (1997), Convention for the Protection of Human Rights and Dignity of the Human Being with Regard to the Application of Biology and Medicine, Oviedo, April 4, 1997, (1997), Int. Dig. Health Leg. 48, 99, Official site of the Council of Europe, http://conventions.coe.int/Treaty/EN/cadreprincipal.htm (date accessed: April 15, 2004)

18 Council of Europe (Parliamentary Assembly) (2001), Recommendation 1512 (2001), Protection of the Human Genome by the Council of Europe, 2001, Official site of the European Council, http://assembly.coe.int/Documents/AdoptedText/ta01/erec1512.htm (date accessed: April 15, 2004)

19 World Health Organization (European Partnership on Patients' Rights and Citizens' Empowerment) (2003), Genetic Databases – Assessing the Benefits and the Impact on Human Rights and Patient Rights, Geneva, 2003, http://www.law.ed.ac.uk/ahrb/publications/online/whofinalreport.rtf (date accessed: April 15, 2004)

20 The Nuremberg Code (1946–1949), Trials of he War Criminals before the Nuremberg Military Tribunals under Control Council Law No. 10, October 1946–April 1949, Vol. 2, pp.181–182. Washington, D.C.: US Government Printing Office, 1949

21 World Medical Association (1964, 1975, 1983, 1989, 1996, 2000, 2002), Declaration of Helsinki, Recommendations Guiding Physicians in Biomedical Research Involving Human Subjects, World Medical Assembly, 1964, 1975, 1983, 1989, 1996, 2000, 2002), Official site of the World Medical Association, http://www.wma.net/e/policy/b3.htm (date accessed: April 13, 2004)

22 Le Bris, S., Knoppers, B.M., Luther, L. (1997), International Bioethics, Human Genetics and Normativity, Houston Law Rev. 33, 1363

23 Council of Europe (2001), Additional Protocol to the Convention for the Protection of Human Rights and Dignity of the Human Being with regards to the Application of Biology and Medicine on the Prohibition of Cloning Human Beings, Paris, January 12, 1998 (opening for signature), March 1, 2001 (entry into force), Official site of the European Council, http://conventions.coe.int/Treaty/en/Treaties/Word/168.doc (date accessed: April 14, 2004)

24 European Parliament (2000), European Parliament Resolution on Human Cloning B5-0710, 0753 and 0753/2000, Official site of the European Parliament, http://www.europarl.eu.int/comparl/tempcom/genetics/links/b5_0710_en.pdf (date accessed: April 15, 2004)

25 Commission of the European Communities (2003), Amended Proposal for a Council Decision Amending Decision 2002/834/EC

on the Specific Programme for Research, Technological Development and Demonstration: "Integrating and Strengthening the European Research Area" (2002–2006) (presented by the Commission pursuant to Article 2509 (2) of the EC Treaty), Brussels, November 26, 2003, , COM (2003)749 final-2003/0151 (CNS), Official site of the Eur-Lex, http://europa.eu.int/eur-lex/en/com/pdf/2003/com2003_0749en01.pdf (date accessed: April 22, 2004)

26 United Nations Educational, Scientific and Cultural Organization (International Bioethics Committee) (1997), Universal Declaration on the Human Genome and Human Rights, Paris, November 11, 1997, Official site of the UNESCO, http://www.unesco.org/shs/human_rights/hrbc.htm (date accessed: April 13, 2004)

27 European Parliament (1998), Directive 98/44/EC of the European Parliament and of the Council of 6 July 1999 on the Legal Protection of Biotechnological Inventions (July 30, 1998) L 213 Official Journal of the European Communities p.13, Official site of the European Union, http://europa.eu.int/eur-lex/pri/en/oj/dat/1998/l_213/l_21319980730en00130021.pdf (date accessed: April 15, 2004)

28 (European Union) European Advisory Body on Data Protection and Privacy (2004), Working Document on Genetic Data, Brussels, March 17, 2004, Official site of the European Commission, http://europa.eu.int/comm/internal_market/privacy/docs/wpdocs/2004/wp91_en.pdf (date accessed: April 14, 2004)

29 World Health Organization (1998), Proposed International Guidelines on Ethical Issues in Medical Genetic and Genetic Services, Report of a WHO Meeting on Ethical Issues in Medical Genetics, Geneva, December 15 and 16, 1997, Official site of the World Health Organization, http://whqlibdoc.who.int/hq/1998/WHO_HGN_GL_ETH_98.1.pdf (date accessed: April 20, 2004)

30 Human Genome Organization (1998), Statement on DNA Sampling: Control and Access (February 1998) 6: 1 Genome Digest 8, http ://www.gene.ucl.ac.uk/hugo/sampling.html (date accessed: April 15, 2004)

31 European Society of Human Genetics (Public and Professional Policy Committee) (2000), Population Genetic Screening Programmes: Technical, Social and Ethical Issues, Recommendations of the European Society of Human Genetics, July 2000, Official site of the European Society of Human Genetics, http://www.eshg.org/ESHGscreeningrec.pdf (date accessed: April 20, 2004)

32 World Health Organization (1999), Draft World Health Organization (WHO) Guidelines on Bioethics, May 1999, http://www.nature.com/wcs/b23a.html (date accessed: April 20, 2004)

33 World Health Organization (1998), Resolution WHA51.10 on Ethical, Scientific and Social Implications of Cloning in Human Health, Geneva, May 16, 1998, Official site of the WHO, http://www.who.int/gb/EB_WHA/PDF/WHA51/ear10.pdf (date accessed: April 20, 2004)

34 Clayton, E.W., Steinberg, K.K., Khoury, M.J. et al. (1995), Informed Consent for Genetic Research on Stored Genetic Samples, JAMA 274, 1786

35 Knoppers, B. M. and Laberge (1995), C.M. "Research and Stored Tissues – Persons as Sources, Samples as Persons?" (1995) 274 JAMA 1806

36 Knoppers, B.M., Hirtle, M., Lormeau, S., Laberge, C., Laflamme, M. (1998), Control of DNA Samples and Information, Genomics 50, 385

37 (U.K.) Royal College of Physicians (1999), Research based on archived information and samples. Recommendations from the Royal College of Physicians Committee on Ethical Issues in Medicine, Journal of the Royal College of Physicians, London, May–June 1999, 33(3) 264

38 (Germany) German National Ethics Committee (Nationaler Ethikrat) (2004), Biobanks for research, Opinion (available in English at kontakt@ethikrat.org)

39 European Society of Human Genetics (2001), Data Storage and DNA Banking for Biomedical Research: Technical, Social and Ethical Issues, Birmingham, November 2001, official site of the ESHG, http://www.eshg.org/ESHGDNAbankingrec.pdf (date accessed: April 15, 2004)

40 (Australia) Australian Law Reform Commission (2003), ALRC 96 : Essentially Yours : The Protection of Human Genetic Information in Australia, Sydney, May 2003, http://www.austlii.edu.au/other/alrc/publications/reports/96/ (date accessed: April 26, 2004)

41 (U.S.A.) American Medical Association (2002), E-2.079 Safeguards in the Use of DNA Databanks in Genomic Research, Chicago, June 2002, Official site of the American Association, http://www.ama-assn.org/ama1/pub/upload/mm/369/ceja_opinion_2a02.pdf (date accessed: April 20, 2004)

42 (U.S.A.) Coriell Institute for Medical Research (Coriell Cell Repositories) (2003), Policy fort eh Responsible Collection, Storage and Research use of samples from named Populations for the NIGMS Human Genetic Cell Repository, March 10, 2003 (last update), Official site of the Coriell Institute, http://locus.umdnj.edu/nigms/comm/submit/collpolicy.html (date accessed: May 5, 2004)

43 (Iceland) Iceland Minister of Health and Social Security (2000), Act on Biobanks, no. 110-2000, Iceland, May 13, 2000, http://www.raduneyti.is/interpro/htr/htr.nsf/pages/Act-biobanks (date accessed: May 5, 2004)

44 (U.K.) The Royal College of Pathologists (2000), Guidelines for the Retention of Tissues and Organs at Post-Mortem Examination, London, March 2000, Official site of the Royal College of Pathologists, http://www.rcpath.org/resources/pdf/tissue_retention.pdf (date accessed: May 5, 2004)

45 (Australia) National Health and Medical Research Council (1999a), National Statement on Ethical Conduct in Research Involving Humans, Australia, 1999, Official site of the NHMRC, http://www.health.gov.au/nhmrc/publications/pdf/e35.pdf (date accessed: April 26, 2004)

46 (Netherlands) Health Research Council of the Netherlands (1994), Proper Use of Human Tissue, The Hague, January 1994, publication n°1994/01, for the executive summary see European Commission (Quality of Life Programme) (2001), Survey on Opinions from National Ethics Committees or Similar Bodies, Public Debates and National Legislations in Relation to Biobanks, Brussels, May 2001, Official site of the European Union, http://europa.eu.int/comm/research/biosociety/pdf/catalogue_biobanks.pdf (date accessed: April 22, 2004) at 17–19

47 (Australia) Australian Law Reform Commission, Australian health Ethics Committee (2003), Joint Inquiry into the Protection of Human Genetic Information, February 2003,Official site of the Office of the federal Privacy Commissioner, http://www.privacy.gov.au/publications/genesub03.doc (date accessed: May 5, 2004)

48 (Japan) Council for Science and Technology (Bioethics Committee) (2000), Fundamental Principles of Research on the Human Genome, Japan, June 14, 2000, Official site of the Council for Science and Technology, http://www.sta.go.jp/shimon/cst/rinri/pri00614.html (date accessed: April 15, 2004)

49 (U.S.A.) Council of Regional Network for Genetic Services (1997), Issues in the Use of Archived Specimens for Genetic Research, Points to Consider, Albany, New York (January 1997)

50 (Germany) Deutsche Forschungsgemeinschaft, Senate Commission on Genetic Research (2003), Predictive Genetic Diagnosis – Scientific Background, Practical and Social Implementation, Bonn, March 27, 2003, Official site of the Senate Commission, http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/2003/download/predictive_genetic_diagnosis.pdf (date accessed: May 5, 2004)

51 (Sweden) Ministry of Health and Social Affairs (2002), Biobanks [Health Care] Act (2002:297), Sweden, May 23, 2002, http://www.jus.umu.se/Elsagen/Biobanks%20%5BHealth%20Care%5D%20Act%202004-01-01.doc (date accessed: May 5, 2004)

52 (U.K.) Medical Research Council (2001), Human Tissue and Biological Samples for Use in Research: Operational and Ethical Guidelines Issued by the Medical Research Council. London, April 2001, Official site of the MRC, http://www.mrc.ac.uk/pdf-tissue_guide_fin.pdf (date accessed: April 29, 2004)

53 (U.S.A.) American Society of Human Genetics (1996), American Society of Human Genetics Statement on Informed Consent for Genetic Research, Am. J. Hum. Genet. 59, 471

54 (Canada) Medical Research Council of Canada, Natural Science and Engineering Research Council of Canada, Social Science and Humanities Research Council of Canada (1998), Tri-Council Policy Statement – Ethical Conduct for Research Involving Humans, Canada, August 1998 (with 2000 and 2002 updates), Official site of the Tri-Council, http://www.pre.ethics.gc.ca/english/pdf/TCPS%20June2003_E.pdf (date accessed: April 19, 2004)

55 (Estonia) Estonian Parliament (2000), Human Gene Research Act, December 13, 2000, http://www.legaltext.ee/text/en/X50010.htm (date accessed: May 5, 2004)

56 (U.K.) UK Biobank (2003), Ethics and Governance Framework, Draft for Comment, London, September 24, 2003, Official site of UK Biobank, http://www.ukbiobank.ac.uk/documents/egf-comment-version.doc (date accessed: May 5, 2004)

57 (France) National Consultative Ethics Committee for Health and Life Sciences (1995), Opinion and Recommendations on "Genetics and Medicine: from Prediction to Prevention", Opinion 46, Paris, October 30, 1995, Official site of the CCNE, http://www.ccne-ethique.fr/english/pdf/avis046.pdf (date accessed: April 20, 2004)

58 (France) Consultative Ethics Committee for Health and Life Sciences (2003), Regarding the Obligation to Disclose Genetic Information of Concern to the Family in the event of Medical Necessity, Paris, April 24, 2003, Official site of the CCNE, http://www.ccne-ethique.fr/english/pdf/avis076.pdf (date accessed: April 20, 2004)

59 (Canada) Pullman D., Latus A. (2003), Policy Implications of Commercial Human Genetic Research in Newfoundland and Labrador, Prepared for the Newfoundland and Labrador Department of Health and Community Services, January 2003, http://www.nlcahr.mun.ca/dwnlds/DP_Final_%20Report.pdf (date accessed: May 5, 2004)

60 Litman, M., Robertson, B.M. (1996), The Common Law Status of Genetic Material in Knoppers, B.M., Caulfield, T. and Kinsella, T.D., eds., Legal Rights and Human Genetic Material (Toronto: Edmond Montgomery, 1996) 51

61 Human Genome Organization (1996), Statement on the Principled Conduct of Genetic Research (May 1996) 3: 2 Genome Digest 2, http ://www.gene.ucl.ac.uk/hugo/conduct.htm (date accessed: April 15, 2004)

62 (Canada) Network of Applied Genetic Medicine (2000), Statement of Principles: Human Genomic Research, Montreal, Quebec, April 2000, http://www.rmga.qc.ca/en/index.htm (date accessed: April 20, 2004)

63 Knoppers, B.M. (2003), ed., Populations and Genetics: Legal and Socio-Ethical Perspectives (Martinus Nijhoff, Leiden, 2003)

64 Steering Committee on Bioethics – Working Party on Reseach on Stored Human biological Material (2004) Abstract of the 26th Meeting of the CDBI, March 16–19, 2004, official site of the Council of Europe, http://www.col.int/T/E/Legal%5FAffairs/Legal%5Fco%20operation/Bioethics/CDBI/08abstract_mtg26.asp#TopOfPage (date accessed: April 20, 2004)

65 (Iceland) The Icelandic Supreme Court, Ragnhildur Gudmundsdottir vs. The State of Iceland, Case No. 151/2003, available on the Mannvernd Website, http://www.mannvernd.is/english/. See also Ministry of Health and Soocial Security, Act on a Health Sector Database no 139/1998, 1998–1999, available through the Mannvernd Website, http://www.mannvernd.is/english/ (date accessed: April 20, 2004). See also deCODE genetics, http://www.decode.com (date accessed: April 20, 2004)

22
Biobanks and the Challenges of Commercialization

Edna Einsiedel and Lorraine Sheremeta

22.1
Introduction

From prehistoric times mankind has demonstrated an interest in and understanding of the principles and practice of genetic science. Historically, the development of agriculture and the cultivation of plants and animals for domestic and for commercial use is evidence of this interest. Greek philosophers including Hippocrates, Aristotle, and Plato wrote about the passing of human traits from generation to generation, noting that some are dominant and passed directly from parent to child [1]. Embedded in this history are humanity's hopes and fears, its noblest and most horrific aspirations.

With the mapping of the human genome [2, 3] there is optimism that society will benefit profoundly from innovations stemming from the Human Genome Project (HGP). The next major challenge for the HGP is to translate this knowledge into tangible health benefits [4]. A series of technological advances, including high-throughput DNA sequencing methods, improved modes of data storage and new bioinformatics tools will enable the translation of science to clinical practice. It is expected that analysis of data procured in large-scale studies of population genetics will enable researchers to gain a better understanding of the genetic bases of disease, hereditary transmission patterns, and gene–environment interactions that are implicated in complex diseases such as heart disease, diabetes, multiple sclerosis, and Alzheimer's disease [5, 6]. Such studies depend on large-scale population genetic research – often involving many thousands of research participants – and requiring national and international collaboration [7].

Population genetic research can be based either on previously collected and stored biological samples or on the creation of new (and often very large) repositories. We use the term "biobank" to describe a collection of physical specimens from which DNA can be derived, the data that have been derived from DNA samples, or both. This paper focuses on the issues relevant to prospectively collected population genetic biobanks that will be used for translational and basic scientific research and for clinical research. We assume such collections will be permanent in nature.

As with other technological advances, there are attendant risks and benefits associated with population genetic research. Despite the optimism that human health and

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

well being will ultimately be improved as a result of population genetic research, numerous ethical, legal and social concerns have been raised. For example, concerns about the role of informed consent [8–10], the relevance of community or group consent [11, 12], ownership of human biological materials [13–15], privacy and confidentiality [16–19], genetic discrimination and stigmatization [20–22], and eugenics [23–25] are often raised. Researchers focusing on the "ethical, legal, and social" issues (ELSI) of the HGP have commented extensively on these topics. Numerous professional organizations, including the Human Genome Organization (HUGO), the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the Council for International Organizations of Medical Sciences (CIOMS) have issued guidance documents and policy statements relevant to these topics.

The issues raised by large-scale biobank projects are particularly relevant to those countries that have biobanks, and to others like Canada that are considering launching such an initiative. The Canadian Lifelong Health Initiative is at the planning stage. It is expected to be a two-pronged population genetic research initiative comprising the Canadian National Birth Cohort and the Longitudinal Study on Aging (Canadian Institutes of Health Research, CIHR). The stated purpose of the studies is to facilitate the analysis of "... the role and interaction of different genetic and environmental exposures involved in the human development and aging processes over the life course, the multi-factorial causes and evolution of common diseases and the utilization of health care services ..." [26]. The infrastructure and data will provide the resource platform from which Canadian (and other) scientists can draw. Promoters of the project expect that the CLHI will "... place Canada at the forefront of modern health research and help attract and retain the best investigators and trainees."

The two Canadian projects expected to comprise the CLHI are similar to, though not identical to, projects already commenced or planned in Iceland (www.decode.com), Estonia (www.geenivaramu.ee), and the United Kingdom (www.ukbiobank.ac.uk). There are many important lessons to be gleaned from studying these planned or existing projects, and from experience in the United States with private-sector biobanks. In this chapter we focus specifically on issues arising from the commercialization of human genetic information derived from population genetic research and its implications for society. The issues we consider include funding arrangements, allocation of research benefits, public opinion and the role of public engagement, and ethical implications for the responsible governance of biobanks.

22.2 Background

The 20th Century has seen an exponential increase in the rate of technological development. The HGP, although initially predicting that the sequencing of the human genome would take fifteen years, was completed under budget and more than three years ahead of schedule. The HGP sparked an interest in elucidating the functions of genes. Large-scale population genetics research initiatives are aimed at uncovering gene–environment interactions implicated in a variety of complex human diseases. A shift has occurred in medical genetics away from traditional linkage analysis that focuses on specific heritable conditions, impacting relatively small numbers of affected individuals and their families, to population

genetic research initiatives that focus on large heterogeneous populations.

Practically speaking, population genetic research requires the simple collection of biological samples from individual participants. Typically a blood sample or a buccal swab is obtained. Genotype data are then derived from the sample and stored in a database as "sequence data". To a large extent the value of the sequence data resides in its persistent linkage to associated health information about the person from whom the sample is obtained. In Canada, as in the United Kingdom, Iceland, and Estonia, the linkage to health information is necessarily facilitated by, and through, publicly funded healthcare systems.

Table 22.1 illustrates the broad range of population genetic initiatives that have commenced or are at various stages of development. Funding arrangements range from predominantly public to predomi-

Table 22.1 Summary of various existing and proposed large-scale population genetic/genomic repositories and databases. (Adapted from Ref. [27])S

Project	*Financial base*	*DNA sample size*	*Budget*	*Website*
American Cancer Society Cancer Prevention Study, Lifelink Cohort (CPS-II) (US)	Public	110k	–	www.cancer.org/docroot/RES/content/RES_6_2_Study_Overviews.asp
Parents and Children (ALSPAC) (UK)	Public (Wellcome Trust)	25k(?)	£3 M	www.alspac.bris.ac.uk/welcome/index.shtml
BioBank UK	Public	500k	$66M	www.ukbiobank.ac.uk/
CDC National Health and Nutrition Examination Survey (NHANES III) (US)	Public	7300	–	www.cdc.gov/nchs/nhanes.htm
Estonian Genome Project	Public/Private (Estonian Genome Project/eGeen	1M	$150M	www.geenivaramu.ee/
European Prospective Investigation into Cancer and Nutrition (EPIC) (Europe)	Public	520k (10 countries)	–	www.ism.uit.no/kk/e/EPIC%20international.htm
Genomic Research in the African Diaspora (US)	Public/Private	25k	–	www.genomecenter.howard.edu/grad.htm
Icelandic Health Sector Database	Public/Private (Decode genetics)	280k	$212M	www.decode.com/
Latvian Genome Database	Not Known	60k	$1.7M	bmc.biomed.lu.lv/gene/ (under construction)
Marshfield Personalized Medicine (US)	Public/Private	40k	$3.8M	www.mfldclin.edu/pmrp/default.asp
Mayo Clinic Life Sciences System (US)	Public/Private (IBM)	100k	–	www.mayoclinic.org/ (no project-specific website)

Table 22.1 Continued

Project	*Financial base*	*DNA sample size*	*Budget*	*Website*
National Children's Study (US)	Public	100k	–	www.nationalchildrensstudy.gov/
Nurses' Health Study (US)	Public (NIH)	66k	–	www.channing.harvard.edu/nhs/
CARTaGENE (Canada)	Not known	50+k	$19M	www.cartagene.qc.ca/en/

nantly private, and various combinations of the two. The trend toward large sample sizes is indicative of the shift noted above away from linkage analysis to large-scale genomic research requiring data from heterogeneous populations.

This surge of interest in biobanking has occurred for several reasons. First, we have previously noted mankind's enduring interest in understanding the genetic basis of human health and disease. In addition, advances in bioinformatics have enabled countries or regions with ready access to archived health information referable to a population to mine the data in association with existing or prospectively collected genetic data. In some instances large-scale biobank initiatives have been developed, at least in part, in response to a perceived commercial potential of such resources.

The speed with which scientific and socio-political developments are occurring in the context of population genetic research is cause for concern. This is especially true given the recognition that our abilities to deal with the ethical fallout appear not to be keeping pace with technological development. This is evident in a number of key scientific areas including genetic research, stem-cell research [28], and nanotechnology [29].

22.3 Population Genetic Research and Public Opinion

The consideration of public opinion is critically important to the social and political validity and viability of large-scale scientific enterprises including biobank initiatives. Although desirable, ascertaining meaningful responses is difficult given that the concept of biobanking is largely unknown to the general public. In Canada, for example, a survey conducted in 2003 revealed there is almost no understanding of how population health research or genetic studies are conducted [30]. In general people have no idea whether biobanks are prevalent and, if so, who might be administering them. In addition, the knowledge that the public does have about biobanks seems to focus on genetic information as opposed to the physical samples that are the source of the information. Despite this, there seems to be an increasing awareness of research efforts to trace genetic histories through families and to gather data from related people.

The perception by the general public of the relative importance of genetic information serves as an important indicator about how biobanks will be viewed. Genetic information has two features that people generally identify as being distinct from other types of medical or biological information.

First, it is imbued with a predictive quality for the health of the individual from which the sample is derived. Second, although directly referable to a particular individual, genetic information has implications for family members. Public opinion in favor of greater protection for genetic information is related to a belief that genetic information is somehow unique and should be afforded greater protection than other types of health information [31].

Public views are also modulated by the intended uses for which genetic information is collected. Canadians are usually quite open to research uses of genetic information – especially where the focus is to find cures for genetic diseases [32]. Canadians seem to be more willing than Europeans for genetic information to be used in criminal investigations and health research. Over 70 % of Canadians support development of DNA testing technologies for criminal investigations versus 45 % of Europeans [33].

It is also important to note that there are significant cross-cultural variations concerning the acceptable uses of genetic technologies. An international survey asked respondents in a number of countries whether "parents should be allowed to use gene technology to design a baby to satisfy their personal, cultural, or aesthetic desires". The percentages of those who said they disapproved ranged from 97 % in Denmark, 92 % in the UK, and 87 % in the US to 76 % in Mexico, 67 % in Taiwan, and 53 % in Turkey (www.genetics-and-society.org/analysis/opinion/summary.html).

Understanding the evolution of public opinion and the reasons for those opinions in the context of population genetic research is a critical first step in assessing the acceptable uses to which the highly sensitive and personal data contained in biobanks can be put. It is also essential to consider the opinions, values, and priorities of the public when designing the necessary and most appropriate governance structures to ensure that the research participants are protected, the potential benefits for society are maximized, and that commercial involvement is appropriately managed and aligned with society's interests. With regard to this last point, efforts to create biobanks and to exploit them for research and commercial drug development have not been without controversy [34–36]. We now turn to consideration of the challenges associated with commercialization of biobank resources.

22.4 The Commercialization of Biobank Resources

A major objective of population genetic research is the development of beneficial and marketable preventative, diagnostic and therapeutic products. From this perspective, commercialization of the results of genetic research is both necessary and desirable. Commercialization is the means by which socially useful innovations are developed into products that are marketed and used in society. The commercial process is a complex, iterative process that requires innovation, research and development (including clinical trials), product development, market definition and analysis, regulatory approval, and post-approval marketing. The commercial process is frequently viewed simply as business interactions – with profit-seeking the primary motivator, and attention to ethical issues as a detractor from profit. In reality, corporate actors are becoming increasingly aware that to maintain public trust, they must devote resources (including time, energy, and money) to consideration and solution of ethical issues

[37]. Companies that fail to consider relevant ethical issues and fail to behave as responsible corporate citizens face the risk of market failure.

Of necessity, the private sector is assuming a greater role in earlier stages of the commercial trajectory in the biotechnology sector. Industry is increasingly involved in performing and funding basic and translational research. The private sector performs much of the research and development required as part of the drug and medical device approval processes. Government funding of research remains significant and frequently provides the initial incentive that spurs the innovative and commercial processes. Identification and promotion by government of innovation strategies that focus on biotechnology and information technology justify the supportive role of government in funding research and in providing support to researchers – in both the public and private sectors. The implications of this increased intermingling of academia and industry continue to be a dominant theme in medical, legal, bioethics, and policy literature [38–42].

In the context of human genetic and genomic research, the blurring of the private and public sectors has stirred controversy. Specific concerns have been raised, including the following:

1. the commercial process will inevitably lead to the undesirable commodification of the human body;
2. the commercialization process (and the patenting process specifically) will result in academic secrecy and unwillingness to collaborate on research efforts that are potentially commercializable;
3. commercial pressures will result in the premature implementation of new technologies in the marketplace (i.e., before the clinical, ethical, legal, and social issues have been appropriately considered); and
4. intellectual property (especially patents) might adversely and unduly hinder patient access to new technologies.

We specifically consider the role and perceptions of intellectual property in the commercialization of human genetic research.

22.4.1 An Emerging Market for Biobank Resources

Developments in genetic technology have made human biological materials and the genetic information derived from them increasingly valuable as raw materials for biomedical research. Biomaterials can provide insight into biological processes that cannot be gleaned from other sources. Molecular profiling and clinical validation of specific biological targets necessitates high-quality human biological materials and relevant clinical information from hundreds or even thousands of individuals. Accordingly, a commercial market for human biological materials has emerged [43, 44]. Notably, "... the number of high-volume tissue banking efforts around [the US] has gone from "a handful" 10 years ago to thousands today ..." [44]. Firms established for this purpose in the US include, among others, Ardais (www.ardais.com), Asterand (www.asterand.com), DNA Sciences (substantially all assets purchased by Genaissance Pharmaceuticals, online at www.dna.com), First Genetic Trust (www.firstgenetic.net), Genomics Collaborative (www.genomicsinc.com), and TissueInformatics (www.tissueinformatics.com, now www.paradigmgenetics.com).

The corporate history of DNA Sciences, an applied genetics firm with headquarters in Fremont California, provides an example of the potential problems associated with the commercialization of human genetic

materials and associated clinical data. The company was described as "... a large-scale consumer-focused, Internet-based research initiative designed to discover the links between genetics and common diseases." [45]. As part of its business efforts, DNA Sciences created the "Gene Trust" – a biobank to facilitate genomic research. It solicited biological samples from donors over the internet who were asked to assist in the common fight against human genetic diseases. More than 10,000 samples were obtained by DNA Sciences [45] before it was forced to sell off all of its assets, including the Gene Trust, to Genaissance Pharmaceuticals in an effort to avoid bankruptcy [46]. Although Genaissance opted against continuation of the Gene Trust for business reasons, this case raises important issues about the commodification of biomaterials and the appropriate governance of biobanks [47]. Questions remain as to whether the samples and associated data should be destroyed, whether the samples could be resold, and whether the donors who offered their samples to the Gene Trust face ongoing risks that their genetic data might be at some point used against their interests. This situation highlights a particular need to examine legal options that could be used to protect donor-participants in genetic and genomic research. Specific legal mechanisms including legal trusts might be useful in the protection of donor interests *vis a vis* public or private biobank entities [48, 49].

22.4.2 Public Opinion and the Commercialization of Genetic Resources

In addition to private biobank initiatives, local, national, and regional governments are increasingly entering the "business" of biobanking. In virtually all of the initiatives that have been widely reported, the issue of commercial gain has been a contentious issue. For example, in Iceland serious concerns have been raised over the grant of an *exclusive license* to deCODE Genetics to exploit the national Health Sector Database for profit [34, 50]. In the United Kingdom, public consultations in preparation for UK Biobank have revealed public concerns over *any* commercial involvement in Biobank. One study sponsored by the Human Genetics Commission revealed the British public's strong preference that databases should not be owned or controlled by commercial interests and that products developed through the initiative should be publicly owned [51].

In Canada, public opinion data suggest that, although the public generally supports biotechnology and the development of the biotech sector, there are serious reservations about certain aspects of commercialization and ownership of human genetic material. For example, among Canadians there seems to be deep resistance to the idea that biobanks could sell genetic data to other parties – even if informed consent is obtained [32]. In the Canadian context this might be more indicative of a judgment against the role of profit in association with healthcare than a considered decision about the particular circumstances of the use of genetic information. On the whole, however, studies of opinion in several countries suggest the public generally lacks trust in corporate participants in biomedical research [52]. This is certainly true in Canada where it has been determined that "... medical researchers are far more trusted to do what is right and are given more latitude in their ability to access personal information." [32].

22.5 Genetic Resources and Intellectual Property: What Benefits? For Whom?

It is hoped that research on banked samples will lead to commercializable diagnostic and therapeutic products. Typically, a researcher who discovers a disease gene might seek to file a patent application over the gene and its specific uses in diagnosis and/or therapy. Who, ultimately, should be entitled to benefit from research discoveries of this nature? Although an inventor has a legal right to obtain intellectual property protection of his or her invention, when a patentable invention depends on donated biological samples, novel ethical issues emerge. What are and what ought to be the proper limits of "patentable subject matter"? Should the sequencing and identification of a gene sequence implicated in specific biological processes be construed as an "invention"? Are individuals or communities that participate in population genetic research entitled to share in any benefits arising from the use of their biological material? If not, should they be so entitled? Do the benefits that patents bestow on society outweigh the anticompetitive behavior they might also inspire? What is the actual relationship between patenting and innovation?

22.5.1 Patents as The Common Currency of the Biotech Industry

In essence, patents are contractual agreements made between an inventor and the state. The ultimate objective of patent statutes and the intellectual property regime is to benefit society. By granting market exclusivity for a limited period (typically 20 years), patents provide inventors with an opportunity to recoup their research and development costs and to enjoy a period of time to exploit their inventions commercially without infringing competition. In return, the state is provided with an enabling description of the invention, sufficiently detailed to allow a person skilled in the art to work the invention. Descriptions of all inventions, including gene sequence listings, are made publicly available in patent databases at a specified time after filing. Accordingly, patent databases are valuable repositories of technical information to which others in academia or industry can turn to stimulate their own innovative efforts. The rationale underlying publication is that duplication will be avoided and improvements on the state of the art are promoted.

In the pharmaceutical industry and in the life sciences, in which research and development costs are high and the likelihood of developing a product that will generate revenues is small, patent applications and granted patents are seen to be critically important to industrial success. Patents, some argue, ensure the development of necessary drugs and medical devices that would not be developed in the absence of patents [53–56]. Patents and patent applications have become the *de facto* currency of the biotech industry. The unsettling reality is that we do not understand the impact of patents on the economy or on society more broadly [57]. Questions persist about the overall effect of patents on innovation and social welfare in the life sciences and other industrial sectors. Allan Greenspan, Chairman of the United States Federal Reserve Board has recently questioned whether "... we are striking the right balance in our protection of intellectual property rights ..." and whether the protection is "... sufficiently broad to encourage innovation but not so broad as to shut down follow-on innovation." [58].

22.5.2 The Debate over Genetic Patents

The extent to which industry has embraced gene patenting has caused much consternation. In 2001, Robert Cook-Deegan and Stephen McCormack reported in Science magazine that more than 25,000 DNA-based patents had been issued in the US by the end of 2000 [59]. The intense scientific focus on, and public investment in, the HGP combined with the sharp increase in the rate of genetic sequence patenting (including whole gene sequences, EST, and SNP) stimulated much debate about the relative benefits and risks of genetic patenting [60]. Over the past several years numerous commentaries and reports suggesting a variety of reforms, both inside and outside patent law, have been promulgated by international agencies, research institutes, non-governmental organizations, government agencies, and individual academics around the world [61–65].

Despite the frequently heralded benefits of patents, their utility in the field of genetics and genomics remains contested. In an oft-cited paper, published in Science magazine, on the effects of patents on innovation in biomedical research, Michael Heller and Rebecca Eisenberg describe the "tragedy of the anticommons" wherein a scarce resource is prone to under use when multiple owners each have a right to exclude others and no one has an effective privilege of use [66, 67]. In conclusion, the authors warn that:

> "... privatization must be more carefully deployed if it is to serve the public goals of biomedical research. Policy makers should seek to ensure coherent boundaries of upstream patents and to minimize restrictive licensing practices that interfere with downstream product development. Otherwise, more upstream rights may lead paradoxically to fewer useful products for improving human health." [66]

Although it is accepted that genes, including human genes, are patentable subject matter in most, if not all jurisdictions, the debate continues to rage about the ethical appropriateness of patentability. In its report on gene patenting, The Nuffield Council on Bioethics aptly concluded on this point that:

> "... exclusive rights awarded for a limited period are, in the main, defensible and that the patent system has in general worked to the benefit of the people. Nonetheless, we consider that in the particular case of patents that assert property rights over DNA, consideration should be given to whether the balance between public and private interests has been fairly struck [63]."

A growing body of data tends to support Heller and Eisenberg's thesis. It has, for example, been shown that transaction costs resulting from high licensing fees and royalties might deter laboratories from providing genetic tests [68, 69]. In France, an economic study of the cost-effectiveness of genetic testing strategies including BRCA1/2 testing showed that the latter had the highest average cost per mutation detected. The authors of this study conclude that the broad scope of the patent inhibits healthcare systems from choosing the most efficient testing strategy [70]. There is evidence of these same effects in Canada. Though these specific concerns seem valid, caution must be applied in presuming that the effects are, *in sum*, negative.

At least one commentator has taken issue with the approach taken by Heller and Eisenberg and argues that many of the problems described in a series of papers authored by Eisenberg and Rai [66, 71–74] would be worse if patents were not available [75]. Patents, argues Kieff, increase output by increasing both input and efficiency and, although not perfect, they are the best option and they act to ensure that biological

research will be funded, to some extent, through the private sector. Similarly, an OECD report published in 2002 entitled "Genetic Inventions, Intellectual Property Rights and Licensing Practices" concludes that:

> "... the available evidence does not suggest a systematic breakdown in the licensing of genetic inventions. The few examples used to illustrate theoretical economic and legal concerns related to the potential for the over-fragmentation of patent rights, blocking patents, uncertainty due to dependency and abusive monopoly positions appear anecdotal and are not supported by existing economic studies [65]."

More research in this area is clearly needed. Several studies sponsored by the National Institutes of Health in the US aim to provide evidence that will enable rigorous assessment of the impact of university-held gene patents [76, 77]. Without clear evidence that a problem exists, or without a clear understanding of an identified problem, attempts to fix the system seem ill-advised. The case of Myriad Genetics is frequently cited as justification for patent reform.

22.5.3 Myriad Genetics

Around the world, numerous patents have been granted to Myriad Genetics over two genes – BRCA1 and BRCA2 – associated with familial breast and ovarian cancer. Myriad has been harshly criticized for failing to broadly license testing methods covered by its patents. In hindsight, the decision to refuse to broadly license its patented genetic testing methods might have been a poor business decision – if for no reason other than that it has become a public relations disaster. The practice itself is not inherently wrong or immoral and, in fact, many firms across a variety of industrial sectors employ similar business strategies.

The problem for Myriad lies in the fact that the genetic testing services it seeks to promote are of profound interest to individuals who require testing. In addition, in countries with publicly funded healthcare systems, monopolistic business practices and pricing can adversely impact the ability of the state to provide healthcare services. This reality seems to account for the relative enthusiasm shown by countries like France and Canada in opposing Myriad's monopoly. For example, as a result of the Myriad patents the cost of BRCA1 and BRCA2 testing in Canada increased from $1000–1500 to $3850 per test. In response to the price increase, provincial Ministries of Health in Canada were forced to make tough decisions over whether or not to publicly fund testing in their province. To our knowledge, Quebec is the only Canadian province to offer the service to its residents through Myriad Genetics. Other provinces tacitly argue that Myriad's patents are invalid and continue to provide potentially infringing testing in defiance of the Myriad patents. To date, legal action has not been commenced by Myriad against a single Canadian province despite continued testing.

On this same issue, headlines were made in Europe and around the world on May 18, 2004, when the Opposition Division of the European Patent Office granted an appeal against Myriad Genetics over patent EP 699754 for a "Method for Diagnosing a Predisposition for Breast and Ovarian Cancer" [78, 79]. As a prelude to this decision several notices of opposition were filed at the EPO against Myriad Genetics. In France, a notice of joint opposition was filed by the Institute Curie, the Assistance Publique-Hopitaux de Paris, and the Institute Gustave

Roussy. The initiative was supported by the French Hospital Federation, the French Ministries of Public Health and Research, and the European Parliament. Further notices of opposition were filed by a group led by the Belgian Society of Human Genetics and the genetics societies of Denmark, Germany, and the United Kingdom [80, 81]. Myriad's patent was challenged on the basis that there was no sufficient inventive step and that the patent application failed to disclose a sufficient description of the invention. It is important to note that patent laws in Canada and the United States provide no similar summary opposition procedure that can be used to challenge issued patents. Rather, the validity of patents must be tested in costly patent-infringement litigation.

Ultimately, the patent at issue was revoked on the basis that the invention claimed by the applicant was not inventive. Opponents of the patent had uncovered discrepancies between the sequence described in the initial application filed in 1994 and in the patent granted in 2001. The correct sequence was only filed by Myriad after the same sequence had already been published by another party and had thus become part of the "prior art". Myriad has until the end of this year to appeal the decision. An earlier decision of the EPO Opposition Division in February of this year struck down another of Myriad's patents relating to BRCA2, because the charity Cancer Research UK had filed a patent application on the gene first [78]. Myriad faces two additional opposition hearings in 2005 relating to others of its granted patents [78, 79]. Although these decisions continue to make headlines and are seen by many as moral victories against industry writ large, it remains unclear how much weight the ethical concerns over gene patenting potentially hold.

22.5.4 Proposed Patent Reforms

Largely in response to the concerns over gene patents and the patenting of higher life forms, academics have called for substantial reform of patent laws. A variety of changes have been recommended; these include:

- creating a statutory definition of "patentable subject matter" that includes or excludes certain biotechnological inventions;
- adding an "ordre public" or morality clause to the Patent Act;
- adding a statutory opposition procedure similar to that which exists in Europe (and which has been successful thus far in striking down two of Myriad Genetics' breast cancer patents);
- creating a narrow compulsory licensing regime that would facilitate access by others to key patented technologies; and
- creating a specialized court to ensure that only judges with expertise in technology and patent law can hear intellectual property cases.

Of these recommendations, the adoption of an "ordre public" or morality clause is the most contentious. In effect, such a clause would enable patent examiners (or an appointed ethics panel or other agreed upon ruling body) to determine the patentability of inventions on the basis of morality [82]. Gold and Caulfield recommend the creation of an independent, transparent and responsible tribunal made up of specialists in ethics, research, and economics with the power to suspend or withhold patents in certain limited circumstances. The authors envisage a mechanism that "... avoids delays in the patent-granting process, that

leaves ethical decisions to specialists and that prevents frivolous complaints against patentees." On this issue, the Canadian Biotechnology Advisory Committee (CBAC) alternatively recommends the *status quo* be maintained and that:

> "... social and ethical considerations raised specifically by biotechnology should continue to be addressed primarily outside the Patent Act. While some proposals have been made to modify the Patent Act (see Annex D), the existing range of mechanisms available to restrict or prevent activities determined to be socially or morally undesirable, is quite extensive. If new limits are required, it will be more effective at present to modify or expand current regulations than to introduce a completely new mechanism into the Patent Act [83]."

Whatever steps countries opt to take with respect to their patent laws, the issues raised by patenting in the life sciences will inevitably continue to be debated.

Allowing patents over human biological materials, including human genetic material, is frequently criticized on the basis that it will create a demand for such materials and will increase the likelihood that individuals will be inappropriately exploited [84]. Although the connection between patents and commodification of the human body has been explored in the context of gene patenting, to date, the debate has had little impact on patent policy. Indeed, it is important to critically consider whether or not patent law or procedure is the correct forum for addressing morality [85]. It might well be that regulation of specific uses of patented products or processes is a better approach. We believe it is important to recognize the relative values inherent in human biological materials, human life, and the patent system in promoting the development of products and processes that could benefit society. Importantly, the potential of significant benefits accruing to society as a result of the patent system should not be lightly discounted without solid evidence to the contrary.

22.5.5 Patenting and Public Opinion

Public perceptions on patenting in the field of biotechnology are diverse. A variety of groups in Japan, including the public and scientists, were asked whether "people should be able to obtain patents and copyrights" with regard to *new* plant varieties, *new* animal varieties, *existing* plant/animal genes and *existing* human genes. Support for patenting fell among both groups as the focus moved from new plant and animal varieties to patenting existing plant/animal and human genes [86]. This hostility toward patenting genetic material already in existence was also evident among members of the New Zealand public [87].

The Canadian public strongly supports the mapping of the human genome and, with the success of this enterprise, has shown increased support for the idea of patenting genes. Concerns have been raised, however, about the possibility of patents driving up prices of medical products and reducing accessibility. Most Canadians associate genome research with these products and have indicated in a national survey that equality of access should be the primary guiding principle in commercialization, including patenting of the products [88]. Another concern is the patenting of higher life forms. Half of Canadians who were asked about patenting of the Harvard mouse said they were "not very comfortable" or "not at all comfortable" with the earlier Court of Appeal decision to grant a patent on the mouse [88].

Swedish perceptions of commercializing genetic information reveal similar concerns

[89]. Assessment of responses to commercializing technology made it clear that although respondents had little concern over commercializing information technology, gene technology posed serious concerns about "ethics", chief among which was the idea of commercializing such information [89].

22.6 Human Genetic Resources and Benefit-Sharing

In addition to specific concerns about patenting, concerns are frequently raised that the process of commercialization in the field of genomics has the potential to encourage inequitable distribution of the benefits (and burdens) of technology [90]. In response to this potential, the concept of benefit-sharing has emerged in the international law arena. Although conceived in the realm of international law, the principles are useful in the context of population genetic research initiatives at national, regional, and community levels [47].

The human genome is a unique natural resource and one that has qualities that might appropriately render it a "common heritage" resource [91–93]. In addition, genomic information has qualities that have led some to argue it should be characterized as a "global public good" [94, 95]. The effect of characterizing the human genome in these ways is to morally, if not legally, oblige researchers and exploiters to promote the equitable sharing of the resource and of the information gleaned from use of the resource. The justification for benefit-sharing in the context of non-human genetic material is logically extendable to human genetic material [47]. Currently there seems to be a clear ethical imperative that demands benefit-sharing in the context of human genetic research and a nascent legal obligation to do the same.

Despite the lack of a fully crystallized legal obligation to share benefits in the context of human genetic research it is a topic that has sparked discussion amongst scholars in law, medicine, philosophy, and bioethics [11, 48, 96–99]. Benefits are defined broadly to ensure that many types of gain (not necessarily limited to financial gain) are distributed equitably – to the researchers and industry partners who transform research findings into products and services, the international research community, the patients, communities and populations who participate in the research process, and to society in general. The often forgotten corollary is that if the benefits are to be considered shareable, so ought the burdens. On reflection, benefit-sharing seems to be best conceived as a means to effect distributive justice in circumstances where there is evidence that distributive justice is lacking under an existing regime. The concept is remedial and flexible and might vary dramatically from project to project. Fairness is the key consideration when developing a benefit-sharing strategy.

The Convention on Biological Diversity [100] and the associated Bonn Guidelines [101], which apply to nonhuman genetic materials, elucidate requirements for benefit-sharing arrangements that must be employed in that context. Details of benefit-sharing agreements are not strictly determined by the Bonn Guidelines. Rather, the guidelines simply suggest that a variety of types of benefit might form part of a benefit-sharing agreement. Table 22.2 provides a nonexhaustive characterization of monetary, nonmonetary, and hybrid benefits that might be considered and included in a benefit-sharing agreement.

Benefit-sharing mechanisms typically contemplate the equitable distribution of

Table 22.2 Types of monetary, nonmonetary, and hybrid benefits that may accrue as a result of genetic research.

Monetary benefits	*Hybrid benefits*	*Nonmonetary benefits*
Access fees	Joint ventures or other collaborations	Information
Royalties	Joint ownership of IPR	Participation in research
License fees	Technology transfer	Contribution to education
Research funding	Social recognition	

the benefits that arise *as a result* of the research process. Although strongly discouraging individual inducements to participate in genetic research, the HUGO Ethics Committee expressly acknowledges that "... agreements with individuals, families, groups, communities or populations that foresee technology transfer, local training, joint ventures, provision of health care or of information infrastructure, reimbursement of costs, or the possible use of a percentage for humanitarian purposes ..." are not similarly prohibited [102]. In fact, HUGO's Statement on Benefit-Sharing discloses potential mechanisms that might be used to effect benefit-sharing between sponsor companies and communities that participate in population genetic research. In addition to the benefits listed in Tab. 22.2, nonmonetary benefits in the context of human genetic research might also include such things as the provision of health care (including the provision of drugs or treatment developed as a result of research), development of information infrastructures, social recognition, or simply communication of the results of the research [103]. Appropriate sharing arrangements will depend on the particularities of the research, the parties involved, and the pre-existing social, cultural and political environment. Parties to such arrangements might include governmental, nongovernmental, or academic institutions and indigenous and local communities.

It is relevant to note that compensation paid to individual participants for participation in research or to induce individuals, groups, communities or populations to participate in genetic (or other) research are inducements to participate and as such are not properly considered "benefit-sharing". In fact, inducements to participate in genetic research are discouraged in international ethical statements and in national research guidelines and policy statements [102, 104, 105].

To the extent that patents create or promote inequity, benefit-sharing might be used as a corrective mechanism. Benefit-sharing should neither be used nor viewed as a tool to undermine the existing intellectual property regime. Rather, the accrual and exploitation of intellectual property combined with the use of appropriate benefit-sharing mechanisms will provide incentives to innovators and a mechanism by which sustainable development in the context of human biological resources can be equitably achieved [47].

There is currently a clear ethical imperative demanding the consideration of benefit-sharing arrangements in the context of human genetic research. There is, however, no clear or crystallized legal imperative to demand benefit-sharing. Having said this, planners of biobank initiatives are well-advised to recognize that failure to adequately address concerns and expectations of a subject population that participates in the es-

tablishment of biobank resources – including the sharing of benefits – might result in total failure of such initiatives. Prospectively planned and implemented benefit-sharing arrangements can be used to foster and maintain public trust in the commercial process and to enable the long-term viability of biobank initiatives.

22.7 Commercialization and Responsible Governance of Biobanks

Governance refers to "... those processes by which human organizations, whether private, public or civic, steer themselves ..." [106]. It is "... the process whereby societies or organizations make their important decisions, determine whom they involve in the process, and how they render account ..." [107]. In the context of biobanking, governance issues arise in and between organizations, including public and private institutions, sponsor companies, regulatory agencies, research ethics boards, researchers, research participants, and the general public. The roles assumed by these different institutions and agencies, how they interact, the decisions they make, and how they relate to the stakeholders are all part of the governance mix.

The governance of biobanks plays a key role in ensuring accountability and in building and maintaining the requisite public trust. To do this, the various institutional structures and procedures implicated in the governance processes and the interrelationships between the governing entities must be clearly articulated. Biobanks require participation, cooperation, and oversight from numerous entities including research ethics boards, data protection bodies (or privacy commissioners), professional associations, and regulatory agencies. The need to develop oversight mechanisms that are independent of the management of such an enterprise has been stressed [108, 109]. For example, the Biobank UK website (http://www.biobank.ac.uk/ethics.htm) provides that:

> "The Funders [of Biobank UK] have committed to the establishment of an Ethics and Governance Council (EGC) to act as an independent guardian of the project's Ethics and Governance Framework and to advise the Board of Directors on the conformance of the UK Biobank's activities within this Framework and the interests of participants and the public."

Tools for governance also include the applicable law, policy, and ethical norms. A pressing question that remains is whether currently available governance tools adequately address the concerns that arise in the context of biobanking. Are the existing tools sufficiently adaptable or are other tools including legislation, regulations, and policy statements that are specific to biobanks needed? When considering the appropriateness of the current governance framework, public concerns about commercial involvement and the potential misuse of genetic information cannot be underestimated. With the increasing trend to collection and use of genetic information, instances of misuse could also potentially increase unless measures are taken to ensure the integrity of population genetic research, providing an appropriate balance between encouraging innovative research, protecting research participants, and providing beneficial outcomes to society.

Given the focus of research on human genetics and genomics and the trend toward establishment of biobank resources, there is an urgent need for development of guidance documents that can be used by research ethics boards and private industry. Canada's Tri-Council Policy Statement

[105] provides insufficient guidance to ethics boards charged with ethical review of biobank projects. Specifically, guidance on the applicability of privacy law and consent options that might exist in the context of population genetic research needs to be developed for each jurisdiction. It is possible that existing regulatory frameworks might have to be changed to accommodate genomic research. What is needed is an open and accountable regulatory framework that incorporates the legal and ethical norms governing research on human subjects and the evolving ethical norms of corporate governance. In considering what such a framework might look like, the following questions are relevant and require substantial consideration:

- Who can or should own or control a population genetic biobank?
- How is access to biobank resources granted? To whom? Under what conditions?
- Should population genetic research be considered more akin to communitarian public health research or to traditional autonomy-driven human subject research?
- If more akin to public health research, how might this characterization affect the governance framework?
- How should the interests of the community or population and the interests of the research participants be represented in any commercial agreements that flow from biobanking?
- Can the commercialization of products and services developed from population genetic research be simultaneously promoted and aligned with the best interests of society?
- Is there an emerging legal obligation that would require the incorporation of benefit-sharing mechanisms into population genetic initiatives?
- How might benefit-sharing arrangements be implemented? In what circumstances are such arrangements necessary?

Given the overall lack of experience in creating and managing biobank resources and the emerging appreciation of what might be accomplished as a result of genomic research, answers to some of these questions can be garnered only through practical experience. However, proactive consideration of the issues known to be implicated in the establishment of large scale population genetic initiatives is necessary and might encourage the development of innovative approaches to biobank governance. The challenges presented by large-scale biobank initiatives suggest that what might be required is a legislative framework that ensures that the governance of such initiatives proceed in the public interest. It has been argued for the UK Biobank that self-regulatory mechanisms are no longer sufficient in the face of erosion of public confidence and trust [108].

22.7.1 The Public Interest and the Exploitation of Biobank Resources

As noted previously, the main objective of population genetic research is to develop new methods to prevent, diagnose, and treat human disease. It is inevitable that private industry will be involved in the process and that it will seek to accumulate intellectual property rights over innovations. The key is to effectively align the needs of industry with the needs of the broader research community and with the populations that enable such research. Numerous concerns have arisen in the context of the commercialization of genetic research, not least of which is the potential adverse effect on pub-

lic trust. Although the public is generally supportive of genetic research there is a real risk that over-emphasis of the commercial aspect will result in a backlash against population genetic research and the commercial products developed as a result of the research.

In Iceland, for example, DeCode Genetics was granted an exclusive license to develop and exploit a Health Sector Database [34, 50]. On this point, harsh criticism has been levied against the Icelandic government for entering a bad deal on behalf of its citizens. For example, Professor Greely notes that:

> "With the sale or lease of other assets of speculative value such as oil and gas deposits, the government typically keeps a financial interest in the output, often in the form of royalties. This share of the risk may limit the amount initially paid for the concession or lease, but it guarantees that government will share proportionately in a successful enterprise. Iceland has neither negotiated for a substantial initial payment nor a continuing interest. Its arrangement with DeCODE may turn out to be very one-sided [50]."

Similarly, in a paper that is highly skeptical of the commercial focus of the Icelandic Health Sector Database, Merz et al. conclude that:

> "The major ethical concerns posed by the HSD [the Icelandic Health Sector Database] arise because its primary purpose is commercial, and only secondarily does it support legitimate governmental operations. There are simply too many provisions of the overall project that serve DeCODE's interests and not those of the government or individual citizens." [34].

The Icelandic model, the authors argue, "… provides an informative counterexample that must be critically examined by others considering similar ventures." Although, in the end, pure public funding might be an unattainable ideal, an appropriate balance might be struck between public and private funders. Establishment of collaborative consortia of public and private companies that prevents unfair monopolistic behavior is one such possibility.

22.7.2 The Role of the Public and Biobank Governance

It is important that the general public be provided information about how biobank enterprises are funded, the expected benefits for the community or society, how the risks are managed, and the role of the initiative in the national innovation agenda. Research participants should be specifically informed about the foreseeable uses of their biological sample or associated data, consent conditions for future research access or secondary uses of the data, the potential for commercialization, specific information about how privacy and confidentiality will be maintained, storage conditions, and project maintenance and oversight [110]. Any benefit-sharing arrangements with individuals and/or the community at large, should these be relevant, must also be clearly articulated.

The need for the public to understand and to participate in the governance arrangements of biobanks cannot be over-emphasized. We have described governance as the processes by which decisions are made and how accountability, legitimacy and public trust are maintained. Although the need for public consultation and ongoing debate and dialog are recognized, efforts to address the need have not, to date, been adequate. To address this need a complement of innovative and flexible tools including public opinion surveys, focus group analysis, stakeholder consultation, community consultation, and web-based consultations must be developed. In addition, tools to educate the

public (including media and professionals) about genetic and genomic research must be developed. Ongoing research into the effectiveness of these tools is essential.

Public trust in research depends on the ability of governments to appropriately manage the research, development, and marketing phases of the commercial process. Developments in large-scale population genetic research point to a need to implement public education strategies to inform people about the purpose of population genetic research, the risks to participants, and the necessary role of private industry in the commercialization process. They also point to a need to develop reliable data-protection safeguards and liberal access mechanisms so that valuable research can proceed. This liberalization must only occur, however, in the context of robust safeguards and oversight mechanisms to minimize the potential risks to research subjects, many of which are both unknown and unknowable.

22.8 Conclusion

Failure to apply the highest scientific, legal and ethical standards to biobank initiatives will inevitably undermine public trust and confidence in science and in the downstream products of such research [111]. Establishing and maintaining integrity over a project's planning and research/use phases is critical. Errors made in the development of the Icelandic health sector database are instructive. On this topic, one commentator notes that:

> "... the procedural haste, the refusal to solicit the opinions of foreign experts (who have greater experience with industry/science/ethics conflicts), the unwillingness to take domestic criticism into account, the politicized and partisan debate in the case, the crude oversimplifications in the discussions and controversies over the biological processes basic to the inheritance of disease, the power of private interests, the plebiscitarian legitimation procedures in a case of subtle ethical, social and scientific controversy, all this is bound to raise a lot of misgivings, to say the least. It does not augur well for the search for consensual solutions of ethical conflicts in vulnerable domains of social life." [112].

The planners of new biobank initiatives should take every opportunity to learn from both the positive and negative experiences that others have encountered in the planning of existing biobank projects. It is necessary to understand the characteristics of the participating population, the opinion of the relevant public, and the specific regulatory environment in which the biobank will operate. Planners must pay close attention to the complex ethical concerns that surround the commercialization of research and the patenting of genetic material. Benefit-sharing arrangements, specifically tailored to particular initiatives, might be used to solidify the requisite public trust and enable the long-term success of biobank initiatives.

References

1 Lorentz, C. P., Wieben, E., Tefferi, A., Whiteman, D., Dewald G. (2002) Primer on medical genomics: history of genetics and sequencing of the human genome, Mayo Clin. Proc. 77, 773–782.
2 Venter, J. C. et al. (2001), The Sequence of the Human Genome, Science 291, 1304–1351.
3 Lander, E. S. et al., (2001), Initial Sequencing and analysis of the human genome, Nature 409, 860–921.
4 Collins, F., Green, E. D., Guttmacher, A., Guyer, M. S. (2003), A vision for the future of genomics research: a blueprint for the genomic era, Nature 422, 1–13.
5 Chakravarti, A. (1999), Population Genetics – Making Sense out of Sequence, Nat. Genet. 21(1 Suppl.), 56–60.
6 Bentley, D. R. (2004), Genomes for medicine, Nature 429, 440–445.
7 Collins, F. S. (2004), The case for a US prospective cohort study of genes and environment, Nature 429, 475–477.
8 Beskow, L. M., et al. (2001), Informed Consent for Population-Based Research Involving Genetics, JAMA 286, 2315–2351.
9 Caulfield, T. (2002),Gene Banks and Blanket Consent, Nat. Rev. Genet. 3, 577.
10 Deschenes, M., Cardinal, G., Knoppers B. M., Glass, K. C. (2001), Human Genetics Research DNA Banking and Consent: A Question of 'Form'?, Clin. Genet. 59, 221–239.
11 Weijer, C., (2000), Benefit Sharing and Other Protections for Communities in Genetic Research Clin. Genet. 58, 367–368.
12 Weijer, C., Goldsand, G., Emanuel, E. J. (1999), Protecting Communities in Research: Current Guidelines and Limits of Extrapolation, Nat. Genet. 23, 275–280.
13 Harrison, C. H. (2002), Neither Moore nor the Market: Alternative Models for Compensating Contributors of Human Tissue, Am. J. L. and Med. 28, 77–106.
14 Gold, E. R. (1996), Body Parts: Property Rights and the Ownership of Human Biological Materials, Georgetown University Press.
15 Litman, M., Robertson, G. (1996), The Common Law Status of Genetic Material, in: Knoppers, B. M., Caulfield, T., Kinsella, T. D., (Eds.), Legal Rights and Human Genetic Material, pp. 51–84, Emond Montgomery Publications Ltd.
16 Human Genetics Commission (2002), Inside Information: Balancing Interests in the Use of Personal Genetic Data.
17 Robertson, J. A. (1999), Privacy Issues in Second Stage Genomics, Jurimetrics 40, 59–77.
18 Gostin, L. O., Hodge, J. G., (1999), Genetic Privacy and the Law: An End to Genetic Exceptionalism, Jurimetrics 40, 21–58.
19 Annas, G. J. (1993), Privacy Rules for DNA Databanks: Protecting Coded 'Future Diaries', JAMA 270, 2346–2350.
20 Greely, H. T. (1997), The Control of Genetic Research: Involving the 'Groups Between', Hous. L. Rev. 33, 1397–1430.
21 Juengst, E. T. (1998), Group Identity and Human Diversity: Keeping Biology Straight from Culture, Am. J. Hum. Genet. 63, 673–677.
22 Markel, H. (1992), The Stigma of Disease: Implications of Genetic Screening, Am. J. Med. 93, 209–215.
23 Wertz, D. C., Fletcher, J. C. (1998), Ethical and Social Issues in Prenatal Sex Selection: A Survey of Geneticists in 37 Nations, Soc. Sci. and Med. 46, 255–273.

24 Anonymous (1995), Western Eyes on China's Eugenics Law, Lancet 346, 131.

25 Wertz, D. C. (2002), Did Eugenics Ever Die? Nat. Rev. Genet. 3, 408.

26 Canadian Institutes of Health Research (CHIR) (2004), The Canadian Lifelong Health Initiative, online: CIHR <http://www.cihr-irsc.gc.ca/e/strategic/18542.shtml>.

27 Kaiser, J. (2002), Population Databases Boom, from Iceland to the U.S., *Science* 298, 1158–1161 (on p. 1159).

28 Caulfield, T., Sheremeta, L., Daar, A. S. (2003) Somatic Cell Nuclear Transfer – How Science Outpaces the Law, Nat. Biotech. 21, 969–970.

29 Mnyusiwalla, A., Daar, A. S., Singer, P. A. (2003) 'Mind the Gap': Science and Ethics in Nanotechnology, Nanotechnology 14, R9–R13.

30 Pollara and Earnscliffe Research and Communications (2003), Public Opinion Research into Biotechnology Issues in the United States and Canada, 8th Wave, Biotechnology Assistant Deputy Minister Coordinating Committee.

31 Pollara and Earnscliffe Research and Communications, (2001), Public Opinion Research into Biotechnology Issues, 5th Wave, Report prepared for the Canadian Biotechnology Advisory Committee.

32 Pollara and Earnscliffe Research and Communications (2003), Public Opinion Research into Genetic Privacy Issues, Biotechnology Assistant Deputy Minister Coordinating Committee [GIP 2003].

33 Earnscliffe Research and Communications, Secondary Opinion Research into Biotechnology Issues, Report to Canadian Biotechnology Secretariat, Genome Canada, Dec 2003.

34 Merz, J. F., McGee, G. E., Sankar, P. (2004) Iceland Inc.?: On the Ethics of Commercial Population Genomics, Soc. Sci. and Med. 58, 1201–1209;

35 Høyer, K., Lynoe, N. (2004), Is Informed Consent a Solution to Contractual Problems? A Comment on the Article: Iceland Inc.?: On the Ethics of Commercial Population Genomics 58 Soc. Sci. and Med. 58, 1211.

36 Pinto, A. M. (2002), Corporate Genomics: DeCode's Efforts at Disease Mapping in Iceland for the Advancement of Science and Profits, U. Ill. J.L. Tech. and Pol'y 2002:2, 467–496.

37 Dashefsky, R. (2003), The high road to success: how investing in ethics enhances corporate objectives, J. Biolaw and Business 6:3 [no pagination].

38 Blumenthal, D., Causino, N., Campbell, E. (1996), Relationships between academic institutions and industry in the life sciences – an industry survey, NEJM 334, 368–373.

39 Blumenthal, D. (2003a), Academic–industrial Relationships in the Life Sciences, NEJM 349, 2452–2459.

40 Blumenthal, D. (2003b), Conflict of Interest in Biomedical Research, Health Matrix 12, 377–392.

41 Campbell, E., Louis, K. S., Blumenthal, D. (1998), Looking a gift horse in the mouth: corporate gifts supporting life sciences research, JAMA 279, 995–999.

42 Blumenthal, D. (1996), Ethics issues in academic-industry relationships in the life sciences: the continuing debate, Acad. Med. 71, 1291–1296.

43 Anderlik, M.R. (2003), Commercial Biobanks and Genetic Research: Banking without Checks? in: Knoppers, B. M. (Ed.) Populations and Genetics: Legal and Socio-Ethical Perspectives, pp. 345–373, Martinus Nijhoff Publishers.

44 Paxton, A. (2003), Brisk Trade in Tissue for Proteomics and Genomics Research, online: College of American Pathologists <www.cap.org/apps/docs/cap_today/feature_stories/biorepositories.html>.

45 DNA Sciences (2001), DNA Sciences Inc. Registers 10,000 Gene Trust Participants in Less than One Year, online: BioSpace <www.biospace.com/ccis/news_story.cfm?StoryID=5705915&full=1>.

46 Genaissance Pharmaceuticals (12 May 2003), Genaissance Pharmaceuticals' Acquisition of Substantially All of the Assets of DNA Sciences Is Approved, Press Release, online: Genaissance Pharmaceuticals <www.dna.com/investor/archive_releases03.html>.

47 Sheremeta, L., Knoppers, B. M. (2004), Beyond the rhetoric: Population Genetics and Benefit Sharing, Health L. J. 11, 89–117.

48 Winickoff, D. E., Winickoff, R. N. (2003), The Charitable Trust as a Model for Genomic Biobanks, NEJM 349,1180–1184.

49 Ashburn, T. T., Wilson, S. K., Eisenstein, B. I. (2000), Human Tissue Research in the Genomic Era of Medicine, Arch. Intern. Med. 160, 3377–3384.

50 Greely, H. T. (2000), Iceland's Plan for Genomics Research: Facts and Implications, Jurimetrics 40, 153–190.

51 Human Genetics Commission, (2000), Report to the Human Genetics Commission on public attitudes to the uses of human genetic information, online: <http://www.hgc.gov.uk/business_publications_public_attitudes.pdf>.

52 Einsiedel, E. (2003) Whose genes, whose safe, how safe? Publics' and professionals' views of biobanks, Unpublished Paper, commissioned by the Canadian Biotechnology Advisory Committee, April 2003.

53 Bale, H. E. Jr. (1996), Patent Protection in Pharmaceutical Innovation, Int'l J.L. and Politics 29, 95–101.

54 Kolker, K. L. (1997), Patents in the Pharmaceutical Industry, Patent World 88, 34–37.

55 Mossinghoff G. J., Kuo, V. S. (1998), World Patent System Circa 20XX A.D., JPTOS 80, 523–554.

56 Levin, R. C., Klevorick, A. K., Nelson, R. R., Winter, S. G. (1987), Appropriating the returns from industrial research and development, Brookings Paper on Economic Activity 3, 783–831.

57 Sheremeta, L., Gold, E. R., Caulfield, T. (2003), Harmonizing Commercialisation and Gene Patent Policy With Other Social Goals in: Knoppers, B. M. (Ed.) Populations and Genetics: Legal and Socio-Ethical Perspectives, pp. 423–452, Martinus Nijhoff Publishers.

58 Greenspan, A. (27 February 2004), Intellectual Property Rights, Lecture presented to the Stanford University Institute for Economic Policy Research Economic Summit, Stanford, California, online: Federal Reserve <http://www.federalreserve.gov/boarddocs/speeches/2004/200402272/default.htm>.

59 Cook-Deegan, R. M., McCormack, S. J. (2001) Patents, Secrecy, and DNA, Science 293, 217.

60 Doll, J. (1998), The patenting of DNA, Science 280, 689–690.

61 Merrill S. A., Levin, R. C., Myers, M. B. (Eds.) (2004), Patent System for the 21st Century (Prepublication Copy), National Academies Press, online: NAP <www.nap.edu/books/0309089107/html/>.

62 Ontario, Ministry of Health, (2002) Genetics, Testing and Gene Patenting: Charting New Territory in Healthcare, online: Ontario Ministry of Health <www.health.gov.on.ca/english/public/pub/ministry_reports/geneticsrep02/genetics.html>.

63 Nuffield Council on Bioethics (2002), The Ethics of Patenting DNA.

64 Australia Law Reform Commission (2004) Gene Patenting and Human Health, Discussion Paper 68, online: ALRC <www.austlii.edu.au/au/other/alrc/publications/dp/68/>.

65 Organisation for Economic Co-Operation and Development (2002), Genetic Inventions, Intellectual Property Rights and Licensing Practices: Evidence and Policies, online: OECD <www.oecd.org/dataoecd/42/21/2491084.pdf>.

66 Heller, M.A., Eisenberg, R. (1998), Can patents deter innovation? The anticommons in biomedical research, Science 280, 698–701.

67 Heller, M. (1998), The Tragedy of the Anticommons: Property in Transition from Marx to Markets, Harv. L. Rev. 111, 622–688.

68 Henry, M. R., Cho, M. K., Weaver, M. A., Merz, J. F. (2003) A pilot survey on the licensing of DNA inventions, J. L. Med. and Ethics 31, 442–449.

69 Cho, M. K., Illangasekare, S., Weaver, M. A., Leonard, D. G., Merz, J. F. (2003) Effects of patents and licenses on the provision of clinical genetic testing services, J. Mol. Diagn. 5, 3–8.

70 Sevilla, C., Julian-Reynier, C., Eisinger, F., Stopa-Lyonnet, D., Bressac-de Paillerets, B. H., Obol, H., Moatti, J. (2003), The impact of gene patents on the cost-effective delivery of care: the case of BRCA1 genetic testing, Int. J. Tech. Assessment in health care 19, 287–300.

71 Eisenberg, R. S. (1987), Proprietary Rights and the Norms of Science in Biotechnology Research, Yale L. J. 177–231.

72 Eisenberg, R. S. (1989), Patents and the progress of science: Exclusive rights and experimental use, Chicago L. Rev. 56, 1017–1086.

73 Eisenberg, R. S. (1996), Public research and private development: patent and technology transfer in government sponsored research, Va. L. Rev. 82, 1663–1727.

74 Rai, A. K. (1999), Regulating Scientific Research: Intellectual Property Rights and the Norms of Science, Nw. U.L. Rev. 94, 77–152.

75 Kieff, F. S. (2001) Facilitating Scientific Research: Intellectual Property Rights and the

Norms of Science – A Response to Rai and Eisenberg, Nw. U. L. Rev. 95, 691–706.
76 Malakoff, D., (2004), NIH Roils Academe with Advice on Licensing DNA Patents, Science 303, 1757–1758.
77 Cook-Deegan, R., Walters, L., Pressman, L., Pau, D., McCormack, S., Gatchalian, J., Burges, R. (2002) Preliminary Data on U.S. DNA-Based Patents and Preliminary Plans for a Survey of Licensing Practices in: Knoppers, B. M. (Ed.) Populations and Genetics: Legal and Socio-Ethical Perspectives, pp. 453–471, Martinus Nijhoff Publishers.
78 Coghlan, A. (2004), Europe Revokes Controversial Gene Patent, New Scientist, online: New Scientist.com <www.newscientist.com/news/print.jsp?id=ns99995016>.
79 Abbot, A. (2004), Clinicians Win Fight to Overturn Patent for Breast-Cancer Gene, Nature 429, 329.
80 Wadman, M. (2001), Testing Time for Gene Patent as Europe Rebels, Nature 413, 443.
81 Institute Curie (2004), The European Patent Office has revoked the Myriad Genetics Patent, Press Release, online: <www.curie.fr/upload/presse/190504_gb.pdf>.
82 Gold, E.R., Caulfield T., (2002) The Moral Tollbooth: A Method that Makes Use of the Patent System to Address Ethical Concerns in Biotechnology, Lancet 359 2268–2270.
83 Canadian Biotechnology Advisory Committee (2002), Patenting of Higher Life Forms and Related Issues, online: <http://cbac-cccb.ic.gc.ca/epic/internet/incbac-cccb.nsf/en/ah00188e.html>.
84 Andrews, L., Nelkin, B. (2001), Body Bazaar: The Market for Human Tissue in the Biotechnology Age, Crown Publishers.
85 Crespi, R. S. (2003), Patenting and Ethics – A Dubious Connection, J. Pat. and Trademark Off. Soc'y 85, 31–47.
86 Macer, D. R. J. (1992) Public acceptance of human gene therapy and perceptions of human genetic manipulation, Human Gene Therapy 3, 511–518.
87 Couchman, P.K. and Fink-Jensen, K. (1990) Public Attitudes to Genetic Engineering in New Zealand – DSIR Crop Research Report 138, Christchurch: Department of Scientific and Industrial Research.
88 Pollara and Earnscliffe Research and Communications (1999), Public Opinion Research into Biotechnology Issues, 3rd Wave.
89 Høyer, K. (2002), Conflicting notions of personhood in genetic research, Anthropology today 18:5, 9–13.
90 World Health Organization, Advisory Committee on Health Research (2002), Genomics and World Health, Genomics and World Health.
91 Knoppers, B. M. (1991), Human Dignity and Genetic Heritage: A Study Paper, Law Reform Commission of Canada.
92 Baslar, K. (1998) The Concept of the Common Heritage of Mankind, Martinus Nijhoff Publishers.
93 Spectar, J. M. (2001), The Fruit of the Human Genome Tree: Cautionary Tales about Technology, Investment and the Heritage of Mankind, Loy. L.A. Int'l and Comp. L.J. 23, 1–40.
94 Thorsteinsdottir, H., Daar, A. S., Smith, R. D., Singer, P. A. (2003) Genomics Knowledge, in: Smith, R., Beablehole, R., Woodward, D., Drager, N. (Eds.) Global Public Goods for Health, pp. 137–158, Oxford University Press.
95 Varmus, H. (2002), Genomic Empowerment: The Importance of Public Databases, Nat. Genet. 32:3, 3.
96 Berg, K. (2001), The ethics of benefit sharing, Clin. Genet. 59, 240–243.
97 Pullman, D., Latus, A. (2003), Clinical trials, genetic add-ons, and the question of benefit-sharing, Lancet 362, 242–244.
98 Kent, H. (2000) Benefits of Genetic Research Must be Shared, International Genome Organization Warns, CMAJ 162, 1736–1737.
99 Nicol, D., Otolowski, M., Chalmers, D. (2001), Consent, Commercialization and Benefit-Sharing, J. L. and Med 9, 80–94.
100 United Nations (1992), Convention on Biological Diversity, 31 I.L.M. 818.
101 United Nations (2002), The Bonn Guidelines on Access to Genetic Resources and Fair and Equitable Sharing of the Benefits Arising out of their Utilization, UN Doc UNEP/CBD/COP/6/20, online: Convention on Biological Diversity <www.biodiv.org/decisions/default.asp?m=cop-06&d=24>.
102 HUGO (1996), Statement on the Principled Conduct of Genetic Research.
103 HUGO (2000), Statement on Benefit Sharing.
104 Council for International Organizations of Medical Sciences (CIOMS), (2002) International Ethical Guidelines for Biomedical Research Involving Human Subjects, online: <www.cioms.ch/frame_guidelines_nov_2002>.

105 Medical Research Council, Natural Sciences and Engineering Research Council of Canada and the Social Sciences and Humanities Research Council of Canada (1998), Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans.
106 MacDonald, M. (2000), The Governance of Health Research Involving Human Subjects, Executive Summary, online: Law Commission of Canada <www.lcc.gc.ca/en/themes/gr/hrish/macdonald/macdonald_main.asp>.
107 Graham, J., Amos, B., Plumptre T. (2003), Principles for good governance in the 21st century, Policy Brief No. 15, Institute on Governance, Ottawa.
108 Kaye, J., Martin, P. (2000), Safeguards for research using large scale DNA collections, BMJ 321, 1146–1149.
109 Cardinal, G., Deschenes, M. (2003), Surveying the Population Biobankers, in: Knoppers, B. M. (Ed.) Populations and Genetics: Legal and Socio-Ethical Perspectives, pp. 37–94, Martinus Nijhoff Publishers.
110 People Science and Policy Ltd. Biobank UK: A Question of Trust. A Consultation Exploring and Addressing Questions of Public Trust. London, UK: Medical Research Council, Wellcome Trust <www.ukbiobank.ac.uk/documents/consultation.pdf>.
111 Lowrance, W. W. (2001), The Promise of Human Genetic Databases: High Ethical as well as Scientific Standards are Needed, BMJ 322, 1009–1010.
112 Edelstein, W. (1998) The Responsible Practice of Science: Remarks About the Cross Pressures of Scientific Progress and the Ethics of Research, online: Mannvernd <www.mannvernd.is/english/articles/we.twim.html>.

23
The (Im)perfect Human – His Own Creator? Bioethics and Genetics at the Beginning of Life

Gebhard Fürst

I would like to introduce you to my thoughts about a topic the importance of which for our time cannot be underestimated. Since cloning of the sheep Dolly, and as a result of other related news, or as a result of direct confrontation with the topic, many people are challenged by the questions which arise from new discoveries and technology in the field of biotechnology. All of us now have to be more responsible for our actions than ever before. This responsibility needs standards and guidelines.

German Bundespresident Johannes Rau made an impressive and clear statement at the conference "Ethics and Disability" in December 2003. He also defined the scope:

> "For the first time in the history of mankind the question arises, if we should make use of the option to change the human being and redesign man genetically. I can understand that many people are fascinated by the new possibilities opened by gene analysis and gene technology. If we discuss which possibilities we should utilize and which not the issue is not first and foremost about scientific and technical questions. We have to come to a qualitative decision. We have to decide, which technological possibilities can be matched with our value system, and which cannot [1]."

The key question is defined by the above statement. In view of current developments in gene technology and biomedicine, far-reaching changes in culture and civilization can be expected. Never before did we know so much, never before were we as capable as today. But do we want to know everything that we might be able to know? And should we do, or are we allowed to do, everything we could do, or might be able to achieve?

I represent the German Catholic Conference of Bishops on the German National Ethics Advisory Board, which has the mandate to represent the ethical questions and arguments regarding life sciences to the German government. Life sciences include, among others, biosciences and agricultural sciences, and bioinformatics, biomedicine, and pharmacy. The life sciences will deepen our knowledge about humans dramatically and therefore raise many expectations, hopes and fears. New scientific findings thus require examination if their utilization can be justified ethically [2].

I would like to mention a few examples to demonstrate that these thoughts are not speculative generalizations: In Korea a nucleus which was removed from a cell of a

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

twelve year-old boy, was implanted into a female rabbit oocyte, which first had its own nucleus removed: the utopic vision of a chimera between human and animal might soon be realized. In the United States, two deaf lesbian women ordered, through conscious selection of a deaf sperm donor, an embryo, which – as planned – developed into a deaf child. The children are not allowed to be different from those who ordered them.

Reproduction technology now makes it possible for a child to can have five parents: biological parents, social parents, and the surrogate mother who carries the embryo. It cannot be predicted which influence these possibilities might have for the child. A last example, as spectacular as it is recent, was the bold announcements of the American Raelian sect, who claimed to have given birth to the first cloned children.

It is now demonstrated that "... all, who played down the dramatic of the can do delusion are caught in a naïve attitude towards progress". Right now, it is most probable that the announced births of the first cloned humans are not genuine, as distinguished scientists assured me that it would be impossible to clone a human being.

But technical limitations like this have often been overcome in the past. The probability that cloned children will be disabled or will develop diseases later in life is great; the risk that cloning technology will be further perverted by dictators and other egomaniacs is obvious. In view of all these risks, the creation of a human clone is an "unsurpassable cynicism". It shows what science without ethical orientation is capable of. The common outrage [about human cloning] was great and many political and clerical organizations distanced themselves strongly from the idea and asked for a worldwide ban of reproductive cloning. I also want to emphasize that German researchers are unanimous in their view that therapeutic and reproductive cloning are inevitably connected. During "therapeutic" cloning a patient is cloned with the goal of breeding a microscopic embryo which bears replacement tissue. During "reproductive" cloning, the embryo is implanted into the mother. Both techniques are therefore identical in their most important steps. Scientists advancing therapeutic cloning thus create the know how for the cloning of babies [3].

It is essential to create an international ban of *all* forms of human cloning. This first point is very important to me – the possibility of cloning humans not only poses new challenges for politics. It is society as a whole, reflecting on dreams, wishes and ideals, which creates the request for human cloning. With all sympathy for particular situations, I would like to mention as an example the wish of infertile couples for a child at any cost. Another example is the egomaniac individualism of those who would like to see themselves replicated. Finally I would like to mention certain imaginations of eternal life, which are not founded in scripture, but based on the promises of biotechnology [3].

Please allow me to once again cite from the talk of Bundespresident Johannes Rau: to "... hold on to ..." moral standards "... is not always easy at a time when we are exposed every day to the message that it should be our goal to stay young forever, become more beautiful, and never sick, and stay productive indefinitely. Too many internalize these slogans without reflection or bad feelings, although they could see with open eyes that this picture of mankind has nothing to do with reality ..." [1].

23.1 Life Sciences and the Untouchable Human Being

I would first like to question the term "life sciences" critically, because I believe it usually is based on a simplistic, positivistic understanding of "life" and that because of this, predeterminations are made already when the term is used, which have far-reaching implications for the context.

The church therefore has to assert her convictions about human beings, life, and dignity, and social, economical, and political order in the discourse of society. The church understands herself as the advocate of humanity and the untouchable human being. This is implicated by the Christian belief that life is more than just a biological fact, because God created man in his own image. In addition, "non-theological" reasoning also leads to the realization that human dignity is intrinsic to humans, solely based on being human, and cannot be regulated by law. It is therefore important for me to state that we do not just represent a moral standpoint internal to theology or the church. "We are the group, which advocates that human dignity has to be protected by law and the constitution right from the beginning. Therefore we are the best alley of the law of the land [4]." The German Bundespresident Rau has also made an unambiguous statement about this:

> "There is a debate since a while, which attempts to distinguish between human dignity on one side and the protection of human life on the other side. But if human life at an early stage is denied human dignity, so that there can be a consideration of legal issues, the goals of the constitution are neglected. Legal considerations cannot replace basic ethical considerations.
> Human dignity, which for good reasons is placed at the beginning of the opening chapter of our constitution, is inseparable from the obligation to protect human life." [1].

In this context, the principle of human dignity, in which the untouchable human being, including the bodily existence, is anchored, forms not only the foundation of the democratic constitution, but is also the basis for a broad consensus within society.

This dimension of the untouchable, which defines the human being, is now threatened with being replaced in favor of secondary goals – this has been pointed out many times by the catholic church. Doesn't a human being get torn apart in the conflict of playing God, judging over life and death on one hand, and on the other hand valuing humans so low that human life becomes only biomaterial, which can be utilized? "We would rob ourselves of our humanity, if we would attempt to improve ourselves by our own power. The imperfect belongs to human nature and human creation [4]." The perfect human, who becomes his own creator, is a dangerous utopia, which might soon lead to classification of humans into different categories of perfect and imperfect, of the especially well created and valuable and in contrast the unworthy and avoidable. In view of the complexity of the questions, further reflection is necessary to determine how to act in any particular case. The first question deals with the justification of the goals: Can the road which leads to the objective be morally justified? Of great importance also is estimation of the consequences of genetic activities – what benefit can be expected and what damage might result?

> "I want to emphasize that the church is not averse to science and research. On the contrary: we request from the researchers to make every effort to find therapies for the major diseases of our time.
> All technologies, which do not endanger or destroy human life even in the early stages are ethically justified [4]."

The church is not against progress, but has a positive attitude towards life. Therefore,

she supports gene technology and bio medicine, as long as human dignity is recognized and protected; but she cannot abstain from pointing out the dangers and consequences resulting from such activity. Genetics and gene technology can have large benefits, but they can also come back to haunt people. This happens when they try openly, or secretly, to create a new human being. This creature quickly becomes an idol, at whose altar human sacrifices are made. In our country children with Down's syndrome, elsewhere girls because of their gender, today children with genetic defects, tomorrow those who lack intelligence, beauty, or simply chances for success [5].

The refusal of people to accept themselves as an image of God or, in other words, to accept the untouchability of humans, creates a wearing and destructive unhappiness and hands over humanity to the perfectionism of our own ideals without protection. In what kind of society, and with what kind of ideals, do we want to live? Shouldn't we have the goal

> " ... to live in a society, which utilizes scientific progress, without being handed over to it, which is conscientious on an every day basis that differences are part of human life and that nobody can be excluded because they are different?" [1].

On the other hand, it is questionable if one can preserve the untouchability of humans, if one denies or ignores the fact that humans are created in the image of God. On the slippery slope of constant movement and pushing the borders of mankind, it is no longer sufficient to find fulfillment, we have to find and recreate ourselves. To realize that we are God's creation and created in His image, enables us humans to accept ourselves in our human dignity. This attitude allows humans to grow beyond themselves without having to show off [5]. Human dignity cannot be earned, it is also impossible to loose. We cannot assign dignity and therefore we also cannot deny it either. Dignity is preassigned and must not be touched. Dignity also cannot be made dependant on criteria for performance, happiness or social compatibility defined by third parties.

23.2 Consequences from the Untouchability of Humans and Human Dignity for the Bioethical Discussion

Human dignity and being human cannot be separated, because the development of humans and being human are a continuous process, which spans from the fusion of the egg and the sperm, over the embryonic phase to birth, over childhood to being an adult, to illness, dying, and death. In human development there is neither from the scientific nor the anthropological view a real and evident break, independent from judgment calls. The key question of the debate is always the same – is the embryo already a human being? When does life begin, which is always also personal life? Does it begin, as defined in German law to protect embryos, from 1991, and in the German law on stem cells, from 2002, at the earliest possible time, i.e. the creation of the gamete, resulting from the fusion of the nuclei, or later? Is it entitled from the beginning to be protected by human dignity, or is it just a "mass of cells", and therefore a non-personal thing, or at the most a being on the level of an animal which can be made an object of research for "important" causes and therefore can be traded if needed and killed in the process?

According to German law for the protection of embryos, the embryo is protected from the beginning, according to criminal law (§218) only after embedding in the uterus (approximately the 10th day). In neither

is the embryo judged as being owned by the mother, or as a thing. Therefore human life (embryos) and ownership of it are incompatible.

The blurring of the border between thing and person is what makes biotechnology and research on stem cells so dangerous for us. Even a four-celled or eight-celled stage is not only potential but already real. On the other hand are they currently existing humans also (hopefully) still capable of development, and therefore have not yet realized their full potential. Nobody would deny them their existence and human dignity because of this fact.

German minister of justice Brigitte Zypries questioned in her talk at the Humboldt University in Berlin exactly this fact. According to Zypries the embryo, which was created in a test tube, is not "just any mass of cells", with which we can do as we like. But it is questionable to her, if it has human dignity, as defined in article 1 of the constitution, because it cannot develop into a human on its own, because this would require a woman. The minister therefore distinguishes between the basic right to live, which she will also concede to the embryo, which was created in a test tube, and the basic right to human dignity, which in her opinion can only be assigned to a child in the womb. Mrs. Zypries enters dangerous territory – the environmental conditions are necessary, but not essential, for the embryo to be a human being. The womb does not create new potential for the embryo. Also, the remark that human dignity requires "essential elements", the possibility for "self responsibility" and "self-determined organization of life" is a risky thesis – severely disabled people, patients in a coma or suffering from dementia caused by old age mostly cannot organize their lives in the self-determined way defined by the justice minister.

The embryo is a human being and an individual person right from the beginning – there is a lack of arguments for stepwise protection, because there is no moment in the development where one could define that the embryo becomes a human being. Also, therefore, there are not first and second-class embryos, and it is wrong to distinguish between embryos worthy of protection and those that are not, as the German justice minister attempts. It is not possible to assign human dignity to the unborn life in the womb and deny it to the artificially created embryo. The only possibility of not becoming unfocused is to accept the parallelism of the beginning of life and human dignity.

Human dignity is an expression of the meaning of life, which cannot be a means to achieve other goals, but is an end in itself. Freedom of research and the search for health can only be a goal as long as human dignity is not touched. No human being has the right to ask for a cure at the cost of the life of another human being. Freedom of research and the search for health can only be a goal as long as human dignity is *not* touched. If the embryo is created *in vitro* and discarded after diagnosis, it is used solely as a means for other goals, the end in itself is ignored. Therefore a weighting between the right to live for an embryo and the expected advantages from research, which uses embryos "for the benefit of others" is unacceptable. The formula "Embryos have to be killed, so that already born humans can survive" is proven untenable, even if it is used to in an attempt to justify research using human embryos and embryonic stem cells with the immense promises to find cures. To make a lasting contribution to the development of our society, medical and genetic research first and foremost have to be used to secure human life and improve the quality of life. From a

Christian point of view, this goal is ethically legitimate and desirable. But at the same time, we always have to be aware that research has to be conducted hand in hand with responsibilities toward society and that even a "high ranking goal" like "healthy life" cannot be pursued at all cost.

The church does not try to defend a nostalgic view of life here, rather one based on the moral principles of the age of enlightenment, and views with great concern that especially in so-called modern and enlightened societies these principles are questioned and become available in view of economical equations. From a Christian and humanistic point of view objections must be raised if embryos are to be "wasted" for research purposes. The language used gives away a lot about our society: The idea of procreation is superimposed with the association of "industrial production". Anyone who talks in this way about the creation and transmission of human life is in danger of dealing with the results of "production of humans" in a similar way to manufactured goods or "biomaterials", and to assign the seal "worthy to live" to specific embryos and to condemn others to extermination. Our language already gives a lot away. We have to live up to human dignity and cannot confuse this with what is often called, without much thought, the value or diminished value of human life [5].

Scientific and technological experiments, especially in medicine and pharmacy, have to *be proven* to be for the benefit of humans, including future generations. The principle of human dignity implies that the human being has to be the goal and purpose of all social and scientific development, but can never be used as a means for any goal. It has to be demonstrated why something is done in research and application and not why it should not be done, or that it *might* be useful some time in the future and *might* not be harmful. Again, we deal here with the wishes and dreams of our society for health, cures, and the perfect life. Especially from the church's point of view caution has to be demonstrated here. Christian belief prevents us dreaming that anything is doable and prevents visions of salvation which are connected to technical achievements [2]. Christian belief can also provide orientation in cases of morally questionable goals and morally wrong means which are used to achieve goals. Health cannot be guaranteed, not even through pre-implantation genetic diagnosis or the breeding of replacement organs for humans. In the end, we remain finite beings who will eventually pass away and therefore are ailing humans who generally become sick and without question will die.

Let me interject a remark, which demonstrates how fast some promises in medicine and biotechnology become obsolete. The TV Show "Report aus Mainz" (06.10.2003) reported ground-breaking results which tremendously reduced the potential of stem cell research. We all still remember how much interest there was at the beginning of 2003 in the debate about stem cell research. Medical researchers, people working on ethical aspects, and politicians dealt with this topic and almost everyone had raised hopes that this would get us closer to healing severe diseases like Parkinson's or multiple sclerosis. A laboratory in Cologne implanted embryonic stem cells into live and healthy mice. Fourteen days later the scientists at the Max Planck Institute in Cologne discovered rapidly developing tumors in the animals. The tumors in the brain of the mice mean a fiasco for stem cell researchers. This is because the same biological rules apply to the implanting of mouse cells into mouse brains as to the transfer of human cells between humans. Destruction instead of a cure. With regard to the promis-

ing, but ethically controversial, research on embryonic stem cells the results from Cologne mean it is currently not possible to exclude the possibility that embryonic stem cells applied to humans could cause cancer. The philosopher and ethics researcher Honnefelder, Bonn, Germany, commented:

> "I believe that we have learned in the last two decades, that we have to be careful with bold announcements. Research is always difficult, and the discovery of useful opportunities is always connected to the potential for side-effects and averse effects."

The example mentioned above seems to be significant to me for the basic problems related to biotechnology and related medical research. Very often the debates are hampered by unfounded promises for cures and a complete lack of understanding of the short-term and long-term consequences of the new technology.

As already mentioned, the language used is an indicator of the developments – in the same way as "therapeutic cloning" is not a new therapy, but means the breeding of human biomaterials, pre-implantation diagnostics is not much help for childless couples, but means the prevention and selection of the diseased and disabled. It is a contradiction in itself if research and development, which is supposed to be for the benefit of humans, "consumes" human life in the process. In other words, in the same way that any human has the right to be not only the product of the art of genetic engineering, but the child of his or her parents, conversely parenthood does not mean just the creation of an heir with especially worthy features, but is an expression of a moral and social relationship. Humans are fathered and not created. A child is not a ware, which can be returned when it develops defects. It is a cruel imposition for any human to have to live with the knowledge that they have been selected for detailed promising criteria and are expected to develop accordingly.

23.3 Conclusion

My remarks are central points for an ethical orientation in questions of biological and medical research and its application. Let me, at the end, close the circle by relating my experiences with the current political discussion. Everyone is asked to participate – every citizen, (especially) the scientists, teachers, students, artists, and all who participate in coming to conclusions in our society and culture, so that we can derive a responsible opinion on central bio-ethical questions. The catholic church has always and repeatedly in the last few years asked for a broad and transparent discussion of these questions. This discussion has begun in the meantime and we, including myself, welcome this development.

Another point is also very important – ethics cannot be delegated to ethics centers, national ethics advisory boards or ethics committees in clinics or elsewhere. Nobody can be substituted on the topic of ethical responsibility. In the meantime, this is also shown in the work at the National Ethics Advisory Board in Berlin – this board (like similar boards elsewhere) is in no position to simply give a green or red light in response to certain questions. The National Ethics Advisory Board is no legislative instrument. It does not work like a traffic light, where red means stop, green means go, and yellows means to cross quickly. It is important to collect all technical information and the ethical and moral arguments to enable those who carry autonomous responsibility to come to a solid conclusion. The perceived benefits of gene technology

lead some to visions that everything is doable, and others to a complete rejection, both extremes are wrong. We have to develop high sensitivity and moral competence. The goals and methods of gene technology, which can be ethically justified have to be supported; improper goals must be understood and we should not believe everything that is promised, nor should we do everything that is possible. It is especially important to preserve human dignity, the basic right to life and untouchability, and the rights of self determination and personal rights, and therefore support a culture of life. Such a culture of life requires a categorical imperative, to always keep the human perspective in mind. For the scientists this means, for example, weighing the chances and risks of a research topic in a responsible manner, carefully calculating the risks and fully disclosing the activities. Our parliaments are asked to recognize the complexity, the dimension of the risk, the effects on the future and the ethical implications of gene technology by drafting appropriate laws.

Let me conclude with remarks about the title of my paper. It was inspired by an exhibit in the German Hygiene Museum in Dresden about disabled people in our society. The title of the exhibit was "The (im)perfect human being", in which the syllable "im" was enclosed in brackets in the same way as in the title I chose. In a certain sense, the basic lines of argument are summarized in this title as I have shown in my discussion. The "imperfect" is, of course, on the one hand connected to human history. Humans are characterized by their history and tradition, they carry the traces of time, of their ancestors, nobody begins life at absolute zero – inheritance is always part of the baggage. Imperfect: on the other hand the title tries to establish humans to be non-perfect in the sense of beings with shortfalls. The title therefore points to the old dream of humans, told in so many science fiction films, and which now seems to become reality through gene technology. The biotechnology boom and the fantasies about what may now be possible, increase the enormous pressure on society to be normal, or even better, perfect. Humans dream about loosing imperfections, ugliness, diseases, and disabilities. The double meaning of the "imperfect human" leads to a dramatic culmination: The stars of cinema, television, or magazines, create new dreams for humanity – and new images of the human being. Through the ideals of our time, for example beauty and health, ability to perform and enjoy, autonomy, and rationality, humans are exposed to enormous pressure to be perfect. Categories, which form our understanding of life create the measure for the self-realization of humans at the same time. Enormous promises for cures from research and medicine put the individual under enormous pressure, which might even kill. Nobel laureate Heinrich Böll (literature) wrote prophetic-sounding words years ago:

> "I prefer even the worst Christian world over a perfect god-less world, because the Christian world gives room to those, which no god-less world allows for: the crippled and sick, old and weak. And beyond room, the Christian world gives love to those who seem to be of no use to the god-less world."

My title also reminds us of the message which Christians, as followers of Jesus of Nazareth, represent in this world – everything must done to support the imperfect human, create room for those that do not match any of the norms, dedicated love and sympathy for the humans at the fringes. Let us fight for a world in which human beings, as imperfect as we might be, have a home. Instead of attempting to create more perfect humans and re-create them, let us acknowl-

edge the imperfect, because standardization for impeccable beauty and defined specifications will strangle us soon. My final plea is therefore of special importance. All who are involved in church and society with a better understanding of the problems discussed above are asked to accompany the progress of life sciences with responsibility, sensitivity, and critical-constructive engagement.

References

1 Rau J (2003) Talk at the Conference Ethics and Disability, cited from KNA, 8.12.2003.
2 Wort der Deutschen Bischofskonferenz (DB69) zu Fragen der Gentechik und Biomedizin (2001): Der Mensch: sein eigener Schöpfer.
3 Schwägerl C (2002) Auf dem Weg zum Klonkind. Frankfurter Allgemeine Zeitung 2.12.2002.
4 Interview with Bishop Fürst (2002) PURmagazin 04/2002:16.
5 Kamphaus F (2002) Der neue Mensch. Nicht suchen, finden. Frankfurter Allgemeine Zeitung 27.11.2002.

Part V
Outlook

24
The Future of Large-Scale Life Science Research

Christoph W. Sensen

24.1
Introduction

Genome research is a very rapidly evolving field, making it hard just to keep up with all the emerging developments. Although predicting several years into the future is essentially impossible, on the basis of current trends we can at least try to project some future developments. The final chapter of this book is an attempt to do that, and, as with all predictions, might well generate some controversy. To predict what might happen in the next few years, it is worth looking at the past and extrapolating from the pace of previous developments. Reviewing some of the other chapters in this book, it becomes evident that nothing in the field of genome research is really new – it only takes a much larger-scale, automated, and integrated approach to biochemistry, biology and molecular medicine. Almost all the technology and techniques were in use before the advent of genome research, only on a smaller scale and with less automation.

Two developments were crucial for the emergence of genome research. The first development was laser technology, which enabled environmentally friendly versions of several existing technologies. For example, the use of radioactivity for DNA sequencing has all but been abandoned in favor of fluorescent biochemistries. Lasers have "invaded" many aspects of the molecular biology laboratory, they are part of automated DNA sequencers, high density DNA array scanners, MALDI mass spectrometers, and confocal microscopes to name but a few. The development was that of the computer. Computers and the internet have played a major role in the development of genome research. All major machines in a molecular biology laboratory are connected to computers, and often the data collection system is directly coupled to a laser-based detector. At the same time, data exchange in the genome research world is almost completely computer-based. "It is on the web." is now a common notion and the interface that is provided through web browsers is considered the major work environment for many scientific analyses. Genomic databases are shared through the web, to the point that all new data are entered into web-accessible databases on a daily basis, often before publication. Many large-scale projects are now conducted as international collaborations, there is even a new word for this – collaboratory.

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

A typical genome project involves many individuals. This is a dramatic change from the pre-genomic era when most molecular biological publications included from two to five authors. The large size of the laboratories involved in genome research has resulted in new modes of operation, often more or less along the lines of a factory operation. Tasks are distributed in a defined way, allowing few degrees of freedom for an individual. We can easily predict that this trend will continue, making single-author publications more or less a "thing of the past".

24.2 Evolution of the Hardware

Hardware for genome research has developed at an astonishing rate. It is impossible to predict details of future machine developments, but we see several trends where new approaches might emerge in the near future. The following paragraphs contain speculations about some of these trends.

24.2.1 DNA Sequencing as an Example

As an example of the potential for future development, we would like to look at automated DNA sequencing. Whereas it was quite good to obtain 1000 to 2000 base pairs (bp) of raw sequence per day from a radioactive DNA-sequencing gel, today's capillary sequencers can produce up to 1,000,000 bp per machine in the same time. A combination of robotics, which enables up to six automated machine loads per day, 384 capillary machines, enhancements in biochemistry that enable the labeling of DNA fragments, and automation of data processing has resulted in this increased throughput. Radioactive sequencing gels all had to be analyzed more or less manually whereas today's data can be assembled automatically, enabling the researcher to spend most of the time on data analysis. DNA sequencing laboratories are well on the way toward resembling the "Ford Model-T factory".

In addition to the improvements described above, there is certainly also room for improvement in DNA separation technology. The DNA-sequencing reaction typically can cover more than 2000 bp whereas detection systems can only use between 500 and 1200 bp. Use of more sophisticated separation strategies and better detection systems could lead to another two to fourfold increase in DNA sequence production on a single device, without any changes to the biochemistry.

Further enhancements of DNA sequencing can be predicted if the shortcomings of acrylamide-based separation can be overcome. The DNA polymerases used for DNA sequencing today are selected for efficiency in the first 2000 bp. If a separation technology with a much longer separation range could be established, it is easy to foresee that DNA sequencing could see another order-of-magnitude increase, because the biochemistry could then be adjusted to accommodate the new technical possibilities. Routine future DNA sequencing reactions could yield readouts many thousands of base pairs long.

24.2.2 General Trends

The DNA sequencing example above highlights some general trends we can see occurring right now. Some of these general trends are discussed in more detail in the paragraphs below and in much more detail in other chapters in this book. In Germany, we have a saying: "Das Bessere ist der Feind des Guten", which translates as: "The better is the enemy of the good". There will be a

continuation of gradual updates and upgrades, that result in dramatic enhancements of performance, and completely new approaches that will outperform known technology by orders of magnitude. While general laboratory technology, e.g. spectrophotometers, liquid chromatography equipment, and the like have a useful lifetime of many years, this will probably not be true for some genome research-related hardware, which will need replacement at a much faster pace because of unforeseeable jumps in technology development. This probably will not pose a major threat or problem to genome research, because only approximately 10 to 15% of the total cost of a genome project is related to hardware. Even with the current level of automation, most of the expense in a Genome project is related to consumables and human labor. In future, the human component will be even further reduced by increased automation.

24.2.3 Existing Hardware Will be Enhanced for more Throughput

Almost none of the machines currently used in genome research are stretched to their physical limits. Single-photon detection systems, which can dramatically increase the detection sensitivity of genome research hardware and bring the resolution to the theoretical physical limit, higher-density data processing (many devices still detect only on the 8-bit or 16-bit level), and faster data-processing strategies can be used to enhance current technology. Some systems, e.g. DNA sequencers and high-density DNA arrays can be scaled, which has been shown through the development of 384 capillary sequencers, which have replaced models with 96 capillaries, or an increase of the number of spots on gene chips. Separation-based machines (e.g. sequencing gels or protein gels) have seen some improvements in capability over time. They might be replaced by other technology that enables higher resolution. The crucial make or break for the establishment of a completely new technology will always be the cost factor. Even if a new concept were proven to be technically superior, it still has to compete on production cost. At the end of the day, it does not really matter how data are produced, as long as the data quality is comparable.

24.2.4 The PC-style Computers that Run most Current Hardware will be Replaced with Web-based Computing

One of the biggest problems in molecular biology and genome research laboratories today is that most machines are operated with the help of a PC-style computer. Currently, there are almost as many different operating systems in use (from MacOS to Linux to Windows) as there are machines and today it is not unusual to see all of these operating systems in the same laboratory. This situation causes several problems. The PC-style computers age quickly and require high levels of system maintenance. For example, it was quite difficult to network many of the original ABI DNA sequencing machines using TCP/IP networks, which have now become the standard, requiring major computer upgrades to achieve the integration. A typical sequencing machine, for example, might last 10 years, making it necessary to replace the controlling PC two to three times to keep up with the pace of development.

To address the problems of PC-style computers, we predict that client–server models will replace the stand-alone systems currently in use. Thin clients that collect the

data and post them to the web will be complemented by platform-independent analysis software that can be executed on large servers or the latest workstations. An early example of this approach was the LiCor 4200 Global system (Chapt. 6). Data collection on the LiCor 4200 Global system is performed by a Linux-based Netwinder thin client, which operates an Apache web server for data access. Users can control and monitor this kind of machine from any web browser (e.g. Netscape or Internet Explorer). The analysis software for the LiCor 4200 is written in the Java programming language, making it platform-independent, thus enabling execution on any Java enabled platform. LiCor has now adopted this technology for their entire machine line. Other manufacturers at least have web servers built into their systems, which allows them to offer their client software through the internet.

24.2.5 Integration of Machinery will Become Tighter

In the past, laboratories have operated many devices that could not directly "talk" to each other. Incompatible operating systems and the lack of data exchange possibilities prevented a high degree of automation. We see attempts to change this, machines from different manufacturers are starting to "speak the same language". Interestingly, this does not necessarily mean that all machines understand a common, standardized data format, instead they are capable of exporting and importing results generated by other devices. For example, spectrophotometers have become capable of exporting data sheets that can be imported into pipetting robots, saving time in the setup of PCR reactions. Sequence assembly programs can "pick" DNA sequencing primers and export the lists to oligonucleotide synthesizers, saving the time of retyping them before oligonucleotide synthesis. Laboratories are starting to build more and more "assembly lines" for genome research and, logically, the trend for integration and data exchange will continue. International standards bodies are now working on the definition of data standards for molecular biology and genome research. One example is the development of the MIAME (minimum information about microarray experiment) standard (http://www.mged.org/Workgroups/MIAME/miame.html), a joint development between academia and several industrial partners, which is now generally used for gene-chip data.

24.2.6 More and more Biological and Medical Machinery will be "Genomized"

The initial "genome research toolkit" consisted mainly of machines for DNA sequencing. The next add-on of major machinery was proteomics-related mass spectrometers, followed by machinery for expression studies, including DNA high-density arrayers and DNA chip readers. As more and more aspects of biological, biochemical, biophysical and medical research are automated, we can expect more and more of the mostly manual devices used today to be automated and developed further for high throughput. With the complete blueprints of organisms at hand, it is logical that the current set of genome research tools will be expanded by *in-vivo* and *in-vitro* studies of organelles, cells, organs, and organisms. Microscopy equipment, and imaging equipment in general, will be part of the genome research toolkit of the future. Physiological research will be increasingly automated to achieve a level of throughput that enables study of the global biological system rather than a single aspect. Structural aspect of molecules will

become increasingly important to genome research, protein crystallization factories might emerge to enable a dramatic increase in the throughput of protein structure determination. The latter will be a prerequisite for efficient use of synchrotron facilities for determination of protein structure.

The lessons learned by automating procedures in the molecular biology laboratory will certainly be applied in the integration and automation of an increasing number of biological techniques. Although there have been some attempts to create fully automated laboratories, we predict there will always be a human factor in the laboratory. Manned space flight, even though considered dangerous and costly, and thus not feasible, is still a major factor today and we expect that high-tech genome-research laboratories will develop under similar constraints.

24.3 Genomic Data and Data Handling

Bioinformatics is the "glue" that connects the different genome research experiments. The role of computer-based analysis and modeling cannot be overestimated. Currently, several hundred genome research-related databases exist. Many of these databases are updated daily and growing exponentially. New databases are being created at an enormous pace, as more and different experiments are added to the collected works of genome research. It is completely impossible to host all genomic data on personal computers or workstations; thus, there is a requirement for high performance computing environments in every serious genome research effort. We predict that in the future the computational infrastructure for most bioinformatics laboratories will be organized as a client–server model. This will address the performance issue that comes with the exponentially growing data environment, while controlling computer maintenance aspects and thus the major cost of the computational infrastructure.

To date, computer chip development has been according to Moore's law (which predicts that the computing power of a CPU doubles approximately every 18 months) and genomic data have been produced at a rate that could be accommodated with the current computer advancements. It is conceivable that the pace of computer development and genomic data production will in the future lose this synchronization, causing problems with the amount of data that must be handled. The likelihood of this seems small, however, because there are other large-scale databases (e.g. astronomical databases or weather data) that are much more extensive than all the genomic data under consideration.

In addition to scaling current computational environments, an increasing amount of dedicated hardware is being developed that can assist genome research. The goal for future bioinformatics environments should be to provide real-time analysis environments. We are certainly far away from this goal today, but as more systems like the Paracel GeneMatcher (http://www.paracel.com) or TimeLogic Decypher boards (http://www.timelogic.com), which accelerate database searches by up to a factor of 1000, enter the market, they are poised to revolutionize the computational environment.

Networking between genome research laboratories is a basic requirement for the success of any genome project, thus biological and genome research applications have been a major application domain for the development of advanced networking strategies. For example, the Canadian Bioinformatics Resource, CBR-RBC (http://www.

cbr.nrc.ca) is the major test bed for new networking developments in Canada. The trend to establish advanced networking connections in genome research laboratories will continue and ultimately all these laboratories will be connected to broad-band systems. Distributed computational infrastructure has been established in many genome projects and efforts are in progress to create nation-wide bioinformatics networks in several countries, including European countries and Canada. Of special importance to these efforts are computational GRIDs; several countries in Europe, the United States of America, and Canada are now establishing dedicated bioinformatics GRID architectures. The ultimate goal of these activities is to offer powerful distributed computational environments to users who do not need to know anything about their particular architecture but can use all resources as if reside on their own machine.

A rather new development is the establishment of web services (http://biomoby.org). Similar to a GRID, these services enable users to use a large number of services without particular knowledge about their physical location or details of their installation. We expect many software packages to integrate either GRID technology or web services into their functionality within the next few years.

Almost all current databases are initially organized as ASCII flat files, with no relational database infrastructure to support them. Data access is currently handled through web-based database integration systems such as ENTREZ and SRS, or tool-integration systems such as MAGPIE (http://magpie.ucalgary.ca) or PEDANT (http://pedant.gsf.de) In these examples, reformatting of the original data into standardized HTML files creates the illusion for the user that they are dealing with a single database. HTML has been accepted as the main working environment for biologists because web browsers (e.g. Netscape and Internet Explorer) enable a platform-independent graphical view of results from genome analyses.

The use of HTML for biological and medical data is certainly very limited, because HTML was originally designed for text files, rather than data files. Several approaches are currently being pursued to create a more standardized approach to genome research-related data. The most promising candidate seems to be the XML language. XML is an extendable web language, new data types and display modes can easily be introduced into the system. Many current ASCII flat-file databases are also being provided in XML format, even Medline now offers a XML formatting option. Genomics data browsers of the future will be XML compatible, enabling multiple views of the same data without the need for the reformatting of the dataset or reprogramming of the display interface (Chapt. 16).

Similar to HTML and XML, multidimensional imaging data, for example from confocal microscopy, serial section electron microscopy, functional medical resonance imaging or micro-computer tomography experiments, need standards to make them widely accessible. Two formats are quite common, VRML (the virtual reality markup language) and the commercial OBJ format. Both formats can be read by many software packages. In the last couple of years Java 3D has been added to the fold. Java 3D can import VRML and OBJ files (and a large number of other formats). It has been already used for a wide variety of tasks, including the software operating the Mars rovers (http://www.sun.com/aboutsun/media/features/mars.html) and in oil exploration. Many multi-dimensional display units, including CAVE automated virtual reality environments have now been adapted to Java 3D technology, enabling bioinformaticians

to develop multidimensional data integration systems on any computer platform (including laptops, Linux machines or MacIntosh computers) and then execute them using high-end display units up to the CAVE, which enables ultra-high resolution through a cubed display (billions of voxels) and immersive interaction with the displayed objects. Figure 24.1 shows an example of the CAVE experience.

We predict that these breakthroughs in technology will be utilized rapidly by the bioinformatics community. We are already seeing a trend toward establishing virtual reality systems at many universities.

In summary, the bioinformatics tools of the future will integrate as many different biological data types as possible into coherent interfaces and enable online research that provides answers to complex queries in real time environments. Computer models of biological systems will become sophisticated enough to conduct meaningful "in silico" biology. This will certainly help reduce the cost of genome research and assist in the design of smarter wet-laboratory experiments.

24.4 Next-generation Genome Research Laboratories

24.4.1 The Toolset of the Future

The first generation of genome research laboratories focused on the establishment of "sequencing factories" that could determine the genomic sequence most efficient-

Fig. 24.1 Students at the University of Calgary, Faculty of Medicine, exploring a small molecule in the immersive CAVE automated virtual reality environment. More information about this technology can be obtained at: http://www.visualgenomics.ca/

ly. Initially, many of the sequencing projects, e.g. the European yeast genome project, were set up in a distributed fashion. This proved to be almost unworkable, because the entire project depended on the slowest partner. Thus, most recent projects involve relatively few partners of equal capabilities. This has led to the development of very large laboratories in industry and academia that are capable of producing up to hundreds of megabases of finished sequence per year. Examples of such development are the Wellcometrust Sanger Institute in Cambridge, UK (http://www.sanger.ac.uk), the Genome Sequencing Center at Washington University in St Louis (http://www.genome.wustl.edu/gsc/), the formation of TIGR in Maryland (http://www.tigr.org/) and companies such as Human Genome Sciences (http://www.hgsi.com/).

The exceptions to the rule that genomic DNA sequencing is conducted in large-scale laboratories are projects in developing countries, which are just entering the field. Projects in these countries are sometimes conducted in a set-up similar to the European yeast project, but even here a concentration eventually takes place which restricts genome research funding to the more advanced laboratories. The trend to larger and larger DNA sequencing laboratories has probably peaked, because the existing laboratories are capable of handling any throughput necessary. It can be expected that most, if not all, DNA sequencing in future will be contracted to these large-scale laboratories, similar to film-processing laboratories in the photographic sector before the advent of digital photography.

Over time it has become apparent that much of the DNA sequence produced initially stays meaningless. Many of the potential genes that are identified through gene searching algorithms do not match any entries in the public databases, thus no function can be assigned to them. Without functional assignment the true goal of any Genome project, which is to understand how genomes are organized, and how the organism functions, cannot be achieved. It is also not possible to protect any intellectual property derived from the genomic sequence if the function of the molecule is not known (Chapt. 21). The logical consequence is that many genome laboratories today are trying to diversify and broaden their toolset.

Proteomics was the first addition to the genomic toolset and is also a very rapidly evolving field. Today, many expressed proteins can be identified via the 2D-gel–mass spectrometry route (Chapt. 7), but there are still many limitations that make it impossible to obtain and examine a complete proteome. Proteins that occur in very small quantities in a cell, and proteins that are rarely expressed can elude current detection methods. Separation techniques have limitations which are difficult to overcome (e.g. membrane proteins and 2D protein gels). It is to be expected that new proteomics-related techniques and approaches will be introduced at a fast pace. Protein chips and new separation techniques that will bypass 2D-gel systems will be introduced and will help to advance protein studies dramatically. At the same time, current technologies will constantly be improved. As an example, we have seen the development of dedicated MS–MS-ToF systems coupled to a MALDI front-end. More and more dedicated technology for large-scale characterization of proteins can be expected within the next few years. The large-scale study of proteins will ultimately reach the same level as the large-scale study of DNA molecules today.

Expression studies (macro arrays, high-density arrays, chips) have become a very popular addition to the genomic toolset (Chapt. 11). The major bottleneck for this technology is still the computer analysis

component, especially with regard to normalization of the data, which has not yet fully caught up with the establishment of the hardware for this technology. We expect it will be seamless in the future to go from an RNA extraction to functional analysis of the DNA array or DNA chip experiment. Whereas today's chip readers are about the size of a medium-sized printer, we expect integrated handheld devices will be developed over time; these will enable field testing with diagnostic chips, returning results instantaneously.

Certain organisms, for example yeast, *Caenorhabditis elegans*, and mice are being subjected to intense knockout studies to gain insight into the function of as many genes as possible. This technology is not without pitfalls, because the knockout of many genes does not cause any "visible" effect, whereas the knockout of other genes is lethal to the organism, which causes great difficulty in interpretation of the function of these genes. Moreover, most of the characteristics of an organism are derived from the actions of several genes. We predict that this technology will not be applied widely outside the current model systems, because of the cost and the sometimes uninterpretable results. More promising is the development of knock-down technologies using siRNA and related methods. These technologies are already widely used and we expect this trend to continue.

For many years, researchers have attempted to use structural information to deduce the function of proteins (Chapt. 12). Most of these attempts have been unsuccessful, yet we predict there will be a better chance of deduction of function from structure in the future, because then the entire dataset (many complete genomes, and all biologically relevant folds) will be known. This knowledge can then be used to predict the function of a similar gene with much more accuracy. "Structure mining" projects attempting to determine all biologically relevant structures are currently in progress, and even though this process is tedious and slow, it is likely that within the next five to ten years these efforts will be successful.

Although it will certainly be possible to identify the function of many genes using the technology mentioned above, existing technology in other fields will need to be applied to genome research in innovative ways, and entirely new technologies will have to be added to existing biological, biochemical, and medical tools, to gain a complete understanding of the function of all genes in the context of the living organism.

Imaging technologies are a prime candidate for systems that must be added to the current tool to furnish three dimensional data and time-related information about processes in cells, organs, and organisms. Automation has started to "invade" the microscopy sector, assisted by laser technology that enables micromanipulations unheard of twenty years ago. Today, automated microinjection, optical tweezers, and confocal microscopes that can be switched to high-speed imaging, shooting thousands of images per second are reality. Flow cytometry is now capable of screening hundreds of thousand of cells per minute, sorting candidate cells with desired features in an axenic manner. This equipment will soon be introduced to genome research to enable the location of elements in a cell, monitor them over time, and create the input for more realistic computer models of life (virtual cell).

24.4.2 Laboratory Organization

There will probably be three types of genome research laboratory in the future:

- large-scale factories dedicated to specific techniques;

- integrated genome-research laboratories; and
- laboratories that coordinate data production rather than produce data.

Large scale factories will continue to exist. The advantage of lower production cost that comes with the scale and integration level of these operations will justify their existence for a long time. It is crucial to understand that most of these factories will do most of their business based on contracts with other parties. The terms of these contracts (i.e. who owns the rights to the data) have been and will be the issue that determines the success of such a factory. Most of these large-scale factories are dedicated entirely to a particular technology. We predict that new large scale technology will lead to the rise of further "factories" built around certain expertise, rather than being internalized by the existing laboratories.

Most medium-scale genome research laboratories of the future will be tightly integrated units that employ as many different techniques as possible on a large scale, to furnish a "complete" picture that helps to build more precise models of how cells function. This is the logical consequence of the lessons learned from first generation genome-research laboratories. Development in this direction is probably best described with Leroy Hood's term "systems biology". The integrated laboratories will not necessarily have to produce any data on a large scale; they might, for example, collaborate with factory-style operations for large-scale data collection in most of their projects. The crucial aspect of an integrated laboratory is the computer setup. The bioinformatics infrastructure of a "systems biology" laboratory must be quite large and sophisticated and therefore capable of handling and integrating many different data types into coherent models. "Systems biology" laboratories will thus employ large bioinformatics development teams that can provide custom software solutions to address research and development needs.

We envisage the emergence of a new type of laboratory without any capability to produce data, but rather a coordination office with data-analysis capability. This kind of laboratory will outsource all wet-laboratory activity to third parties, resulting very low overheads and high flexibility, because such a laboratory can always draw on the latest technology by choosing the right partners. Many start-up companies have begun to work this way, at least for part of their activity, and large pharmaceutical companies are now outsourcing many of their genome-research-related activity.

The risk in this arrangement is almost completely on the side of the data producer. This is an area in which governments will have to involve themselves in genome research and development, because many of the risks in high technology environments are hard to calculate and will probably not initially be assumed by industry. The countries most involved in genome research today all have large Government programs to address this need and almost certainly this involvement will have to continue for the foreseeable future.

24.5 Genome Projects of the Future

As already mentioned, the goal of genome research is to obtain the "blueprint" of an organism to understand how it is organized and how it functions. To achieve this goal, genome projects of the future will have to become truly integrated. Connecting all the different bits and pieces available in many locations around the world is certainly one of the major challenges for future genome

research. To facilitate this, data will have to be available instantaneously. Already DNA sequences are often produced under the "Bermuda agreement", which calls for posting of new data to the web immediately after data generation. Complete openness before publication is certainly a new approach in biological and medical research. The notion that only complex data analysis leads to new insights and that any particular data type in itself remains meaningless without connection to all the other types of data is becoming increasingly established.

Genome research will focus increasingly on biological questions. Almost 20 years after the first plastid genomes were completely sequenced in Japan, there are still certain genes in plastids, which have no identified function. This can certainly be attributed to the fact that much of the plastid genome research stopped when these genomes were released, because of the false notion that now "the work was done". The real work starts with completion of the sequencing of the genome and it might never stop because of the complexity of biological systems.

A key factor in the continuation of genome research will be the public perception and level of acceptance of this kind of research. Scientists have not yet been very successful in two key areas – education of the public about ongoing activity and evaluation of basic science using ethical criteria. Companies have been perceived as creating "Frankenfoods" that are potentially harmful rather than better products that serve mankind. News such as that released in 2003 by the American Raelian sect that the first humans might have been cloned have shocked and outraged the general public. Molecular biology experiments are very different from atomic research, because they can be conducted almost anywhere with little effort and thus it is quite hard to exercise control. Future genome projects will have to employ higher standards for evaluation of ethical aspects of the research. A more intense discussion with the public, and better education of the public, are absolutely necessary to generate the consensus needed for continued support. All parts of society have to be involved in the discussion (Chapt. 23) and the legislature has to react quickly as new technology emerges.

The openness asked for between scientists in order to address the goals of genome research should be a starting point for relationships with the general public. If everyone has free access to the data, public control can be exercised more readily than in a secretive environment. Much of today's genome research is funded with tax dollars; this is an easily forgotten but very important fact. The openness being requested should certainly not go as far as to conflict with the protection of intellectual property, but the general approaches to genome research science should be common knowledge and certain approaches should be outlawed because they do not lead to improvements acceptable to our society.

24.6 Epilog

One of the most frequently asked questions in genome research is: " How long is genome sequencing going to last?". Genome sequencing will be around for a very long time to come. The sequencing process has become so affordable that it is feasible to completely characterize a genome to answer very few initial scientific questions. Genomic data are now typically shared with the entire scientific community. Unlike many other biological data the genomic sequence is a fixed item, new technology will not help to improve it further, and a com-

plete genome of good quality is therefore the true end product of the sequencing process. In addition to sequencing individual genomes, scientists are now characterizing complete ecosystems by DNA sequencing. "Environmental genomics" will be used to gain insights into the large part of any habitat, which cannot be cultured. This approach is another reason why DNA sequencing will not only persist, but increase.

Several million different species occur on earth today, and each individual organism has a genotype that is different from any other individual. Comparative genome research will lead to many new insights into the organization of life. Knowledge about the diversity of life will lead to new products and cures for diseases at an unprecedented pace. The need to characterize genomic data will therefore be almost endless (Chapt. 2).

In a related research field, in study of the phylogeny of species, many thousand 18S rDNA sequences have been obtained to date, even though the basic dataset that could answer most questions had been generated more than 10 years ago. Today, probably more 18S rDNA sequences get produced in a week than in all of 1994. While it took many years to produce the first 5S rDNA sequence, today this technology is standard in any sequencing laboratory around the world.

The almost endless possibilities of genome research are very exciting for all of us involved in this kind of research, but also pose challenges every day, which go far beyond the laboratory that we usually work in. This keeps us going through the enormous efforts that it takes to complete any large-scale genome project.

Subject Index

Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

D