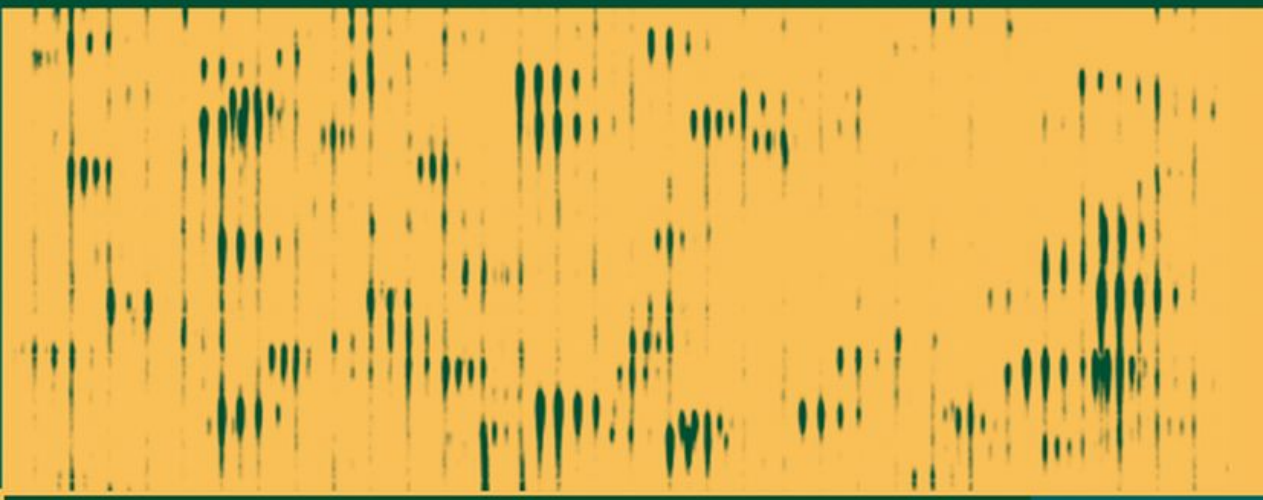
PRINCIPLES AND PRACTICE



Proteome Research

Concepts, Technology and Application

Marc R. Wilkins, Ron D. Appel,
Keith L. Williams, Denis F. Hochstrasser
Editors

Second Edition

Springer

Principles and Practice

M.R. Wilkins R.D. Appel
K.L. Williams D.F. Hochstrasser (Eds.)

Proteome Research
Concepts, Technology and Application

Second Edition

With 46 Figures, 31 in Color and 16 Tables

Springer

Professor Marc R. Wilkins, PhD
School of Biotechnology and
Biomolecular Sciences
University of New South Wales
Sydney NSW 2052
Australia

Professor Keith L. Williams, PhD
BKG Group
P.O. Box 580
Avalon NSW 2107
Australia

Professor Ron D. Appel, Ph.D.
Proteome Informatics Group
Swiss Institute of Bioinformatics
Computer Science Department
University of Geneva
CMU, rue Michel-Servet 1
CH-1211 Geneva 4
Switzerland

Professor Denis F. Hochstrasser, M.D.
Department of Structural Biology
and Bioinformatics
Department of Genetic & Laboratory
Medicine
Laboratory Medicine Service
Geneva University and University Hospital
24, rue Micheli-du-Crest
CH-1211 Geneva 14
Switzerland

Library of Congress Control Number: 2007927822

ISBN-13: 978-3-540-71240-4 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag is a part of Springer Science+Business Media

springer.com

Editor: Dr. Sabine Schreck, Heidelberg
Desk Editor: Anette Lindqvist, Heidelberg
Cover Design: WMXDesign GmbH, Heidelberg
Production and typesetting: SPi

Printed on acid-free paper 39/3180 5 4 3 2 1 0

To Sobet, Catherine, Brynnie and Anne-Catherine
To Adrien, Naara, Jarrah, Asheley, Lucile, Virginie, Sandrine and Nelson

Preface to Second Edition

This book is the result of a long-standing friendship between two research groups – one in Sydney, Australia, and the other in Geneva, Switzerland. It was stimulated by a previous book on proteomics which we produced together in 1997. Many of the authors who contributed to the original book have also written for this new book, but with an additional 10 years of experience. What is interesting about the authors of this book is that many of them have developed new proteomic technology and techniques, commercialized this technology via different routes, established proteomics and/or bioinformatics companies, and applied proteomics to large numbers of problems of scientific, clinical and industrial importance. We believe this body of experience is unusual and unique and makes this book of relevance to proteomic researchers in all areas of academic and industrial biology and medicine.

For it to be possible to write and produce this book, we are grateful for the efforts and patience of the authors of all chapters. We also acknowledge support from Australian and Swiss universities and research institutes, the Swiss Institute of Bioinformatics and the companies Proteome Systems and Geneva Bioinformatics (GeneBio) which employed some authors during their writing. We also acknowledge support from public funding agencies, including the Australian National Health and Medical Research Council and the Swiss National Science Foundation which have supported our research in recent years.

Finally, we would like to acknowledge the efforts of proteomic researchers worldwide, whose work we draw on, discuss and occasionally critique. Proteomic research is a fast-paced, growing, yet challenging area. We hope that this book will serve to further grow the field and to encourage many new researchers (young and old) to join in this endeavour. There remains much work to be done!

April 2007

Marc Wilkins, Ron Appel,
Keith Williams, Denis Hochstrasser

Foreword to Second Edition

Ten years has elapsed since the publication of the first book on proteomics by the editors of the present book. Rather than 'proteomics', the book was entitled *Proteome research: new frontiers in functional genomics*. The idea was to establish a continuity with the Genome Analysis Project, and especially the sequencing of the human genome which was under way. However, it was already clear to some of us that a new revolution in biology was being launched: the introduction of a new paradigm permitted shifting the focus of investigation from DNA sequences to structures and functions of proteins, interacting between themselves and with other molecules, including DNA, in ways not encoded in DNA sequences. After completion of the sequencing of DNA of human and other species, the picture became even clearer. As is often the case in the history of science, the previous paradigm dominated by DNA technologies allowed for discoveries which turned this paradigm upside down. 'Proteomics' – the study of the proteome, i.e. the complete set of proteins in a cell or tissue – is one of the words being used today to name the new paradigm, together with the more general expressions 'biocomplexity' and 'systems biology'. But one should not be mistaken: proteomics is not a plain continuation of genomics. DNA sequences are being used now as an indispensable source of data regarding the first level of protein structures. However, this only marks the beginning of an entirely new story. Moreover, the same protein may have completely different functions in different tissues, even in the same cell, depending upon its localization in the cell and the state of activity of the latter. Expressed DNA sequences do not tell much about three-dimensional structures of proteins or their modifications in cellular microenvironments, nor about the dynamics of their synthesis, activation and inactivation, all of these determining their functions. Knowledge of the proteome is not limited to the pattern of expressed proteins identified from DNA sequences in DNA microarrays. This has prompted a change in the whole of biological thinking. For several decades, after the extraordinary discoveries of DNA structures and functions in the 1960s, molecular genetics and genomics were a source for *explanations*, giving answers to century-old questions regarding the nature of processes specific to living beings, such as metabolism and reproduction.

These explanations were based mainly on the metaphor of a computer program written in DNA sequences, the so-called genetic program. In spite of

their being relatively simplistic, such explanations were accepted by the majority of biologists owing to their heuristic value. Protein physicochemistry, a very active field in the 1950s, was not fashionable anymore and had been almost abandoned. Among the reasons advocated were DNA technologies were easier and looked more promising. At the same time, the authors of these two books were developing two-dimensional gel electrophoresis techniques and mass spectrometry dedicated to the analysis of proteins and global protein expressions in cells and tissues. Thus, they emerged at the front line of biological research when it became clear that genomics by itself was able to provide knowledge of one-dimensional structures only, and very little knowledge of function.

This second book describes the progress made during the last 10 years. The main efforts aimed not only at improving the techniques, with the help of bioinformatics and data bases of DNA libraries, but also at tackling the more difficult problems of following protein modification and function in conditions as close as possible to their in vivo states. Progress has been made in developing reliable techniques to provide catalogues of proteins used as signatures of different normal and pathological cellular states, under well-defined conditions such as cancer versus normal cells in a given tissue. However, while this work was under way, it became clearer and clearer that post-translational modifications of proteins had to be taken into account as exhaustively as possible if protein structures were to be related to biological functions. In addition to phosphorylation, glycosylation, methylation and other covalently related modifications, more subtle intermolecular interactions are being looked for and protein–protein interaction maps are already being investigated. All these tools provide additional data which question more and more the complexity of functional regulations. How are all these interaction networks being regulated and how do they produce the observed functions? It is unlikely that a simple universal answer will be given to this question, in the form of the universal genetic code, 'identical from bacteria to elephants' according to the saying. Rather, local, ad hoc models will have to be designed and adapted to particular questions. Some medical applications are already being reported in diagnosis and drug development. Hopefully, they will develop into individualized medicine if not only individual genomes but also proteomes are made available in some distant future.

In any case, proteomics belongs to a world of postgenomics. This world has opened up a new era where more and more questions are raised rather than answers given owing to the formidable complexity being revealed. For example, there are different proteomes to be studied in more than 200 cell types (for humans only) expressing protein patterns differently, at different times, and in different conditions.

As George Klein put it nicely in a seminar on the cellular signaling pathways possibly disturbed in cancer: "Biologists must not only accept to live with complexity but to love complexity". He was quoting Tony Pawson

on cell signal transduction who pointed out that the complexity we see is nothing compared to the real complexity that exists.

Proteomics, as it is presented in this book, will most likely help biologists to 'love complexity', i.e. to be stimulated by the new problems and by the technical and theoretical tools being developed to approach them more and more efficiently.

February 2007

Henri Atlan
Professor Emeritus of Biophysics in Paris and Jerusalem
Director of the Human Biology Research Center at the Hadassah University Hospital in Jerusalem and Director of Research at the Ecole des Hautes Etudes en Sciences Sociales in Paris

Contributors

Allard, L.
Dpt R&D Immunoessais et Protéomique, BioMérieux S.A., Chemin de l'Orme, 69280 Marcy L'Etoile, France

Appel, R.D.
Proteome Informatics Group, Swiss Institute of Bioinformatics, Computer Science Department, University of Geneva, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Bairoch, A.
Swiss-Prot Group, Swiss Institute of Bioinformatics, Department of Structural Biology and Bioinformatics, University of Geneva, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Binz, P.-A.
Geneva Bioinformatics (GeneBio) S.A., 25, avenue de Champel, 1206 Geneva, Switzerland, and Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Boschetti, E.
Ciphergen Biosystems Inc., Fremont, CA 94555, USA

Bougueleret, L.
Swiss-Prot Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Citterio, A.
Department of Chemistry, Materials and Chemical Engineering "Giulio Natta", Polytechnic of Milan, Via Mancinelli 7, Milan 20131, Italy

Corthals, G.L.
Protein Research Group, Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, P.O. Box 123, 20521 Turku, Finland

Couté, Y.
Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, University of Geneva, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Gavin, A.-C.
EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Gooley, A.A.
SGE Analytical Science Pty Ltd., 7 Argent Place, Ringwood, VIC 3134, Australia

Herbert, B.R.
Proteomics Technology Centre of Expertise, Faculty of Science, University of Technology, Sydney, P.O. Box 123, Broadway, NSW 2007, Australia

Hernandez, P.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Hochstrasser, D.F.
Department of Structural Biology and Bioinformatics, Department of Genetic & Laboratory Medicine, Laboratory Medicine Service, Geneva University and University Hospital, 24, rue Micheli-du-Crest , 1211 Geneva 14, Switzerland

Hoogland, C.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Lescuyer, L.
Department of Genetic & Laboratory Medicine, Laboratory Medicine Service, Geneva University Hospital, 24, rue Micheli-du-Crest , 1211 Geneva 14, Switzerland

Lisacek, F.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Packer, N.H.
CORE of Functional Proteomics and Cellular Networks, Macquarie University, Sydney, NSW 2109, Australia

Palagi, P.M.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Righetti, P.G.
Department of Chemistry, Materials and Chemical Engineering “Giulio Natta”, Polytechnic of Milan, Via Mancinelli 7, Milan 20131, Italy

Rose, K.
Department of Structural Biology and Bioinformatics, University of Geneva, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Sanchez, J.-C.
Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, University of Geneva, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Walther, D.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Wilkins, M.R.
School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia

Williams, K.L.
BKG Group, P.O. Box 580, Avalon, NSW 2107, Australia

Zimmermann-Ivol, C.G.
Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, University of Geneva, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Contents

1 Ten Years of the Proteome

Marc R. Wilkins and Ron D. Appel

Abstract

The concept of the proteome is now over 10 years old. As with all anniversaries, it is a good time to look back and reflect on what has been achieved in the area that we now call proteomics. What has been done well? What has been done not-so-well? What has been achieved, and what still eludes us? This review will briefly explore some of these questions, with respect to protein separations, mass spectrometry, and proteomic bioinformatics.

1.1 Introduction to the Proteome

The editors of this book have been carrying out research and development in proteomics for more than 20 years. They developed techniques for the analysis of proteins and global protein expression (Williams et al. 1991; Hochstrasser et al. 1988) and software algorithms and tools for the interpretation of the results obtained using such analytical tools (Appel et al. 1988; Wilkins et al. 1995). While the idea of observing the protein expression of genomes in a holistic manner rather than one protein at a time arose with the advent of 2-D gels, the concept of the proteome itself was only introduced by Marc Wilkins in 1994 at a conference in Siena, Italy[1], having coined the term earlier that year in association with his then PhD supervisor Keith Williams. The first papers that began to use the term were published shortly thereafter (Wilkins et al. 1995; Wasinger et al. 1995), and the first book on proteomics was published in 1997 (Wilkins et al. 1997). Ten years has now passed since the publication of that first book, and as with all anniversaries, it is a good time to look back and reflect a little on what has been achieved in the area we now refer to as proteomics. What has been done well? What has been done not-so-well? What has been achieved, and what still eludes us? Here we will suggest answers to these questions. At the same time, we will comment on what we have sought to achieve in this book, and provide a brief précis on its contents.

[1]First Siena conference, 2D electrophoresis: from protein maps to genomes, 5–7 September 1994.

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

1.1.1 What's in a Word?

The words 'proteome' and 'proteomics' have been widely adopted by the biological community. In the 10 years since their introduction, their use has grown very rapidly (Fig. 1.1). In fact over 4,000 proteomics research and review articles were published in 2005. This has been fuelled by increasing numbers of journals that have arisen to serve the field, including *Proteomics*, *Proteomics-Clinical Application*, *Practical Proteomics*, *Journal of Proteome Research*, *Molecular and Cellular Proteomics*, *Proteome Science*, *Current Proteomics*, *Genomics and Proteomics*, *Briefings in Functional Genomics and Proteomics*, *Genomics Proteomics Bioinformatics* and *Expert Review of Proteomics*. In addition, proteomics research is increasingly published in a variety of other journals, so it has become established as a valuable means to obtain insight into the complexities of biological systems.

If we are simply measuring the progress of a field by its use of language, we might ask if the growth of proteomics is just a reflection of the so-called -omics revolution, or does it show a true growth in the field? The volume of work published in two other newer -omics areas, metabolomics and glycomics, is tiny by comparison, with 433 and 115 manuscripts having been published in 2005, respectively. Proteomics is clearly more widespread and established.

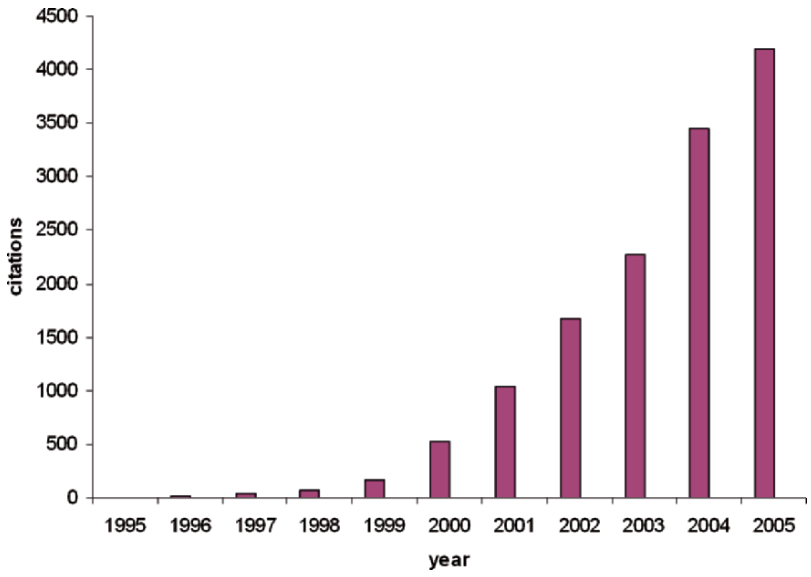


Fig. 1.1 Publications in the field of proteomics and proteome research have grown rapidly in the last 10 years. This was measured by querying the NCBI PubMed database for each year with the words 'proteome' or 'proteomics'. Note, however, that some articles may have been counted twice by this approach

1.1.2 Could Things Have Been Different?

So, would the world have been a different place had the term 'proteome' not been coined? Some commentators have argued that a combination of technical advances in separations technology (gel-based and chromatography-based), in mass spectrometry, and the explosion of information available from genome sequencing efforts have largely driven an increased interest in protein chemistry (Blackstock 2004).

While this is certainly true, it may be argued that the new language has brought renewed focus and legitimacy to protein chemistry that had previously been absent, largely due to the enormous shadow cast by genomics and other nucleic acid based approaches. The new language has also influenced biochemical thinking to move from a one-protein-at-a-time perspective to a more global view. Linguistically, it has been argued that thought cannot exist without language.[2] The proteome and proteomics are examples of this, as are other -omic words which were coined thereafter.[3] The new language and terminology has already helped a gamut of analytical technology to find its place in science and literature. New language in other fields will likewise legitimise emerging technology, focus thinking and also assist the funding of research in these areas.

1.2 Proteomics Is Technology-Driven

If we are to ask what has been done well in proteomics to date, one would have to pay particular attention to the development and dissemination of new technology. In a 10-year period, there have been a number of significant advances that, together, have transformed protein chemistry into the science of proteomics. Importantly, it has been a combination of conceptual breakthroughs and technical advances in separations techniques, mass spectrometry, protein chemistry and bioinformatics which have made this possible. The flood of nucleic acid sequence and genomic information, made available in sequence databases, was another essential co-requisite.

1.2.1 Protein Separations

Initially, proteomics researchers had a goal of visualising all proteins from a proteome on a single, or perhaps one acidic range and one basic range (2-D) polyacrylamide gel. This was happening in the late 1980s, and there was enormous excitement about the possibility of being able to see all

[2]Ferdinand de Saussure, Professor of Linguistics at Geneva University 1901–1913.
[3]See Chitty (2006) for an amusing list of new -omic words.

proteins in a proteome. However, it did not take long to realise that the separation and visualisation of all proteins from a proteome was not a straightforward task. In the mid-1990s, the availability of the first genome sequences and predicted proteomes allowed theoretical 2-D gels to be calculated, showing where each protein spot should be found (Urquhart et al. 1998). This revealed a bimodal distribution of proteins, with the majority of proteins having isoelectric point (pI) 4–6.5 and another group of proteins having pI 8–12. Most proteins had a mass of less than 100 kDa. The comparison of these theoretical maps with experimental 2-D gel separations immediately highlighted shortcomings with 2-D gels in that they were poor in resolving very acidic, very basic or very high mass proteins. A meta-analysis of proteins seen on 2-D gels and those predicted theoretically from genomes of *Escherichia coli*, *Saccharomyces cerevisiae* and *Bacillus subtilis* highlighted two additional issues (Wilkins et al. 1998). The first was that hydrophobic proteins were largely absent from the 2-D gels and that low-abundance proteins present at less than 1,000 copies per cell were likely to be undetectable, owing to limitations on the loading capacity and staining sensitivity of the 2-D gel process.

Since that time, a series of important technical advances have been made to help us see more proteins in the proteome. The latest advances associated with 2-D electrophoresis are discussed in Chap. 2. Broadly speaking, a number of strategies have been adopted. These include the running of narrow pI range gels to 'zoom in' on a particular region of the proteome, the fractionation of samples into either biologically (e.g. organelles) or physicochemically distinct fractions (e.g. membrane proteins) that can then be analysed appropriately, the enrichment or depletion of proteins of interest from a sample, along with new solubilisation and gel running techniques to assist in the analysis of the more difficult proteins. Importantly, fractionation has provided an avenue to load more of the relevant portion of samples of interest onto 2-D gels, thus assisting in the detection of lower-abundance proteins.

To completely bypass many of the challenges of working with complex mixtures of proteins, a conceptually different strategy emerged for protein analysis in proteomics. Called 'shotgun proteomics', probably inspired by the shotgun DNA sequencing approaches that were developed by Venter et al. (1998), it involves taking complex mixtures of proteins or indeed a whole proteome, and digesting all proteins to peptides with endoproteinases of known specificity. The resulting mixtures of peptides, which are physicochemically more homogenous than their parent proteins although greater in number, are then analysed using 2-D liquid chromatography and tandem mass spectrometry. Peptide fragment data are matched against sequence databases (Wolters et al. 2001) to determine the proteins present in a sample. Whilst this approach has limitations, notably the loss of protein isoforms (see Chap. 5), it provides an alternative to gel-based analyses for the separation and identification of large numbers of proteins from a proteome.

1.2.2 Mass Spectrometry

The last 20 years has brought astonishing advances in mass spectrometry technology. These advances have helped establish the science of proteomics. Mass spectrometers, whilst remaining expensive, now have remarkable mass accuracy and resolution, can analyse femtomolar quantities of peptides and proteins, and are increasingly automated. Two means of ionisation of proteins and peptides are in widespread use, electrospray ionisation and matrix-assisted laser desorption/ionisation, and these are teamed with a variety of mass analysers and detectors (see Chap. 3).

Mass spectrometers have all but superseded Edman degradation as the method of choice for protein identification. Two techniques, namely peptide mass fingerprinting and peptide fragmentation, can be used. Peptide mass fingerprinting has been used in a number of massive projects, for example more than 20,000 proteins were analysed as part of a large-scale analysis of yeast protein complexes (Gavin et al. 2002). However, peptide mass fingerprinting is losing favour to higher-confidence peptide fragmentation approaches that are able to fragment multiple peptides from the same protein. Nevertheless, it should be noted that mass spectrometers typically do not sequence peptides or proteins *per se*. They instead allow us to infer sequences by matching peptide fragmentation data against sequence databases. Routine *de novo* sequencing remains complex and is thus a work in progress (see Chap. 3).

In addition to protein identification, a myriad of new mass spectrometry approaches have been developed for the quantitative analysis of two or more samples. Such comparisons are of great scientific interest for the detection of biomarkers and the understanding of the multiplicity of changes that can occur when a proteome is perturbed by intrinsic or extrinsic forces. Previously, the comparison of protein expression in two or more samples was done by 2-D gel electrophoresis and computer image analysis (see Sect. 4.2). This approach has been successfully used in a large number of studies and remains widespread. The newer mass spectrometry based approaches are a significant advance and essentially use different stable isotopes to label proteins from two or more samples (Gygi et al. 1999). The samples are then mixed together and co-analysed. The high mass accuracy of the mass spectrometers allows the isotopic variants to be separated and relative quantitation to be undertaken. This concept has now been developed in a number of different ways (see Sect. 4.3) and whilst not perfect is providing a new means to undertake comparative analysis of two or more complex samples.

A final area in which mass spectrometry is now playing a major role is in the characterisation of proteins. Post-translational modifications of proteins are of increasing interest as they are key to the control and modulation of many processes inside the cell. Our recent appreciation of their roles in protein–protein interaction networks, whereby interactions between many proteins require the presence of certain post-translational modifications (Pawson and Nash 2003), is providing even greater impetus for their study.

Many sophisticated analytical strategies have been developed for the analysis of modifications (see Chap. 5) and these have now been applied, in some cases, on a proteome-wide scale. Protein phosphorylation has been a particular focus (Beausoleil et al. 2004). These analyses of modifications, whilst of large scale, remain incomplete. Yet they are giving the first glimpses of the dynamics of post-translational modifications in the proteome.

1.2.3 Making Sense of All the Data

New strategies for proteomic analysis and improvements in separation and analytical technologies have, without doubt, increased the amount and complexity of proteomic data. However, it is the combination of analytical approaches with sophisticated new bioinformatics that has allowed researchers to better generate, analyse, visualise and contextualise proteomic data and thus better understand their biological system under study.

Software for the quantitative analysis of protein expression on 2-D gels, particularly in association with new fluorescent stains, has drastically improved our capacity to find qualitative and quantitative expression differences between two gels or two populations of samples (see Chaps. 4, 6). Protein-identification software, vital to most aspects of proteomics, has incorporated statistical methods to allow identification confidence to be calculated. Bayesian and non-Bayesian statistics have been applied to the problem of protein identification by peptide mass fingerprinting (Perkins et al. 1999). For shotgun proteomics experiments, where thousands of protein identifications cannot possibly be verified by a human, searching against 'normal' and 'randomised' sequence databases is now used to estimate false-positive rates and thus overall identification confidence. The issue of protein-identification confidence has been the subject of much discussion, and proteomics journals have now released guidelines on protein identification which authors are expected to follow for their work to be published (Wilkins et al. 2006; Carr et al. 2004). In addition to improved strategies for protein identification, data-processing pipelines have been developed to automate the peak-picking and peak-matching processes for the hundreds to thousands of mass spectra that may be generated from the larger proteomics experiments. Workspaces have also been developed for the management and storage of the huge volume of data produced (Rauch et al. 2006).

Dramatic advances in the bioinformatics of post-translational modifications have also been made in recent years. Software tools for the discovery of protein modifications in mass spectrometry data are available, and are used for the analysis of peptide mass and peptide fragmentation data (see Chaps. 3, 5). Modifications such as methylation, acetylation, oxidation and phosphorylation can thus be found. The analysis of protein glycosylation, which produces enormously complicated mass spectrometry fragmentation spectra, is expected to become commonplace now that glycan structure databases and 'glycan mass fingerprinting' structure assignment tools have been developed.

The most profound advance in proteome bioinformatics has been its capacity to bridge the gap between technology and biology. Bioinformatics has been developed to allow the visualisation of cells and tissues after their direct laser scanning with mass spectrometry. This is a stunning new advance that is giving insight into the micro- and macroheterogeneity of protein expression in cells (see Chap. 6). In differential-display experiments, visualisation tools have become indispensable to highlight small changes that are undetectable when analysing each data item separately (see Chap. 6). A bioinformatics capacity to map all differentially expressed proteins onto the gene ontology also provides a 'big picture' understanding of the molecular function and biological processes that may be changed in association with a phenotype (see Chap. 7). It can reveal which changes in proteins may be functionally related. Where proteomic studies find differential expression of enzymes, the bioinformatic contextualisation of such proteins in the metabolome or 'reactome' (Reactome 2006) can reveal direct links between the proteome and metabolites in the cell. Bioinformatics is also allowing us to better understand the complexities of protein–protein interactions and interaction networks and how these change in association with disease (see Chap. 8). Figure 1.2, for example, shows the result of mapping protein function onto an interaction network. It is expected that these and other increasingly rich visualisations will assist in understanding the complexities of the proteome and the cell.

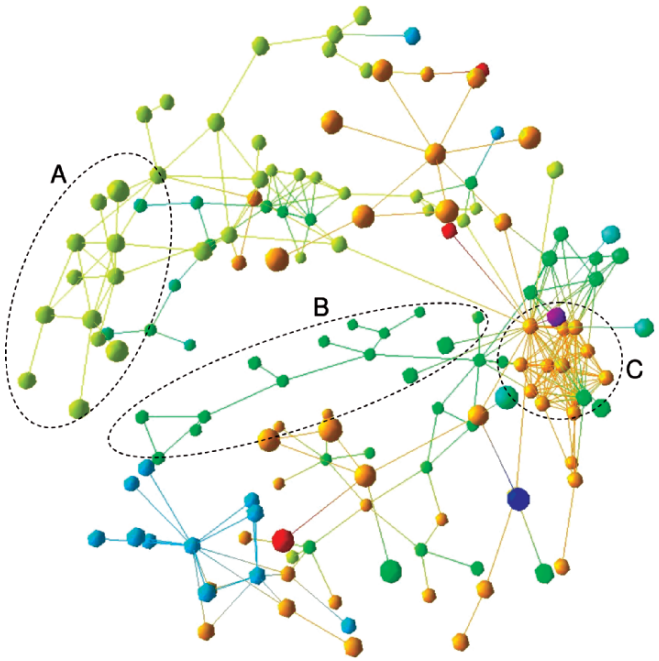


Fig. 1.2 Co-visualisation of protein–protein interaction and protein function. Some groups of directly interacting proteins have the same colour, indicating a common molecular function. Examples of molecular functions performed by such groups include *A* RNA binding (*yellow*), *B* structural molecule activity (*green*) and *C* protein binding (*orange*). (From Ho 2006)

1.3 What Has Proteomics Delivered?

A question we must always bear in mind when assessing emerging biomedical technology is what has it delivered to date or what is it likely to deliver? For proteomics, has the excitement associated with new methods translated into biological insights of scientific importance? This is a difficult question at the best of times, and since the technology has shifted the biological paradigm from a one-protein-at-a-time view to a new 'global' view, the question becomes almost impossible to answer without the benefit of the passage of time. However, it is clear that proteomics has already provided insight in a number of key areas. These may be enunciated as follows:

1. The proteome is no longer largely unknown. Substantial audits of protein expression, from gel-based studies coupled with mass spectrometry to those based on shotgun proteomics and tandem mass spectrometry, have given and will continue to give insight into which proteins are present in a particular cell or tissue. This is not to say we know the function of each protein in the proteome, but at a minimum we now have great insight into what proteins are present at any one time.
2. The 'higher order' of the proteome, obtained from large-scale studies of protein–protein interactions in the cell using proteomic techniques, is just starting to be revealed (see Chap. 8). The widespread adoption of this view will require another paradigm shift as it requires a global protein-based view of the cell and an acknowledgement that proteins do not act alone but participate in protein–protein interactions to form functional units in the cell.
3. Proteomics is providing a major new avenue for the discovery of medical biomarker proteins of diagnostic and/or prognostic significance. As proteomic technology is supremely well suited to the analysis of soluble proteins, the analysis of proteins from body fluids has been and will continue to be a fruitful endeavour. This is explored in detail in Chap. 9.
4. Proteomics is providing high-resolution data to supplement existing biomedical techniques. Toxicology, which has traditionally relied on histopathology and the evaluation of a small number of blood-associated proteins and metabolites, is using proteomics to better understand the effects and side effects of drugs (reviewed in Wilkins 2006). Immunoproteomics, the application of proteomics to the discovery of immunoreactive proteins and peptides, is starting to give stunning insight into how the body distinguishes self from non-self and what happens when this goes wrong (reviewed in Purcell and Gorman 2004).
5. Metaproteomics, a new term used to describe the shotgun proteomics analysis of mixtures of microbial species, is providing insights into microbial diversity and interactions that would otherwise be impossible to achieve. Microbial species that are difficult to culture in the laboratory can

be studied directly by a shotgun proteomics analysis of environmental samples. Whilst this currently requires a parallel metagenomic analysis (Venter et al. 2004) to allow protein identification, it is expected that this approach will become increasingly widespread.

1.4 What Still Eludes Us?

Finally, we may wish to ask which aspects of the proteome remain unexplored, and where has proteomics yet to be effectively applied? Whilst not an exhaustive answer to this question, the following points may be made:

1. The separation and detection of all proteins in the proteome remains a challenge. Low-abundance proteins are particularly elusive, owing to the large differences in concentration of proteins in many samples. Fractionation of samples can help with this, as may new 'equaliser' technology (see Chap. 2), but new approaches are still required to address this issue. Proteins that are very large, very basic and very acidic also remain problematic for 2-D gel analysis.
2. It is not possible to compare one interactome with another. The incredible complexity of the interactome, and the fact that interaction networks are built from the results of thousands of individual experiments, makes it impossible to currently compare one interaction network with another. Blue native electrophoresis, which separates large numbers of protein complexes under gentle conditions (Schagger 2001), may provide a means to address this.
3. *De novo* sequencing of proteins remains difficult. Researchers studying unusual organisms for which there is little nucleic acid sequence data cannot identify proteins of interest. They are also precluded from using shotgun proteomics techniques. De novo sequencing could address this issue; however, it remains a work in progress. Improvements in mass spectrometry and bioinformatics such as 'open-modification search' strategies (see Chap. 3) are required before this can become a robust and widespread technique.
4. We cannot monitor changes in the proteome in real time. The need to destroy cells for proteomic analysis, and a lack of alternative technology to mass spectrometry, makes it impossible to understand the myriad of changes that continuously occur in the cell. Whilst it is not clear how we may achieve such a feat, advances in high-magnification microscopy of living cells may prove to be a fertile ground for future developments.
5. Proteomics is currently semiquantitative, not quantitative. A capacity to undertake absolute rather than relative quantitation is desirable. Immunoassay techniques have been used to quantitate a large proportion of the *S. cerevisiae* proteome in copies per cell (Ghaemmaghami et al. 2003).

This has provided a first, although incomplete, molecular definition of a proteome. It has also been demonstrated that the spiking of samples with a known quantity of a labelled peptide enables quantitative analysis during mass spectrometry (see Sect. 4.3.6). Unfortunately, neither of these approaches presents a means for routine quantitative proteomics; thus, we await further technology developments to address this issue.

6. Our capacity to generate proteomic data of a common type and standard, and thus to have truly large, collaborative projects, is limited. Early approaches to federate 2-D gel databases and to share information on the Internet (Appel et al. 1996) partially addressed this issue. However, because most laboratories use different 2-D analysis and mass spectrometry techniques and because most proteomics data management platforms and analysis software use different file formats or incompatible algorithms, the direct comparison of 2-D maps from laboratory to laboratory or the sharing of data from shotgun proteomics experiments remains difficult. The large differences in data seen between laboratories in multicentre Human Proteome Organisation (HUPO) projects has also been worrying. Data standards, such as those designed by the HUPO Proteomics Standards Initiative (Human Proteome Organisation 2006) for molecular interactions, mass spectrometry, gel electrophoresis, proteome informatics or protein modifications, provide a means of sharing data (see Chap. 7). However, these are of little use if analytical techniques from laboratory to laboratory provide inconsistent results. Together, this is a major impediment to the field.
7. The field remains slow to embrace advanced statistics for experimental design and data analysis. For biomarker studies, this is of serious consequence and it is likely that many biomarker studies have analysed too small a number of patients, or have analysed data with insufficient statistical rigour for the outcomes to be reliable (Hunt et al. 2005). For other studies, the biological interpretation of results may be difficult unless statistical techniques are used to find what differences in protein expression are of significance. In the same way that the field of gene-expression analysis had to adopt statistics to understand the complexities of microarray data, proteomics must increase its use of statistics. The tools for this, in most cases, are available but there is an unwillingness on the part of researchers to take the necessary steps. It may only be through increasingly stringent peer-review processes that this can be encouraged to happen.
8. The interpretation of data, that is the extraction of knowledge from the large volume of data produced by analytical proteomics experiments, remains a major bottleneck. In particular, proteomics experiments produce up to thousands of tandem mass spectrometry spectra per day that are computationally matched to protein or peptide sequences in databases. Whilst several identification programs are routinely used, they are not a complete solution to this problem. First, several aspects of protein identification are handled manually. Spectra must be fed into identification software and many identification parameters must be set. Often those

parameters are set only once for a given experiment, while spectra vary greatly in quality and content. Expected post-translational modifications of peptides must be selected, which excludes the chance to detect those that might be novel. Results must be visually validated, a step too often neglected by researchers owing to overwhelming complexity and lack of time. Second and most importantly, many (and often a majority of) tandem mass spectrometry spectra cannot be identified at all. The possible causes are numerous, such as the spectra being of insufficient quality, incorrect use of search parameters, peptides carrying unexpected modifications or mutations, sequences in the database containing errors, or inadequate mass-error tolerance. New algorithms are required to identify and characterise at least a larger proportion of mass spectrometry data. Finally, while bioinformatics for protein identification has been the focus of much research and development in recent years, we have no means to validate and compare the results obtained using the many identification tools available.

1.5 This Book and Some Conclusions

This book is about the concepts, techniques and practice of proteomics. We have sought to capture the state-of-the-art thinking, to provide some perspectives gained with the benefit of hindsight and to provide some views on what the proteomic future might hold. It is not a methods book *per se*, but there is an emphasis on the ever-changing nature of technology. This technology continues to provide new avenues for investigation and thus the capacity to generate new biological insights. As with our first book on proteomics (Wilkins et al. 1997), we think this book provides a timely milestone for the field and a point of reference for the future. We hope you find proteomics and this book as exciting as we do.

Acknowledgements. Keith Williams and Denis Hochstrasser are visionary individuals who have had a profound influence on our thinking and have given us remarkable opportunities and sage advice at key points in our careers. We are fortunate to have had the chance to work with them for such a long time, and hope to continue to collaborate with them in the future.

References

Appel R, Hochstrasser D, Roch C, Funk M, Muller AF, Pellegrini C (1988) Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. Electrophoresis 9:136–142

Appel RD, Bairoch A, Sanchez JC, Vargas JR, Golaz O, Pasquali C, Hochstrasser DF (1996) Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data. Electrophoresis 17:540–546

Beausoleil SA, Jedrychowski M, Schwartz D, Elias JE, Villen J, Li J, Cohn MA, Cantley LC, Gygi SP (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc Natl Acad Sci USA 101:12130–12135

Blackstock W (2004) A surfeit of '-omics'. Drug Discov Today Targets 3:125–127

Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A (2004) The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. Mol Cell Proteomics 3:531–353

Chitty M (2006) -Omes and -omics gloassary. http://www.genomicglossaries.com/content/omes.asp. Cited 1 Nov 2006

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

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003) Global analysis of protein expression in yeast. Nature 425:737–741

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17:994–999

Ho E (2006) Visualization of protein-protein interaction networks using GEOMI. Honours thesis, University of New South Wales, Sydney

Hochstrasser DF, Harrington MG, Hochstrasser AC, Miller MJ, Merril CR (1988) Methods for increasing the resolution of two-dimensional protein electrophoresis. Anal Biochem 173:424–435

Human Proteome Organisation (2006) HUPO Proteomics Standards Initiative. http://psidev.sourceforge.net/. Cited 1 Nov 2006

Hunt SM, Thomas MR, Sebastian LT, Pedersen SK, Harcourt RL, Sloane AJ, Wilkins MR (2005) Optimal replication and the importance of experimental design for gel-based quantitative proteomics. J Proteome Res 4:809–819

Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300:445–452

Perkins DN, Pappin DDJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567

Purcell AW, Gorman JJ (2004) Immunoproteomics: mass spectrometry-based methods to study the targets of the immune response. Mol Cell Proteomics 3:193–208

Rauch A, Bellew M, Eng J, Fitzgibbon M, Holzman T, Hussey P, Igra M, Maclean B, Lin CW, Detter A, Fang R, Faca V, Gafken P, Zhang H, Whiteaker J, States D, Hanash S, Paulovich A, McIntosh MW (2006) Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. J Proteome Res 5:112–121

Reactome (2006) Cold Spring Harbor Laboratory, Cold Springer Harbor, The European Bioinformatics Institute, Cambridge, and The Gene Ontology Consortium. http://www.reactome.org. Cited 1 Nov 2006

Schagger H (2001) Blue-native gels to isolate protein complexes from mitochondria. Methods Cell Biol 65:231–244

Urquhart BL, Cordwell SJ, Humphery-Smith I (1998) Comparison of predicted and observed properties of proteins encoded in the genome of Mycobacterium tuberculosis H37Rv. Biochem Biophys Res Commun 253:70–79

Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M (1998) Shotgun sequencing of the human genome. Science 280:1540–1542

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O,

Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso sea. Science 304:66–74

Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL, Humphery-Smith I (1995) Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. Electrophoresis 16:1090–1094

Wilkins MR (2006) How proteomics can assist in the detection and avoidance of adverse drug reactions. Transfus Med Hemother 33:97–105

Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL (1995) Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev 13:19–50

Wilkins MR, Williams KL, Appel RD, Hochstrasser DF (eds) (1997) Proteome research: new frontiers in functional genomics. Springer

Wilkins MR, Gasteiger E, Sanchez JC, Bairoch A, Hochstrasser DF (1998) Two-dimensional gel electrophoresis for proteome projects: the effects of protein hydrophobicity and copy number. Electrophoresis 19:1501–1505

Wilkins MR, Appel RD, Van Eyk JE, Chung MC, Gorg A, Hecker M, Huber LA, Langen H, Link AJ, Paik YK, Patterson SD, Pennington SR, Rabilloud T, Simpson RJ, Weiss W, Dunn MJ (2006) Guidelines for the next 10 years of proteomics. Proteomics 6:4–8

Williams KL, Gooley AA, Haynes PA, Batley M, Curtin JH, Stuart MC, Champion AC, Sheumack DD, Redmond JW (1991) Analytical biotechnology: applications for downstream processing. Aust J Biotechnol 5:96–100

Wolters DA, Washburn MP, Yates JR 3rd (2001) An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem 73:5683–5690

2 Sample Preparation and Prefractionation Techniques for Electrophoresis-Based Proteomics

Ben R. Herbert, Pier Giorgio Righetti, Attilio Citterio, and Egisto Boschetti

Abstract

In the 1990s much of the technology development for 2-D gels was centred on attempts to solubilise and separate whole proteomes. In the last 10 years, much has changed in the science of sample preparation, but the art of 2-D gels is almost the same, and remains a method that many love to hate. The improvements in protein solubilisation, the introduction of pH gradients that extend into the far alkaline region, and particularly the hunt for low-abundance proteins have served to hasten the arrival of the most important current area of sample preparation – fractionation. As proteomics matures, there is a new focus on function and the analysis of co- and post-translational modifications. With this in mind, and as the complexity of sample preparations increases, this chapter deals with possible artefacts and how to avoid generating them. Finally, with artefacts banished, this chapter reviews the most appealing prefractionation tools that are currently available. Fractionation is discussed as a formidable tool for mining below the tip of the proteomic iceberg.

2.1 Introduction

In the last 10 years much has changed in the science of sample preparation. On the other hand, the art of 2-D gels is almost the same, and remains a method that many love to hate. In reality, 2-D gels provide unparalleled protein resolution and enable a frozen-in-time view of the proteome. Importantly, this includes differentially modified and processed versions of many proteins.

One of the key reasons that researchers dislike 2-D gels is the difficulty of good sample preparation. As a consequence, poor sample preparation is the cause of most poor-quality 2-D gels. To get the best from 2-D gels, one must understand the sample that is being studied, know how to extract the desired proteins and prepare them for the 2-D gel in the most appropriate way. If you cannot solubilise, you cannot analyse! This is a daunting challenge as the protein diversity within even the simplest proteome cannot be captured by

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

any single extraction and separation step, and consequently the wealth of literature on sample preparation is remarkably diverse. In this chapter we will focus on the changes and improvements in sample-preparation technology over the last decade and summarise the emergence of a major trend in this area. The improvements in protein solubilisation, the introduction of pH gradients that extend into the far-alkaline region, and particularly the hunt for low-abundance proteins has served to hasten studies on the most important current area of sample preparation – fractionation. Researchers found that having more of the pH spectrum available on immobilised pH gradients (IPGs) and greater protein solubility caused even more crowding of protein spots on 2-D gels. To combat this, many forms of fractionation have been adopted to reduce sample complexity and thus enable deep mining of the low-abundance regions of the proteome.

In the 1990s much of the technology development for 2-D gels was centred on attempts to get whole proteomes solubilised and separated. The target proteins were those with alkaline isoelectric points, those of high or low mass or those that were membrane-associated. These groups of proteins were under-represented on 2-D gels and many researchers worked to rectify that. Previously (Herbert et al. 1997), we discussed these issues and summarised the state of the art in solubilisation reagents and alkaline IPG technology. Another broad group of proteins which came to dominate 2-D gel discussions during the late 1990s was the low-abundance proteins. Many of the discussions about detecting low-abundance proteins assumed no fractionation, and considered only conventional sample preparation. This scenario is a most difficult one in theory, and certainly in practice. With very high (10-mg) protein loads on large-format 2-D gels, we showed theoretically that proteins present at 1,000 copies per cell could be detected with silver stain, but all of these proteins were probably not analysable by mass spectrometry. In practice, most researchers load less than 1 mg of protein on a single 2-D gel and the limitations of sample preparation mean that the chances of observing low-abundance proteins with a one-extract, one-gel approach are practically nil.

2.2 Conventional Sample Preparation

Sample preparation for isoelectric focusing (IEF) relies on non-ionic or zwitterionic reagents to disrupt protein complexes and denature proteins. This ensures that only polypeptide monomers are subjected to the subsequent electrophoretic separation. Because IEF separates proteins on the basis of isoelectric point, the single most powerful solubilising reagent, the anionic surfactant sodium dodecyl sulfate (SDS), is not normally used as it strongly attaches to proteins and may cause anomalous focusing and horizontal streaking in IEF. In order to approximate the denaturing power of boiling

proteins in SDS under reducing conditions, IEF practitioners have relied on various cocktails of chaotropes, surfactants and reducing agents. Chaotropes, urea being the most common for IEF, disrupt the hydrogen-bonding at the protein surface and cause partial unfolding. When water forms hydrogen bonds to the chaotropes rather than to the proteins, a protein is more likely to denature and expose its (often hydrophobic) interior. Once the hydrophobic interior of the protein is exposed, the solubility is often compromised in aqueous solution, thus the requirement for surfactants, more commonly called detergents. It is normal to have at least one surfactant present in the IEF cocktail to help solubilise the hydrophobic residues that are exposed as a result of denaturation in chaotropes. Even small amounts of ionic substances (e.g. 20 mM) are not generally compatible with steady-state IEF; thus, the use of strong ionic detergents such as SDS is not recommended. As such, we are restricted to the use of non-ionic or zwitterionic detergents. Traditionally, non-ionic surfactants such as Triton X-100 and octyl glucoside have been used. More recently these have been superseded by the sulfobetaine class of surfactants such as 3-[(3-cholamidopropyl)dimethylammonio]-1-propane-sulfonate (CHAPS), an amido-sulfobetaine (Gianazza et al. 1987). The final reagent type in the classical IEF sample cocktail is a reducing reagent, which breaks disulfide bonds to enable complete protein unfolding and denaturation. The two main types of reducing agents used are the free-thiol reagents such as mercaptoethanol and dithiothreitol (DTT) and the phosphines, a group of trivalent phosphorous compounds. The traditional free-thiol compounds are used at high concentrations (20–100 mM) and work by displacing the equilibrium towards the breakage of disulfides. However, the reagents are charged at alkaline pH and hence reducing conditions are almost impossible to maintain during IEF with the free-thiol reagents. The non-charged phosphines, such as tributyl phosphine, provide improved reducing conditions and thus improved focusing for some samples (Herbert et al. 1998). However, even the phosphines fail to provide reducing conditions for the overnight run-times required for focusing to equilibrium in IPGs. The ultimate sample preparation method for disulfides is to alkylate the reduced cysteines prior to electrophoresis, thus avoiding any requirement for reducing conditions in the IEF. Chemistries associated with cysteine are widely exploited in proteomics; however, they have a number of pitfalls which will be discussed in this chapter.

Since 1996 a number of publications have reported and reviewed the use of novel reagents such as thiourea and new sulfobetaine surfactants, which improve protein solubilisation prior to IEF (Chevallet et al. 1998). Thiourea at 2 M, in combination with urea at 7 M produces a far more chaotropic sample solution than the conventional 8 M urea. However, the increased chaotropic power required a new class of surfactants to cope with the highly denaturing environment. The group of Thierry Rabilloud developed a range of new chaotrope-tolerant surfactants, the best of which, amido-sulfobetaine 14 (ASB-14) and C7bZ0 (Vuillard et al. 1995; Rabilloud et al. 1997) in

combination with urea and thiourea, provide what is currently the highest level of solubilising power for IEF. The increased solubility with these new reagents and separation on narrow-range commercial IPGs has significantly increased the total number of resolvable proteins from a proteome. However, the increased numbers of proteins solubilised from a single sample makes it difficult to resolve all proteins on a single 2-D gel. Complexity reduction via fractionation is essential.

Fractionation has assumed such importance in proteomics that the majority of this chapter will be devoted to it, especially methods which are used in conjunction with downstream electrophoretic separations. However, it must be kept in mind that each additional fractionation step means increased handling of the sample and thus a greater potential for artefactual modifications, protein losses and results which can be difficult to interpret.

2.3 Artefacts

As proteomics matures, there is a new focus on the analysis of co- and post-translational modifications. Modifications are one mechanism for the generation of complexity in the proteome, given the relatively small number of genes in the genomes of higher eukaryotes (see Chap. 5). Thus, separation scientists are under pressure to maximise resolution of modified proteins, while ensuring that artefactual protein modifications are eliminated or at least minimised during protein separation by 2-D electrophoresis (2-DE).

2.3.1 Cysteine Chemistry – Reduction and Alkylation

In standard 2-D gel procedures, reduction is first undertaken before the IEF/IPG separation. This is followed by a second reduction and alkylation as part of the equilibration step between first- and second-dimension separation, in preparation for SDS polyacrylamide gel electrophoresis (PAGE). Owing to incomplete reduction during the IEF, this protocol is far from being optimal and often results in a large number of spurious multimeric spots due to ‘scrambled’ disulfide bridges among like and unlike protein chains. Because of the negative charge on the –SH group of typical reducing agents such as DTT, these compounds are self-buffered and will migrate inside the pH gradient towards the anode. They will be arrested by protonation at around pH 7. Thus, artefacts arising from incomplete reduction are more often observed in the alkaline portion of the IPG. Even tributyl phosphine, a strong non-thiol reducing agent, does not appear to have the reducing power to maintain all proteins as monomeric polypeptides during IEF. We have shown that the number of protein spots arising as artefacts can be impressively large, even in the case of simple polypeptides such as the human

α-globin and β-globin chains, which possess only one (α-) or two (β-) –SH groups (Herbert et al. 2001). Similar results, supporting the use of alkylation prior to IEF, have recently been published (Luche et al. 2004). This work indicated that the alkylation of cysteine decreased horizontal streaking and greatly increased resolution, especially in the basic region of the 2-D gels.

Poor alkylation efficiency can occur when using iodoacetamide as the alkylating agent unless care is taken to add iodoacetamide as a powder to the sample solution or it is dissolved in plain water and added it to the sample just at the moment of alkylation. This problem with efficiency was traced back to the presence of thiourea, which acts as a scavenger of iodoacetamide in the sample. If iodoacetamide is dissolved in the solubilising mixture containing thiourea, but in the absence of sample proteins, it will be destroyed quite rapidly by thiourea (Galvani et al. 2001a). In addition, prolonged alkylation reactions (about 24 h) with iodoacetamide give rise to modifications of lysine and other amino acids, such as methionine (Galvani et al. 2001a, b).

To simplify the methodology of reduction and alkylation, a very appealing option is to use a reagent with an activated double bond (e.g. acrylamide) as an alkylating reagent instead of iodoacetamide (Herbert et al. 2001; Galvani et al. 2001a, b). Iodoacetamide and acrylamide have similar reactivity rates with ionised –SH groups at the alkaline pH values at which alkylation is customarily performed. Acrylamide does not react with tributyl phosphine, so the reaction can be done in a single step, which results in a significant saving in time. Furthermore, acrylamide does not seem to react with thiourea or with amino acid groups other than cysteine.

2.3.2 Cysteine Chemistry – β-Elimination

Aside from preventing re-formation of disulfides, cysteine alkylation also has the benefit of preventing artefactual β-elimination of the thiol during electrophoresis. β-Elimination or desulfuration of thiols, which results in the loss of an H_2S group (34 Da) in cysteines from proteins focusing in the alkaline pH region, has recently been reported (Steen and Mann 2001; Herbert et al. 2003a). Confirmation that β-elimination occurs was obtained by trapping an alkaline protein, lysozyme, at alkaline pH in an electric field using a multi-compartment electrolyser (MCE; Sect. 2.5.3). To closely simulate the IEF[V1] conditions for alkaline proteins in IPGs, lysozyme was loaded into the alkaline pH 8.0—11.0 chamber of a MCE and trapped in the electric field for 48 h. A control sample under static, non-electric-field conditions, was solubilised in 8 M urea, sodium borate pH 9.0 and maintained at a temperature equivalent to that of the MCE sample for 48 h. Aliquots taken from the MCE-trapped sample during the incubation period and analysed by matrix-assisted laser desorption/ionisation (MALDI) mass spectrometry show a decrease in mass that corresponds to the loss of thiol groups (34 Da). As shown in Fig. 2.1 the initial sample produces one clean peak corresponding to 14,314 Da and

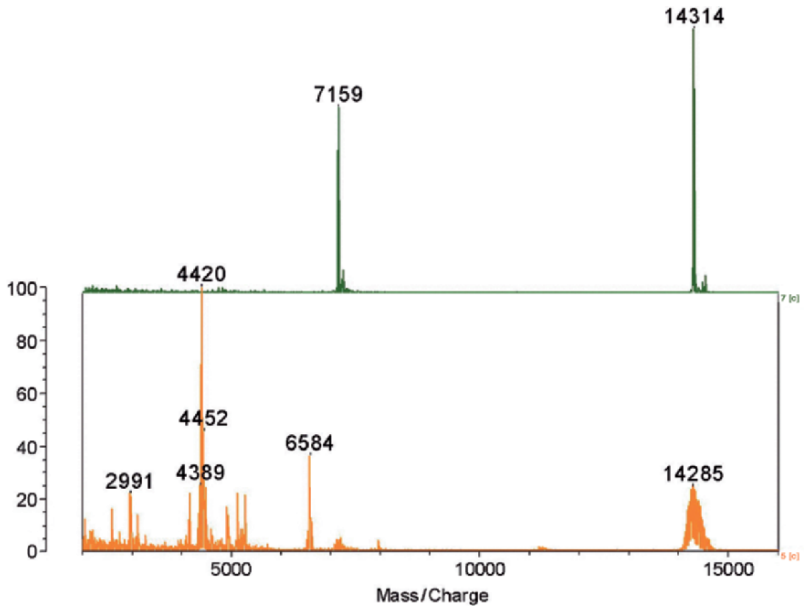


Fig. 2.1 Desulfuration (β-elimination) of lysozyme in an electric field. Mass spectra of control lysozyme (*upper two peaks*, representing the single and doubly charged species) and lysozyme exposed for 48 h to the electric field. Note, in this last case, the massive sample degradation into a series of fragments, with a massive disappearance of the intact lysozyme peak (M_r=14,314). (From Herbert et al. 2003, with permission)

a doubly charged ion of 7,159 Da. After 48 h in the electric field, there has been considerable degradation of the lysozyme and only a very small amount of the original protein remains intact. When the same experiment is repeated with alkylated lysozyme, as shown in Fig. 2.2, there is a small amount of degradation but the majority of the starting material is still present after 48 h in the electric field.

The β-elimination described above seems to be a phenomenon highly accelerated by the electric field. In fact, it was also reported by Steen and Mann (2001) when analysing peptides with a QSTAR Pulsar quadrupole time-of-flight (TOF) tandem mass spectrometer equipped with a nanoelectrospray ion source. At relatively low collision energies (55 V) they noticed loss of sulfenic acid as a major fragmentation pathway of the peptides under analysis.

2.3.3 Lysine Chemistry – Carbamylation

Urea is the most commonly used chaotropic agent and, increasingly, thiourea/urea combinations are used to exploit the improved denaturing ability of thiourea. In solution, urea is in equilibrium with ammonium

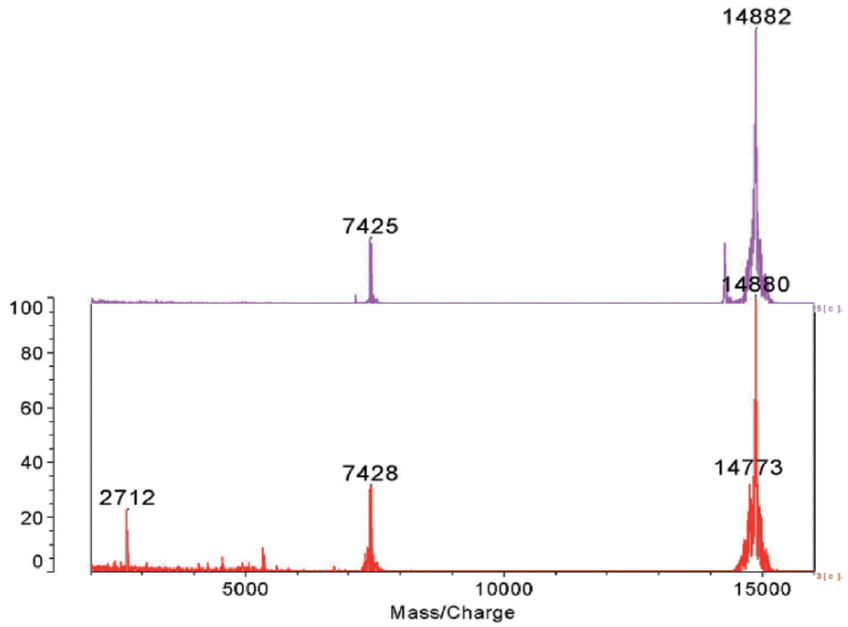


Fig. 2.2 As for Fig. 2.1, but with the sample fully alkylated. Note, in this last case, the maintenance of sample integrity, with negligible formation of lysates at lower M_r values. (From Herbert et al. 2003, with permission)

cyanate (Shapiro 1999). If a cyanate scavenger is present, such as the ε-amino group of lysine, the cyanate will react until either the cyanate or the scavenger is consumed (Anderson and Hickman 1979).

One potential drawback of the reduction and alkylation procedure described in Sect. 2.3.1 is the fact that the reaction requires an alkaline pH to proceed. Unfortunately, the alkaline pH also ensures that lysine residues are deprotonated and reactive towards isocyanate. Under conditions of no electric field, the cyanate is free to react with lysine and cause artefactual modifications. The basic recommendation for sample preparation is to minimise the extraction time and ensure that the sample is kept as cool as possible (below 37°C) without causing urea precipitation during extraction and subsequent storage. Extracts containing urea should always be stored at 4°C for short periods and frozen for long-term storage. Despite the chemical potential for carbamylation, it was clearly shown by McCarthy et al. (2003) that the progression of the reaction is quite slow and carbamylation was not detectable during 12-h incubation at 20°C. However, extended sample-preparation procedures at temperatures above 40°C or prolonged storage at 20°C or above will result in permanent artefactual modifications.

Under IEF conditions where an electric field is applied, the kinetics of carbamylation are quite different from the situation described above. During electrophoresis, the charged products of urea degradation are rapidly transported to the electrodes, providing minimal opportunity for them to react with amino groups on proteins and peptides. McCarthy et al. (2003) electrolysed a myoglobin peptide in 8 M urea in a MCE at alkaline pH for 48 h. No peaks corresponding to the addition of cyanate (43 Da) (IonSource 2007) were observed in the myoglobin peptide trapped in the MCE, even after 48 h of electrolysis in the alkaline chamber. As the only buffering species present, the pH of the alkaline MCE chamber was close to 9.3, defined by the pI of the peptide.

2.4 Multiplexed Approaches to Proteomics

Historically, proteome research has used three different approaches for the analysis of complex protein mixtures. First, and historically the original form of what is described as proteomics, is gel-based protein separation coupled with mass spectrometry analysis of the separated proteins. Second, and what has often been seen as a competing approach, is a range of liquid chromatography based protein or peptide separations and subsequent online electrospray mass spectrometry analysis of separated proteins or peptides. Third is protein chip-based approaches where interactions between molecules are exploited using bait molecules on solid supports to pull out proteins of interest – often followed by mass spectrometry analysis of purified or bound proteins. Evolutionary changes are occurring to these technologies to extend their use into the areas of functional protein analysis, native protein separation and the analysis of protein complexes and interactions.

Proteomics researchers are increasingly blurring the boundaries between these technology platforms as it is increasingly clear that no single sample preparation, fractionation or separation technology can provide a comprehensive view of any proteome. In fact, the different separation and protein chip technologies have particular biases towards certain classes of proteins. False conclusions can thus be drawn from experiments if the data analysis does not consider the bias.

The current best practice in proteomics involves accessing at least two independent methods for sample preparation and protein separation. Ideally, the separated proteins are analysed by different mass spectrometry techniques. This is highlighted by Breci et al. (2005) and in Figs. 2.3–2.5. All illustrate the point that different subsets of proteins are identified with each of the techniques used. An analysis of the merits of different mass spectrometry approaches is beyond the scope of this chapter. This is explored in Chaps. 3 and 4. Next we discuss a variety of sample-preparation and fractionation tools that can be used alone or in combination in proteome research.

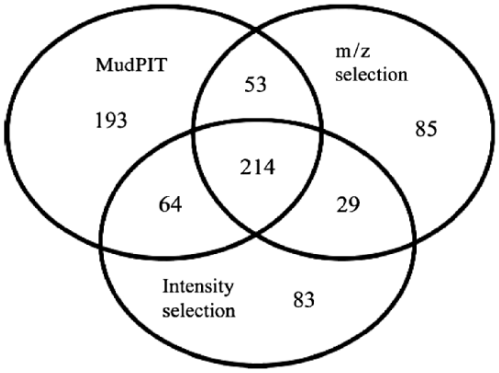


Fig. 2.3 Venn diagram showing overlap of datasets for protein identification using three different liquid chromatography (LC) and mass spectrometry (MS) based peptide fractionation approaches as indicated. (From Breci et al. 2005, with permission)

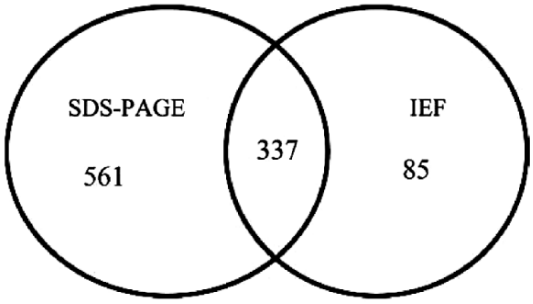


Fig. 2.4 Venn diagram showing overlap of datasets for protein identification using two different gel electrophoresis based protein fractionation approaches, as indicated. *SDS-PAGE* sodium dodecyl sulfate, *IEF* isoelectric focusing. (From Breci et al. 2005, with permission)

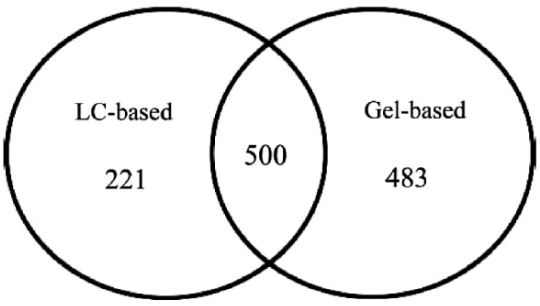


Fig. 2.5 Venn diagram showing overlap of datasets comparing the combined three different LC and MS based peptide fractionation approaches and the combined two different gel electrophoresis based protein fractionation approaches, as indicated. (From Breci et al. 2005, with permission)

2.5 Prefractionation Tools

The major limitations of available technologies for the complete investigation of any proteome have been highlighted in dozens of papers. Classical 2-DE techniques poorly represent strongly alkaline proteins, and they are not effective for the solubilisation, analysis and identification of highly hydrophobic proteins. However, since our first reviews of this area (Herbert et al. 1997), dramatic improvements in the analysis of both hydrophobic and alkaline proteins have been made. A greater appreciation has also developed of the need for IEF for the separation of protein isoforms (see Chap. 5). For polypeptides of masses lower than 5,000 Da, by contrast, mass spectrometry is the analytical tool of choice. All of the above technologies have their limitations for complex proteomic studies.

Prefractionation, with all its variants, makes possible deeper mining of the proteome. A major consideration, with any prefractionation protocol, is that it must interface with downstream processing steps. For example, high-salt prefractionation techniques are not compatible with IEF or liquid chromatography –mass spectrometry analysis. Here we discuss the most appealing prefractionation tools that are currently available. For in-depth analysis of each of the various methods, the reader is referred to a number of our recent reviews (Righetti et al. 2001, 2003, 2005a, b; Herbert et al. 2003b, 2004; Hamdan and Righetti 2005).

2.5.1 Fractional Centrifugation

One of the oldest and still most appropriate methods for simplifying a cellular proteome is the separation of cell organelles and substructures by differential centrifugation. This method, as developed in the late 1950s and early 1960s by de Duve (1971) and others is well ingrained in classical biochemical analysis. Via the use of different centrifugal forces, one can isolate subcellular organelles, such as nuclei, mitochondria, lysosomes, peroxisomes, synaptosomes, microbodies and the like in a reasonably pure form. Clearly, if one is investigating the proteome of such organelles, this is the most direct and proficient method for enrichment of the desired protein fractions as they are part of specific substructures of the cells. Such methods have recently been rediscovered and applied to proteome analysis in a number of reports. For example, differential centrifugation has been applied to the isolation of nuclei. In the case of human liver nuclei, a 2-DE reference map has been established and displays as many as 1,497 spots (Jung et al. 2000). Even nucleoli have been further subfractionated and their reference maps reported, displaying approximately 350 spots (Andersen et al. 2002). With new emphasis on the importance of the mitochondrion and its association with oxidative stress and aging, the analysis of the mitochondrial proteome has become a significant field (Douette and Sluse 2006).

2.5.2 Chromatographic Techniques

Solid-phase chromatographic techniques are tools of considerable interest for proteome analysis as they are based on a very large choice of solid-phase adsorbents. Low or very high selectivity can be implemented just by selecting appropriate sorbents and suitable conditions of adsorption/elution. Adsorbents can be used as a single separation step or in combination as 2-D or multidimensional liquid chromatography.

2.5.2.1 General Chromatographic Methods

Fountoulakis's group (Fountoulakis et al. 1998, 1999a, b; Fountoulakis and Takacs 1998) has developed and described chromatography as a way to simplify the complexity of the proteome. In a first procedure, affinity chromatography on heparin gels was adopted as a prefractionation step for enriching certain protein fractions in the bacterium *Haemophilus influenzae*. Immobilised heparin was deemed suitable for enriching low-copy-number gene products, and about 160 cytosolic proteins bound with different affinities to the heparin matrix to be highly enriched prior to 2-DE. As a result, more than 110 new protein spots were identified, increasing the total number of identified proteins of *H. influenzae* to more than 230. In a second approach, the same lysate of *H. influenzae* was prefractionated by chromatofocusing on polybuffer exchanger. Approximately 125 proteins were identified from the eluate collected from the column, the majority of them being low-abundance enzymes. Thus, with this additional step, a total of 300 proteins could be identified in *H. influenzae* by 2-DE analysis, out of a total of approximately 600 spots visualised on such maps from the soluble fraction of this microorganism. In yet another approach, the cytosolic soluble proteins of *H. influenzae* were prefractionated by hydrophobic interaction chromatography on a phenyl column. In total, with all the various chromatographic steps adopted, the number of proteins identified could be increased to 350.

In another example, hydroxyapatite chromatography was used for the enrichment of low-abundance proteins from *Escherichia coli* (Fountoulakis et al. 1999b). Of the 4,289 possible gene products of *E. coli*, about 1,200 protein spots could be seen on a typical 2-DE map when about 2 mg of protein was separated. By use of the hydroxyapatite chromatography, approximately 800 spots, corresponding to 296 different proteins, were identified. About 130 new proteins that had not been detected in 2-D gels of the total extract were seen for the first time. This chromatographic step, though, was not effective in enriching low-abundance proteins. Instead, it enriched low-mass proteins, such as cold-shock proteins. In a further approach, reactive dye column chromatography was used to enrich *E. coli* low-abundance proteins (Birch et al. 2003). Six reactive dyes were investigated for their potential

to enrich different categories of proteins. The authors demonstrated that reactive dye chromatography was a suitable method for fractionation of complex mixtures.

Other chromatographic tools can also be used for investigation of the proteome. Metal chelate affinity chromatography can be an effective means to separate proteins with exposed histidines (Smith et al. 2004). This method can also be applied to calcium-binding proteins or the separation of phosphoproteins (Lopez et al. 2000; Ficarro et al. 2002). Additionally, lectin-affinity adsorption is extensively used for the separation of glycoconjugates and glycoproteins (Ghosh et al. 2004).

2.5.2.2 Sample Fractionation with Stacked Sorbents

An unusual type of chromatographic prefractionation involves the concomitant use of several columns, with a single buffer for adsorption and a single buffer for elution (Guerrier et al. 2005). The columns have different, serially connected solid-phase chemistries which can all be equilibrated with the same binding buffer. The protein sample is applied to the first column and it crosses all chemistries, each of them capable of capturing a certain group of proteins. When the sample reaches the bottom-most column, most if not all of the proteins have been removed. Once protein adsorption is complete, the sequence of columns is washed with a physiological saline, and then the columns are disconnected. Groups of proteins can then be eluted from each column. This is performed by using a buffer that is compatible with downstream analytical methods such as mass spectrometry and 2-DE. There is thus only one flow-through fraction for the whole experiment. The entire separation has to be done under non-overloading conditions. By choice of orthogonal chemistries, their sequence, the loading volume, and by proper selection of the binding buffer, the bound proteins should be quite different in each of the capturing columns. Providing the columns do not saturate, this chromatographic separation should virtually eliminate protein redundancy (the same protein found in different fractions). It should be noted that with a number of different binding chemistries, care must be taken not to irreversibly deplete certain proteins. When n chemistries are used, $n + 1$ fractions are typically obtained. Eight fractions were recovered from a seven-bed column separation of human serum and the number of protein species observed by mass spectrometry was twofold greater than with classical unimodal/multistep elution chromatography (Guerrier et al. 2005).

2.5.3 Electrophoresis-Based Methods

Preparative electrophoretic methodologies (Righetti et al. 1992), in all of their many forms, have never enjoyed a high popularity compared with their

chromatographic counterparts. However, electrokinetic methodologies have recently experienced a strong revival, owing to their remarkable performance as 'mining tools' for proteome analysis. Some of the major techniques are reviewed next, but it should be kept in mind that the preferred tools will be those that exploit focusing techniques, since the resulting fractions can be directly interfaced with 2-DE.

2.5.3.1 Continuous Electrophoresis in Free Liquid Films

Compared with gel-based separations, this technique has two main advantages: much higher sample loads can be applied, and certain artefactual protein modifications induced by free acrylamide monomers in polyacrylamide gels are eliminated. Present equipment derives from the concepts and instrumentation of Hannig (1978), in which an electrolyte solution flows in a direction perpendicular to an electric field and the mixture undergoing separation is added continuously at a small sample inlet in the flowing medium. Components of the mixture are deflected in diagonal trajectories according to their electrophoretic mobility and can be collected at the bottom of the device into a maximum of 96 fractions. Free-flow electrophoresis (FFE) was born as a technique for purifying cells and subcellular organelles, which could be recovered with high purity owing to their very low diffusion coefficients. The Hannig apparatus went through successive designs and improvements, from an original liquid descending curtain to the present commercial version, dubbed Octopus, exploiting an upward liquid stream (Kuhn and Wagner 1989). FFE was recently reported for purification of *Saccharomyces cerevisiae* mitochondria, previously purified by fractional centrifugation. Many more proteins (129) were identified from FFE-purified mitochondria compared with the number of mitochondrial protein extracts isolated by differential centrifugation (80) (Zischka et al. 2003). In addition, a marked decrease of degraded proteins was found in the FFE-purified mitochondrial protein extracts, suggesting that the organelles were less contaminated by lysosomes.

Whilst useful for the purification of organelles, FFE is not ideal for the prefractionation of proteins. This is due to the higher diffusion coefficients of proteins compared with those of cells and organelles. However, a microfabricated FFE device for the continuous separation of proteins has been reported (Kobayashi et al. 2003). FFE, for protein separation, would work much better in the IEF mode (FF-IEF), owing to built-in forces impeding entropic peak dissipation. FF-IEF can also be used as the first dimension of a 2-DE map, the eluted fractions being directly analysed by orthogonal SDS-PAGE. In use, one advantage of FF-IEF was immediately evident. Large proteins such as vinculin (117 kDa) could be recovered and identified. Recovery of high-mass proteins is problematic in conventional IPG gels (Hoffmann et al. 2001).

2.5.3.2 Rotationally Stabilised Focusing Apparatus: the Rotofor

The Rotofor is a device that separates a liquid protein sample into multiple fractions by IEF. A preparative-scale Rotofor is capable of being loaded with up to 1 g of protein, in a total volume of up to 55 mL. A mini-Rotofor, with a reduced volume of about 18 mL is also available (Bier 1998). The device is assembled from 20 sample chambers, separated by liquid-permeable nylon screens. Cation-exchange and anion-exchange membranes are placed against the anodic and cathodic chambers, respectively, so as to prevent diffusion of noxious electrode-associated by-products into the separation chambers. At the end of a preparative run, the 20 focused fractions are collected simultaneously by piercing chamber septa with 20 needles connected to a vacuum source. The narrow pI range fractions can then be used to generate conventional 2-DE maps. In recent times, this method has taken an unexpected turn – the Rotofor has been used directly as the first dimension of a liquid chromatography based 2-D method. Each fraction was then analysed in a second dimension by hydrophobic interaction chromatography, using non-porous reversed-phase high-performance liquid chromatography (HPLC) (Kachman et al. 2002). Each protein peak collected from the HPLC column was then digested to peptides with trypsin, subjected to MALDI-TOF mass spectrometry analysis and identified by database-searching.

More recently, an unexpected application of the Rotofor was reported. This was the fractionation of peptide digests of an entire proteome in an ampholyte-free environment (Xiao et al. 2004). The peptides themselves act as carrier ampholyte buffers and create a pH gradient via an 'autofocusing' process. There is, however a caveat: the pH gradient is quite poor, since only a few peptides have good buffering power and conductivity in the pH 5–8 range.

2.5.3.3 Sample Prefractionation via Multicompartment Electrolysers with Isoelectric Membranes

MCE were first introduced as a class of instruments based on conventional IEF in the presence of soluble, amphoteric, carrier ampholyte buffers. However, MCE devices based on Immobiline membranes (Rigehtti et al. 1989, 1990) represent a quantum leap over previous techniques. This relies on isoelectric membranes that are fabricated with the same acrylic monomers used in IPG fractionations. The advantages of such a method are immediately apparent. Firstly, such a device produces fractions of the proteome that are fully compatible with first-dimension separation in 2-DE maps, a focusing step also based on Immobiline technology. Secondly, it permits a sample to be fractionated into proteins of precise pI values, such

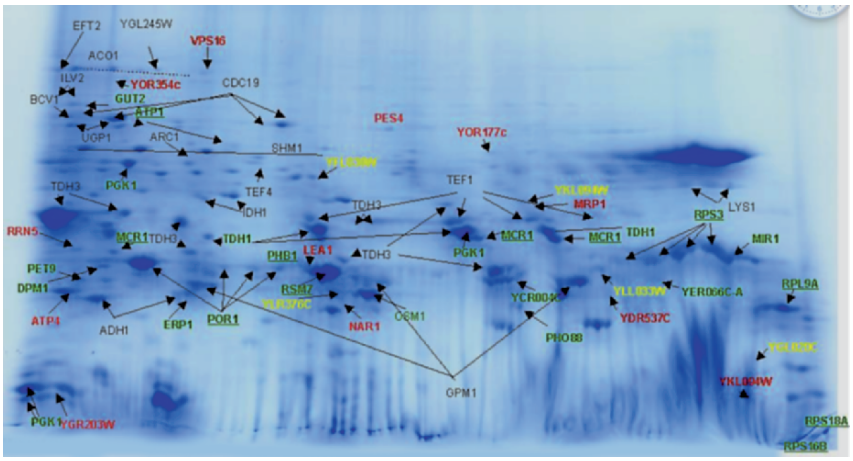


Fig. 2.6 Coomassie brilliant blue stained two-dimensional electrophoresis (2-DE) gel of the alkaline multicompartment electrolyser fraction from a membrane preparation of log-phase yeast. The alkaline fraction was separated in the first-dimension immobilised pH gradient (IPG) using 2% ASB 14 detergent. The excess detergent has combined with SDS to form mixed micelles in the second-dimension gel and caused the smearing observed in the low molecular mass part of the gel. The 2-DE gel is the display of a 1.0-mg membrane protein preparation using an 11-cm pH 7–10.5 IPG for the first dimension and GelChip™ 8–18%T second-dimension gel. The gel is annotated with 237 proteins, representing 93 unique gene products. (From Pedersen et al. 2003, with permission)

that they fall within the pH gradient of any narrow-range or wide-range IPG strip. The chances of protein precipitation occurring will thus be much reduced. In fact, when an entire cell lysate is analysed in a wide gradient, there are few risks of protein precipitation. In contrast, when the same mixture is analysed in a narrow gradient, massive precipitation of all non-isoelectric proteins can occur. Furthermore, owing to the fact that only proteins co-focusing in the same IPG interval will be present, much higher sample loads can be used, permitting the detection of low-abundance proteins. Figure 2.6 gives an example of the large number of *S. cerevisiae* membrane proteins detected, via this prefractionation protocol, in the pH 7–10.5 range (Pedersen et al. 2003). This is an interval that cartographers of the sixteenth century would have described as terra incognita, devoid of any landmarks, and stamped inside the contour of the map the inscription *hic sunt leones*[1], just as they did with the African continent.

The MCE technique continues to undergo improvement. Whilst initial models of the MCE were large, a miniaturised version of the MCE has been

[1]Here are the lions; used by the Romans to indicate the southern limit of the Empire in Africa.

reported (Herbert and Righetti 2000). A remedy for the slow migration of proteins in MCEs because of the sieving effect of isoelectric membranes has also been recently proposed by Fortis et al. (2005), via the introduction of isoelectric beads. These beads were made using ionic acrylamide derivative monomers co-polymerised within the pores of a central ceramic hard core, thus minimising mass-transfer resistance of proteins that are transiently adsorbed onto the beads. As a result, significantly reduced separation time was shown along with a very low electroendosmotic flow.

2.5.3.4 Miniaturised Isoelectric Separation Devices

An interesting variant of focusing by exploiting the IPG technique, called 'off-gel IEF', has been described (Ros et al. 2002). Just like the multicompartment separation technique, the system has been devised for the separation of proteins according to their pI and for their direct recovery in solution without adding buffers or ampholytes. The principle is to place a sample in a liquid chamber, positioned to be touching an IPG gel. Theoretical calculations and modelling have shown that the protonation of proteins occurs in the thin layer of solvation close to the interface of the solution and the IPG gel (Arnaud et al. 2002). Upon application of a voltage, perpendicular to the liquid chamber, all charged species that have pI values above and below the pH of the IPG gel are forced from the liquid chamber into the gel. After separation, only the globally neutral species with pI equal to the pH of the IPG gel remain in solution. In a further extension of this initial work, the system was improved and adapted to a multiwell device, composed of a series of compartments of small volume (100–300 μL) and compatible with current instruments for separation (Michel et al. 2003).

2.6 Other Methods for Prefractionation of Samples

2.6.1 Depletion of High-Abundance Proteins

A major challenge for the direct analysis of proteomes with 2-DE and mass spectrometry is the presence of very high abundance proteins (Anderson and Anderson 2002). In serum, for instance, albumin represents up to 60% of the overall protein present and immunoglobulins another 15%. Accordingly, 2-DE cannot reveal protein species that co-electrophorese with albumin (Pieper et al. 2003). Similarly, the high concentration of albumin is problematic for mass spectrometry analysis of plasma since its peptides suppress the signal of numerous other peptides, which as a consequence are not detected.

In the light of this, it is becoming increasingly common to deplete high-abundance proteins from samples prior to their analysis. It is thus possible to

load more proteins of interest and therefore observe those that may be undetected otherwise. All depletion methods are derived from classical approaches well known in the domain of affinity chromatography. Albumin, the most abundant protein in plasma, can be removed by using specific immunosorbents. Serum antibodies of class immunoglobulin G (IgG) can be depleted by use of immobilised *Staphylococcus aureus* protein A. Whilst these methods have advantages, they may accidentally remove peptides of interest that are associated with the high-abundance protein. Some authors have argued that the removal of high-abundance species may in fact remove an important source of disease biomarkers. In a recent report, Shen et al. (2005) outlined a particularly challenging issue. During albumin depletion, another 815 proteins were co-depleted. When IgGs were captured, another 2,091 proteins were also co-depleted, among which 56% were non-IgG antibody sequences and 44% included low-abundance cytokines and related proteins. Paradoxically, in the final double-depleted sera, only 1,391 proteins could be detected. Of these, 269 proteins were common to those associated with albumin and another 568 were also found in association with IgG. When considering that Shen et al. (2005) detected a total of 4,590 unique gene products, it is ironic that most were discovered in the fractions to be discarded, and not in the fractions that were retained! The potential issues with co-depletion are now becoming common knowledge. For example, Colantonio et al. (2005) discussed the possibility of reaching a balance between total depletion of IgGs, with concomitant loss of other proteins, versus the partial removal of IgGs but minimising the loss of other protein species.

2.6.2 Equaliser Beads: the Democratic Versus the Plutocratic Proteome

A small number of proteins often dominate many proteomes, and obliterate the signal of many others. That is the reason why many scientists lament that, in proteome analysis, the same set of abundant proteins is seen again and again. So the proteome cannot be considered truly democratic! Perhaps the proteome, instead, is oligarchic or plutocratic?

A number of present-day protocols call for a depletion strategy. These seem to derive straight from the French Revolution (e.g. the guillotine) and call for the physical elimination of unwanted protein species. The immunodepletion, especially in sera, of up to 12 of the most abundant species, is a good example of this. There could be another strategy, though, that would call for embracing all proteins in a sample, and seek to give them equal rights. Is this the American revolution, or perhaps the American Dream? This would mean striving for a 'democratic proteome', one in which a dramatic levelling of differences in protein abundance is sought. This approach is now a reality and is embodied in the concept of protein equaliser technology. This has the potential to better understand the 'hidden proteome' and thus discover new biomarkers of clinical importance.

The protein equaliser technology consists of a solid-phase combinatorial library of hexapeptides, synthesised via a short spacer on porous poly(hydroxymethacrylate) beads. The hexapeptides are synthesised on all surfaces of the beads, whereby each bead has a unique ligand that is potentially different from the ligand of any other bead. The basic article, outlining the synthesis of the beads and some of their fundamental properties, was recently published (Thulasiraman et al. 2005; Fig. 2.7), together with some reviews describing the concepts of the equaliser technology (Righetti et al. 2005a, c, 2006).

In the synthesis of the hexapeptides, the 20 natural amino acids are used. Accordingly, the library contains a population of 20^6 linear hexapeptides, i.e. 64 million different ligands. On each bead, the amount of hexapeptide can reach approximately 50 pmol. Such a vast and heterogeneous population of hexapeptides means that, in principle, an appropriate volume of beads should contain a hexapeptide able to interact with just about any protein present in a complex proteome – be it a biological fluid or a tissue or cell

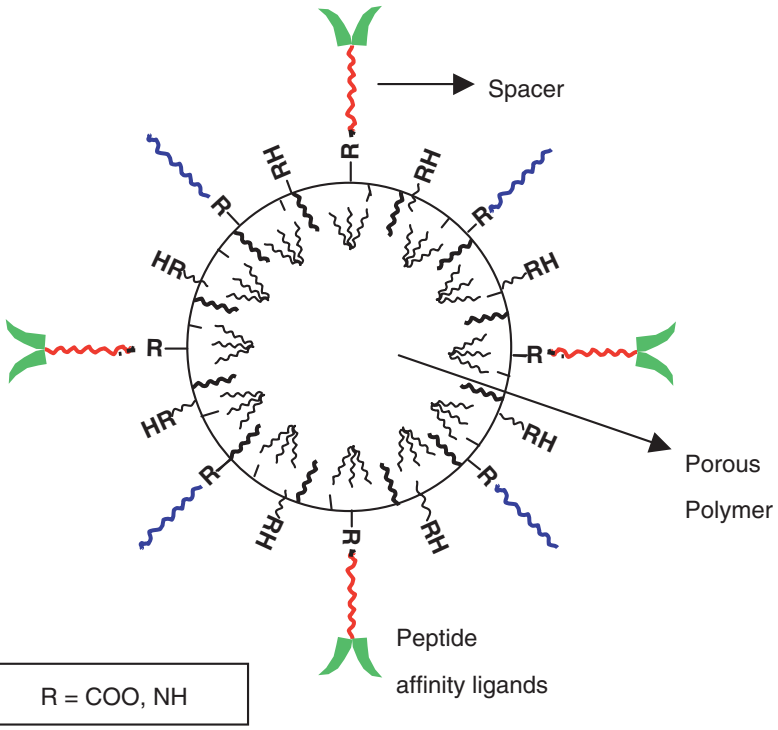


Fig. 2.7 The architecture of beads on which peptide ligands are attached. The structure here is shown to consist of an organic porous polymer on the matrix to which peptide ligands are covalently attached. *R* represents the linker that could be either a primary amino group or a carboxyl group. The peptide can be attached via a spacer. (From Righetti et al. 2006, with permission)

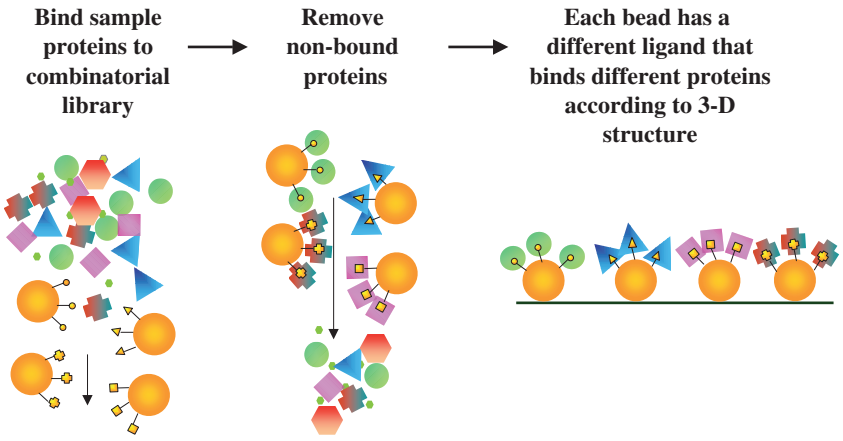


Fig. 2.8 The adsorption/equalisation process on protein equaliser technology (PET) beads. *From left to right*: the sample proteins are bound to the combinatorial library; the unbound proteins are removed by washing under physiological conditions; the final sample caught by the beads is extensively "equalised" and ready for elution (Boschetti et al., unpublished results)

lysate. As schematically represented in Fig. 2.8, when a complex mixture containing proteins of differing abundance, such as human serum, is incubated with ligands of the equaliser library, representatives of each component of the mixture are likely to bind to at least one ligand. In overloading conditions, high-abundance proteins quickly saturate their specific affinity ligand and any excess is removed during the washing step. Low-abundance proteins can, by contrast, continue to concentrate on their specific affinity ligands. After binding and washing, all proteins are eluted from the beads, yielding a mixture of proteins where each protein is still present, but in concentrations much different from in the original mixture. Proteins of high abundance are significantly diluted, while low-abundance species are concentrated.

There are many novel applications of the protein equaliser technology, particularly for human body fluids. For decades, clinical chemistry research has focused on finding new indicators or biomarkers for disease. The search has included many types of tissue specimen but especially the body fluids of plasma, urine, tears, lymph, seminal plasma, milk, saliva and spinal fluid. Body fluids are particularly appealing for analysis since their collection is minimally invasive or, in the case of urine, completely non-invasive. However in the case of urine, any search for biomarkers is aggravated by its very low protein content requiring a concentration step of 100–1,000-fold. Its high salt levels also demand removal prior to any analytical step. Here we outline the results of the protein equaliser technology (PET), as applied to urine analysis. A total of 1.6 L of urine was collected from eight healthy young donors (four male, four female), processed and reduced to a volume of 22 mL dissolved in 25 mM phosphate buffer, pH 7.0. This final volume was adsorbed onto 1 mL of PET beads,

which were then eluted first with 2.2 M thiourea, 7.7 M urea, 4.4% CHAPS and then with 9 M urea titrated to pH 3.8 with 5% (v/v) acetic acid. PET samples and non-PET-treated control samples were then subjected to Fourier transform ion-cyclotron resonance (FT-ICR) mass spectrometry analysis. Control urine revealed a total of 96 unique gene products. In contrast, the PET samples allowed identification of 334 unique protein species in the first eluate and an additional 148 proteins in a second eluate. By eliminating the redundancies and summing up all the species detected, we arrive at a total count of 471 unique protein species (Castagna et al. 2005). The bar graph in Fig. 2.9 shows the increase in species obtained from the sum of the two eluates, compared with the control, while simultaneously expressing their size distribution. The latter is a skewed distribution with a peak in the 40–70-kDa range, as expected because of the filtering properties of the kidney's glomeruli. These results are impressive, when compared with previous findings. Pieper et al. (2004) reported just 150 unique gene products obtained via extensive sample prefractionation and 2-D map analysis.

Even more impressive have been the results with the human serum proteome, from a single, simple experimental protocol using the PET beads. A 300-mL aliquot of serum was processed and subjected to adsorption with the PET ligand library. After elution with three different eluants, 2-D maps showed an impressive increment of spots, compared with control sera (Fig. 2.10). A total of about 800 spots were seen in the three combined eluates with a low sensitivity stain, compared with 115 spots in controls. Upon elution and analysis via FT-ICR mass spectrometry, a total of 3,661 unique gene products could be recognised, with a confidence level of 95% (Righetti et al. 2006). Until 2004, the only extensive data set available on serum proteins was from Anderson et al. (2004), who published a compilation of 1,175 non-redundant species reported in

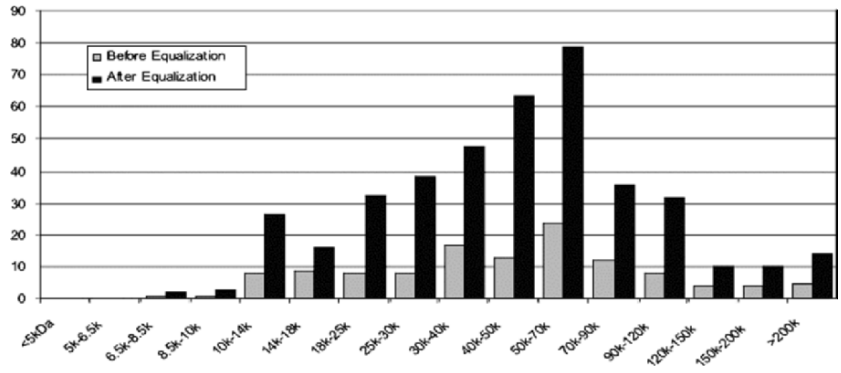


Fig. 2.9 Number of species detected in the two combined eluates (2.2 M thiourea, 7.7 M urea, 4.4% 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate, i.e. TUC, and 9 M urea, pH 3.8, *black bars*) compared with control (*grey bars*) urines, as a function of their respective M_r values in the 5–100-kDa range. There was an overall increase in the combined TUC and urea eluates of 3.9 times. (Modified from Castagna et al. 2005)

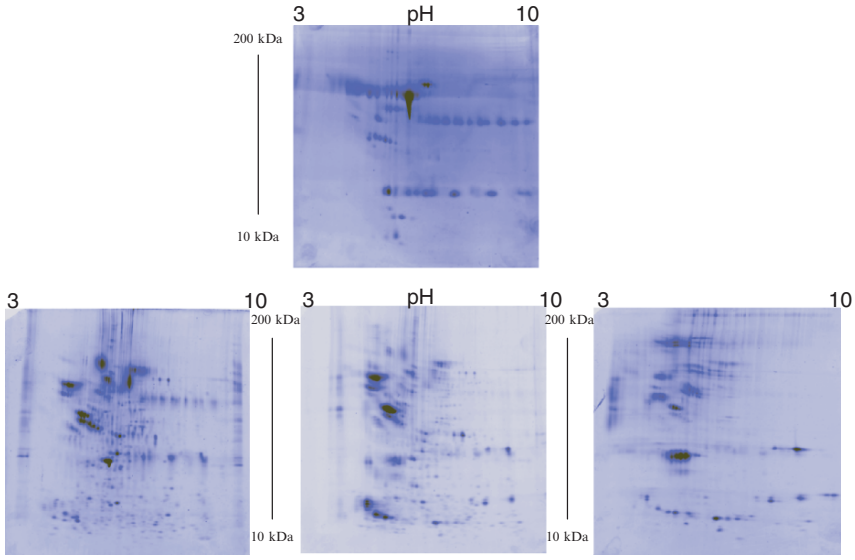


Fig. 2.10 Two-dimensional mapping of human serum before (*upper panel*) and after (*three lower panels*) treatment with PET. First dimension: non-linear IPGs, pH 3–10 range. Second dimension: SDS-PAGE in a 8–15%T porosity gradient. Staining with micellar Coomassie blue. In all cases, 160 μg total protein was loaded. Note that, whereas the control serum exhibits 115 spots, the total number of polypeptide spots counted in the three combined eluates amount to 800. (From Righetti et al. 2006, with permission)

several sources. Although this was an impressive compilation, it was quickly superseded by the massive effort of the Human Proteome Organisation's Plasma Protein Project (PPP). The PPP was started in 2002 as a network of 35 collaborating laboratories and multiple analytical groups. Through these combined efforts, a core set of 3,020 serum plasma proteins could be generated and is now available on public databases (Human Proteome Organisation 2007), together with an additional 6,484 unique proteins identified via only a single peptide (thus with a rather low confidence level, about 60–75%). This brings the total to an extraordinary level of 9,504 proteins. Yet, the validity of this extensive list of serum proteins has been questioned, and may be reduced to only 800–850 genuine proteins of sure identity (States et al. 2006). In contrast, our set of 3,661 serum proteins, based on PET fractionation, while not allowing quantitation, represents a very large data set.

2.7 Conclusions

We believe it is useful to give general concluding remarks and describe what the future may hold. Published work as well as conference presentations indicate a lack of consensus about specific prefractionation

methods. Nevertheless, there is a consensus that protein prefractionation is now a necessary tool for the investigation of all proteomes. The complexity of any proteome is so large that none of the existing technologies can deliver a complete detection and quantification of all proteins that are present. A reduction in complexity, whatever the approach, is not only a useful step but should probably be compulsory. This could theoretically continue until each individual component of a proteome has been separated, but it is clearly not necessary to go that far. Separation by protein groups sharing common physicochemical or biological properties is what will increasingly be applied in the future.

We are particularly excited by the capability of the protein equaliser technology in dealing with the differences in protein abundance in any proteome. We have offered a brief survey of data from two human body fluids – urine and serum – that have important implications for discovery of disease biomarkers. We hope that readers are convinced by the great potential of this technique that has allowed us, for the first time, and with simple manipulations, to truly uncover the hidden proteome. Integration of this technology with classical or novel fractionation methods will undoubtedly increase the capability to discover novel proteins of academic or diagnostic interest. We are especially interested in fractionation methods that directly involve fundamental physicochemical properties of proteins such as their isoelectric points and hydrophobic index.

Acknowledgements. P.G.R. and A.C. are supported in part by grants from FIRB 2001 (grant no. RBNE01KJHT), by the European Community, contract no. 12793, project Allergy Card and by Fondazione Cariplo (Milan). We thank J. Rappsilber, L. Sennels, D. Cecconi and A. Castagna for their valuable help in urine and sera analyses and L. Guerrier and F. Fortis for providing data on protein-depletion methods.

References

Andersen JS, Lyon CE, Fox AH, Leung AK, Lam YW, Steen H, Mann M, Lamond AL (2002) Directed proteomic analysis of the human nucleolus. Curr Biol 12:1–11

Anderson LN, Anderson NG (2002) The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 1:845–867

Anderson LN, Polanski M, Pieper R, Gatlin T, Tirumalai RS, Conrads TP, Veenstra TD, Adkins JN, Pounds JG, Fagan R, Lobley A (2004) The human plasma proteome: a nonredundant list developed by combination of four separate sources. Mol Cell Proteomics 3:311–326

Anderson NL, Hickman BJ (1979) Analytical techniques for cell fractions. XXIV. Isoelectric point standards for two-dimensional electrophoresis. Anal Biochem 93:312–320

Arnaud IL, Josserand J, Rossier JS, Girault HH (2002) Finite element simulation of Off-Gel™ buffering. Electrophoresis 23:3253–3261

Bier M (1998) Recycling isoelectric focusing and isotachophoresis. Electrophoresis 19:1057–1063

Birch RM, O'Byrne C, Booth IR, Cash P (2003) Enrichment of *Escherichia coli* proteins by column chromatography on reactive dye columns. Proteomics 3:764–776

Breci L, Hattrup E, Keeler M, Letarte J, Johnson R, Haynes PA (2005) Comprehensive proteomics in yeast using chromatographic fractionation, gas phase fractionation, protein gel electrophoresis, and isoelectric focusing. Proteomics 5:2018–2028

Castagna A, Cecconi D, Sennels L, Rappsilber J, Guerrier L, Fortis F, Boschetti E, Lomas L, Righetti PG (2005) Exploring the hidden human urinary proteome via ligand library beads. J Proteome Res 4:1917–1930

Chevallet M, Santoni V, Poinas A, Rouquie D, Fuchs A, Kieffer S, Rossignol M, Lunardi J, Garin J, Rabilloud T (1998) New zwitterionic detergents improve the analysis of membrane proteins by two-dimensional electrophoresis. Electrophoresis 19:1901–1909

Colantonio DA, Dunkinson C, Bovenkamp DE, Van Eyk JE (2005) Effective removal of albumin from serum. Proteomics 5:3831–3835

de Duve C (1971) Isolation of cell organelles by fractional centrifugation. J Cell Biol 50:20–55

Douette P, Sluse FE (2006) Mitochondrial uncoupling proteins: new insights from functional and proteomic studies. Free Radic Biol Med 40:1097–1107

Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, Hunt DF, White FM (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. Nat Biotechnol 20:301–307

Fortis F, Girot P, Brieau O, Castagna A, Righetti PG, Boschetti E (2005) Isoelectric beads for proteome pre-fractionation. II: Experimental evaluation in a multicompartment electrolyzer. Proteomics 5:629–638

Fountoulakis M, Takacs B (1998) Design of protein purification pathways: application to the proteome of *Haemophilus influenzae* using heparin chromatography. Protein Expr Purif 14:113–119

Fountoulakis M, Langen H, Gray C, Takacs B (1998) Enrichment and purification of proteins of *Haemophilus influenzae* by chromatofocusing. J Chromatogr A 806:279–291

Fountoulakis M, Takacs MF, Takacs B (1999a) Enrichment of low-copy-number gene products by hydrophobic interaction chromatography. J Chromatogr A 833:157–168

Fountoulakis, M, Takacs MF, Berndt P, Langen H, Takacs B (1999b) Enrichment of low abundance proteins of *Escherichia coli* by hydroxyapatite chromatography. Electrophoresis 20:2181–2195

Galvani M, Hamdan M, Herbert B, Righetti PG (2001a) Protein alkylation in presence/absence of thiourea in proteome analysis: a MALDI-TOF mass spectrometry investigation. Electrophoresis 22:2066–2074

Galvani M, Hamdan M, Herbert B, Righetti PG (2001b) Alkylation kinetics of proteins in preparation for two-dimensional maps: a matrix assisted laser desorption/ionization-mass spectrometry investigation. Electrophoresis 22:2058–2065

Ghosh D, Krokhin O, Antonovici M, Ens W, Standing KG, Beavis RC, Wilkins JA, (2004) Lectin affinity as an approach to the proteomic analysis of membrane glycoproteins. J Proteome Res 3:841–850

Gianazza E, Rabilloud T, Quaglia L, Caccia P, Astrua-Testori S, Osio L, Grazioli G, Righetti PG (1987) Additives for immobilized pH gradient two-dimensional separation of particulate material: comparison between commercial and new synthetic detergents. Anal Biochem 165:247–257

Guerrier L, Lomas L, Boschetti EJ (2005) A simplified monobuffer multidimensional chromatography for high-throughput proteome fractionation. J Chromatogr A 1073:25–33

Hamdan M, Righetti PG (2005) Two dimensional maps. In: Proteomics today: protein assessment and biomarkers using mass spectrometry, 2D electrophoresis, and microarray technology. Wiley-Interscience, Hoboken, pp 341–402

Hannig K (1978) Continuous free-flow electrophoresis as an analytical and preparative method in biology. J Chromatogr 159:183–191

Herbert B, Righetti PG (2000) A turning point in proteome analysis: sample pre-fractionation via multicompartment electrolyzers with isoelectric membranes. Electrophoresis 21:3639–3648

Herbert B, Galvani M, Hamdan M, Oliveri E, McCarthy J, Pedersen S, Righetti PG (2001) Reduction and alkylation of proteins in preparation of two-dimensional map analysis: why, when and how? Electrophoresis 22:2046–2057

Herbert B, Hopwood F, Oxley D, McCarthy J, Laver M, Grinyer J, Goodall A, Williams K, Castagna A, Righetti PG (2003a) Beta-elimination: an unexpected artefact in proteome analysis. Proteomics 3:826–831

Herbert B, Pedersen SK, Harry JL, Sebastian L, Grinyer J, Traini MD, McCarthy JT, Wilkins MR, Gooley AA, Righetti PG, Packer NH, Williams KL (2003b) Mastering proteome complexity using two-dimensional gel electrophoresis. PharmaGenomics 3:22–36

Herbert B, Righetti PG, McCarthy J, Grinyer J, Castagna A, Laver M, Durack M, Rummery G, Harcourt R, Williams KL (2004) Sample preparation for high-resolution two-dimensional electrophoresis by isoelectric fractionation in an MCE. In: Simpson RJ (ed) Purifying proteins for proteomics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp 431–442

Herbert BR, Sanchez J-C, Bini L (1997) Two-dimensional electrophoresis: the state of the art and future directions. In: Wilkins MR, Williams KL, Appel RD, Herbert BR, Molloy MP, Gooley AA, Walsh BJ, Bryson WG, Williams KL (1998) Improved protein solubility in two-dimensional electrophoresis using tributyl phosphine as reducing agent. Electrophoresis 19:845–851

Hochstrasser DF (eds) Proteome research: new frontiers in functional genomics. Springer, Berlin, pp 13–33

Hoffmann P, Ji H, Moritz RL, Connolly LM, Frecklington DF, Layton MJ, Eddes JS, Simpson RJ (2001) Continuous free-flow electrophoresis separation of cytosolic proteins from the human colon carcinoma cell line LIM 1215: a non two-dimensional gel electrophoresis-based proteome analysis strategy. Proteomics 1:807–818

Human Proteome Organisation (2007) Plasma Proteome Project. http://www.bioinformatics.med.umich.edu/hupo/ppp. Cited 26 Mar 2007

IonSource (2007) Carbamylation of proteins. In: Mass spectrometry and biotechnology resource. http://www.ionsource.com/Card/carbam/carbam.htm. Cited 26 Mar 2007

Jung E, Hoogland C, Chiappe D, Sanchez JC, Hochstrasser DF (2000) The establishment of a human liver nuclei two-dimensional electrophoresis reference map. Electrophoresis 21:3483–3487

Kachman MT, Wang H, Schwartz DR, Cho KR, Lubman DM (2002) A 2-D liquid separations/mass mapping method for interlysate comparison of ovarian cancers. Anal Chem 74:1779–1791

Kobayashi H, Shimamura K, Akaida T, Sakano E, Tajima N, Funazaki J, Suzuki H, Shinohara E (2003) Free-flow electrophoresis in a microfabricated chamber with a micromodule fraction separator. Continuous separation of proteins. J Chromatgr A 990:169–178

Kuhn R, Wagner H (1989) Application of free flow electrophoresis to the preparative purification of basic proteins from an *E. coli* cell extract. J Chromatogr 481:343–350

Lopez MF, Kristal BS, Chernokalskaya E, Lazarev A, Shestopalov AI, Bogdanova A, Robinson M (2000) High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. Electrophoresis 21:3227–3440

Luche S, Diemer H, Tastet C, Chevallet M, Van Dorsselaer A, Leize-Wagner E, Rabilloud T (2004) About thiol derivatization and resolution of basic proteins in two-dimensional electrophoresis. Proteomics 4:551–561

McCarthy J, Hopwood F, Oxley D, Laver M, Castagna A, Righetti PG, Williams K, Herbert B (2003) Carbamylation of proteins in 2-D electrophoresis – myth or reality? J Proteome Res 2:239–242

Michel PE, Reymond P, Arnaud IL, Josserand J, Girault HH, Rossier JS (2003) Protein fractionation in a multicompartment device using Off-Gel isoelectric focusing. Electrophoresis 24:3–11

Pedersen SK, Harry JL, Sebastian L, Baker J, Traini MD, McCarthy JT, Manoharan A, Wilkins MR, Gooley AA, Righetti PG, Packer NH, Williams KL, Herbert B (2003) Unseen proteome: mining below the tip of the iceberg to find low abundance and membrane proteins. J Proteome Res 2:303–312

Pieper R, Gatlin CL, Makusky AJ, Russo PS, Schatz CR, Miller SS, Su Q, McGrath AM, Estock MA, Parmar PP, Zhao M, Huang ST, Zhou J, Wang F, Esquer-Blasco R, Anderson NL, Taylor J,

Steiner S (2003) The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins. Proteomics 3:1345–1364

Pieper R, Gatlin CL, McGrath AM, Makusky AJ, Mondal M, Seonarain M, Field E, Schatz CR, Estock MA, Ahmed N, Anderson NG, Steiner S (2004) Characterization of the human urinary proteome: a method for high-resolution display of urinary proteins on two-dimensional electrophoresis gels with a yield of nearly 1400 distinct protein spots. Proteomics 4:1159–1174

Rabilloud T, Adessi C, Giraudel A, Lunardi J (1997) Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 18:307–316

Righetti PG, Wenisch E, Faupel M (1989) Preparative protein purification in a multi-compartment electrolyzer with Immobiline membranes. J Chromatogr 475:293–309

Righetti PG, Wenisch E, Jungbauer A, Katinger H, Faupel. M (1990) Preparative purification of human monoclonal antibody isoforms in a multicompartment electrolyzer with Immobiline membranes. J Chromatogr 500:681–696

Righetti PG, Faupel M, Wenisch E (1992) Preparative electrophoresis with and without immobilized pH gradients. In: Chrambach A, Dunn MJ, Radola BJ (eds) Advances in electrophoresis, vol 5. VCH, Weinheim, pp 159–200

Righetti PG, Castagna A, Herbert B (2001) Prefractionation techniques in proteome analysis. Anal Chem 73:320A–326A

Righetti PG, Castagna A, Herbert B, Reymond F, Rossier JS (2003) Prefractionation techniques in proteome analysis. Proteomics 3:1397–1407

Righetti PG, Castagna A, Antonioli P, Boschetti E (2005a) Prefractionation techniques in proteome analysis: the mining tools of the third millennium. Electrophoresis 26:297–319

Righetti PG, Castagna A, Herbert B, Candiano G (2005b) How to bring the "unseen" proteome to the limelight via electrophoretic pre-fractionation techniques. Biosci Rep 25:3–17

Righetti PG, Castagna A, Antonucci F, Piubelli C, Cecconi D, Campostrini N, Rustichelli C, Antonioli P, Zanusso G, Monaco S, Lomas L, Boschetti E (2005c) Proteome analysis in the clinical chemistry laboratory: myth or reality? Clin Chim Acta 357:123–139

Righetti PG, Boschetti E, Lomas L, Citterio A (2006) Protein equalizer technology: the quest for a "democratic proteome". Proteomics 6:3980–3992

Ros A, Faupel M, Mees H, Oostrum JV, Ferrigno R, Reymond F, Michel P, Rossier JS, Girault HH (2002) Protein purification by Off-Gel electrophoresis. Proteomics 2:151–156

Shapiro R (1999) Prebiotic cytosine synthesis: a critical analysis and implications for the origin of life. Proc Natl Acad Sci USA 96:4396–4401

Shen Y, Kim J, Strittmatter EF, Jacobs JM, Camp DG II, Fang R, Tolié N, Moore RJ, Smith RD (2005) Characterization of the human blood plasma proteome. Proteomics 5:4034-4045

Smith SD, She YM, Roberts EA, Sarkar B (2004) Using immobilized metal affinity chromatography, two-dimensional electrophoresis and mass spectrometry to identify hepatocellular proteins with copper-binding ability. J Proteome Res 3:834–840

States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM (2006) Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. Nat Biotechnol 24:333–338

Steen H, Mann M (2001) Similarity between condensed phase and gas phase chemistry: fragmentation of peptides containing oxidized cysteine residues and its implications for proteomics. J Am Soc Mass Spectrom 12:228–232

Thulasiraman V, Lin S, Gheorghiu L, Lathrop J, Lomas L, Hammond D, Boschetti E (2005) Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. Electrophoresis 26:3561–3571

Vuillard L, Marret N, Rabilloud T. (1995) Enhancing protein solubilization with nondetergent sulfobetaines. Electrophoresis 16:295–297

Xiao Z, Conrads TP, Lucas DA, Janini GM, Schaefer CF, Buetow KH, Issaq HJ, Veenstra TD (2004) Direct ampholyte-free liquid-phase isoelectric peptide focusing: application to the human serum proteome. Electrophoresis 25:128–133

Zischka H, Weber G, Weber PJA, Posch A, Braun RJ, Buehringer D, Schneider U, Nissum M, Meitinger T, Ueffing M, Eckerskorn C (2003) Improved proteome analysis of *Saccharomyces cerevisiae* mitochondria by free-flow electrophoresis. Proteomics 3:906–916

3 Protein Identification in Proteomics

Patricia Hernandez, Pierre-Alain Binz, and Marc R. Wilkins

Abstract

Identifying proteins contained in a biological sample represents a central task in a majority of proteomics projects. The present chapter discusses various attributes that can be used to identify proteins: the species of origin, the isoelectric point, the amino acid composition, sequence tags and mass values. Then, the chapter focuses on protein identification using mass spectrometry data. It introduces the mass spectrometry technology, presents specific instrumentations and describes the major identification approaches, including peptide mass fingerprinting, peptide fragment fingerprinting and de novo sequencing. Available software tools are described, and for each approach, a concrete example of identification results is given. Particular care is attributed to the identification of proteins carrying post-translational modifications.

3.1 Introduction

Protein identification plays a central role in proteomic research. It is the essence of projects that aim to catalogue all proteins present in a biological sample. It is inherent to protein-expression profiling, which seeks to discover and identify differentially expressed proteins. In association with efforts to better understand molecular mechanisms and pathways, it is also linked with the mapping of protein co- and post-translational modifications.

The number of proteins to be identified in proteomics projects is massive. It is therefore essential to use automated identification techniques that generate unequivocal results. Currently, the most widespread method for protein identification is the correlation of spectra of proteins and peptides obtained by mass spectrometry (MS) with protein sequence data stored in databases. These mass spectra are called 'fingerprints' because they represent a unique key for a protein that can be used for its identification. When sequence databases are searched, a number of additional attributes can be used in association with spectral data. For example, taxonomic information about the origin of the sample can be used to restrict a search to proteins

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

from one or more relevant species. Annotations in entries of the Swiss-Prot database, such as information on subcellular compartments, may be of further help. When available, other attributes like the protein molecular weight and isolectric point, as well as knowledge of its sequence or amino acid composition, are also useful. These last attributes are discussed briefly in the following sections. Then, the chapter presents details of protein identification using MS.

3.2 Attributes of Proteins Useful for Their Identification

3.2.1 Species of Origin

Protein species of origin is obviously not, by itself, a sufficiently powerful attribute to allow the unequivocal identification of any protein; however, it can be used to limit a database search to proteins from relevant species. Usually, identification tools allow the choice from a number of kingdoms, phyla and one or more genera. When an organism under study is not listed in an identification tool, the search can be extended to related species whose proteins have high sequence similarity. This process is called cross-species identification.

3.2.2 Protein Isoelectric Point

The isoelectric point (pI) of a denatured protein is a function of its amino acid composition, N- and C-terminal amino acids, and any post-translational modifications. On two-dimensional electrophoresis (2-DE) polyacrylamide gels (see Chap. 2), the experimental (apparent) pI of many proteins can be estimated at the same time. Estimation is typically done by first calculating the theoretical pI of ten or more proteins on a 2-DE gel, using a tool such as Compute pI/MW (1994). Theoretical pI values for these proteins can then be used with image-analysis software to create a pI grid, allowing the pI of other proteins on the gel to be estimated. The accuracy of these estimates will depend largely on the care that is taken in the construction of the grid, and the type of sample being studied.

3.2.3 Protein Mass

In database searches, the molecular mass of a protein can be used to dramatically reduce the number of sequences that might be the identity of a protein. Protein mass can be theoretically determined by summing the mass of all amino acids of a protein together with the mass of any post-translational

modifications. From 2-DE gels, there are two ways to estimate the experimental mass of proteins. Firstly, as with protein pI, this can be done by gel image analysis. The theoretical mass of known proteins on the 2-DE gel can be calculated, and these proteins can be used to create a grid of apparent protein mass. The apparent mass of other protein spots can then be estimated. Whilst this approach can estimate the mass of hundreds of proteins, many proteins migrate anomalously during polyacrylamide gel electrophoresis, and apparent mass values may be in error by more than 30% (Wilkins et al. 1996b). Practically, protein pI and mass are usually used in conjunction with other more specific attributes, notably MS-based identification using peptide mass fingerprints (Sect. 3.3).

3.2.4 Partial Sequence or Sequence Tag

Any amino acid sequence, even if only a few amino acids in length, is a very specific piece of information. For example, there are 8,000 possible combinations of three amino acid sequences, 160,000 combinations of four amino acid sequences and a massive 3,200,000 different sequences of five amino acids. Small stretches of sequence (called sequence tags) can be generated using Edman sequencing or by tandem MS (MS/MS; Sect. 3.3.4). It has been observed that protein N- and C-terminal sequence tags are surprisingly specific, particularly in organisms with small genomes. For example, in *Escherichia coli*, about 60% of proteins have unique N-terminal sequence tags of length four amino acids and 90% of proteins have unique C-terminal tags of length four amino acids. Where terminal sequence tags are not unique for a protein, relatively few proteins share the same tag. The TagIdent tool (Table 3.1) allows searching a database of proteins using a sequence tag in combination with the protein mass and pI. Taxonomic criteria and Swiss-Prot keywords can also be used to narrow the search.

3.2.5 Protein Amino Acid Composition

Amino acid composition is defined by the numbers of each amino acid present in a protein, and is usually described as a percentage. It has been used for the identification of proteins from 2-DE gels, and has been determined in two ways. Latter et al. (1984) estimated protein amino acid composition by radiolabelling whole organisms with one, two or three different radioactive amino acids, and running each singly labelled sample on a separate 2-DE gel. The compositional ratio of the amino acids was then determined by comparing quantitative densitometric measurements for each protein spot from the different 2-DE gels. This approach has the advantage of producing data in parallel for up to hundreds of spots for any sample. However, owing to the difficulties of sample preparation and data interpretation, it did not emerge

Table 3.1 Some protein- and peptide-identification programs available on the Internet, and their URLs. Those shown in *bold* are found on Web sites maintained by the Swiss Institute of Bioinformatics

Tool	URL
Identification by amino acid composition	
AACompIdent	**http://www.expasy.org/tools/aacomp/**
Peptide mass fingerprinting	
Aldente	**http://www.expasy.org/tools/aldente/**
Mascot	http://www.matrixscience.com
MS-Fit	http://prospector.ucsf.edu
PepMapper	http://wolf.bms.umist.ac.uk/mapper/
PeptideSearch	http://www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html
ProFound	http://www.unb.br/cbsp/paginiciais/profound.htm
Peptide fragment fingerprinting (tandem mass spectrometry sequence search)	
InsPecT	http://peptide.ucsd.edu/inspect.py
Mascot	http://www.matrixscience.com
MS-Tag	http://prospector.ucsf.edu
OMSSA	http://pubchem.ncbi.nlm.nih.gov/omssa/
Phenyx	**http://phenyx.vital-it.ch/pwi**
Popitam	**http://www.expasy.org/tools/popitam/**
Sequest	http://fields.scripps.edu/sequest/index.html (no Web interface, not free)
PepProbe	http://bart.scripps.edu/public/search/pep_probe/search.jsp
X!Tandem	http://www.thegpm.org
Spectral library search	
X!Hunter	http://h201.thegpm.org/
SpectraST	http://www.peptideatlas.org/spectrast/
NIST MS Search Program	http://www.nist.gov/srd/nist1a.htm (no Web interface)
BiblioSpec	http://proteome.gs.washington.edu/bibliospec/documentation/ (no Web interface)
De novo sequencing	
Lutefisk	http://www.hairyfatguy.com/Lutefisk (no Web interface)
PEAKS	http://www.bioinformaticssolutions.com/products/peaks/index.php (no Web interface, not free)
pepNovo	http://peptide.ucsd.edu/pepnovo.py
SeqMS	http://www.protein.osaka-u.ac.jp/rcsfp/profiling/SeqMS.html (no Web interface)

(continued)

Table 3.1 (continued)

Tool	URL
Sequence similarity search tools	
CIDentify	http://ftp.virginia.edu/pub/fasta/CIDentify/ (no Web interface)
MultiIdent	**http://www.expasy.org/tools/multiident/**
Mascot	http://www.matrixscience.com
MS-Blast	http://dove.embl-heidelberg.de/Blast2/msblast.html
MS-Pattern	http://prospector.ucsf.edu
MS-Seq	http://prospector.ucsf.edu
PeptideSearch	http://www.mann.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html
PepSea	http://www.unb.br/cbsp/paginiciais/pepseaseqtag.htm
TagIdent	**http://www.expasy.org/tools/tagident.html**

as a general, high-throughput method. Alternatively, numerous groups have generated amino acid composition data by hydrolysing individual proteins from 2-DE gels with strong acid, and then analysing the resulting free amino acids with chromatographic techniques (Eckerskorn et al. 1988; Jungblut et al. 1992; Wilkins et al. 1996a). Amino acid composition based identification can then be performed with the AACompIdent tool (Table 3.1). Again, the search can be narrowed with other relevant attributes, like the protein mass, pI, taxonomic criteria and Swiss-Prot keywords. Whilst this method was in use before many MS techniques became commonplace, it has largely been superseded.

3.3 Protein Identification by Mass Spectrometry

MS is now the technique of choice for protein identification. A number of approaches can be used and these are discussed next.

3.3.1 'Top-Down' Versus 'Bottom-Up' Strategies for Protein Identification

When analysing proteins by MS, we can use a 'top-down' or a 'bottom-up' approach. In the 'top-down' approach, whole proteins are purified and fragmented in the mass spectrometer. This approach is usually restricted to the more powerful Fourier transform (FT) mass spectrometers that have the

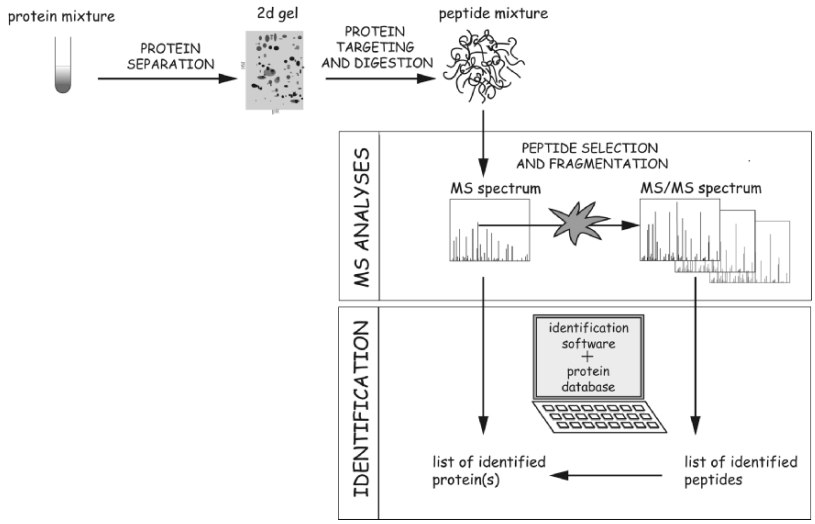


Fig. 3.1 'Bottom-up' protein identification using two-dimensional gel electrophoresis and mass spectrometry. In this workflow, two-dimensional gel electrophoresis is performed on a protein mixture. A protein (spot) of interest is selected, excised and digested. The resulting peptides undergo mass spectrometry analysis of parent ions, which can be used to identify proteins by peptide mass fingerprinting. If desired, peptides of interest can also be fragmented by tandem mass spectrometry (*MS/MS*). The resulting fragment masses can then be correlated with theoretical peptide sequences by a MS/MS protein-identification algorithm

capacity to fragment whole proteins with very high mass accuracy, allowing the facile interpretation of fragmentation results. In the 'bottom-up' approach, whole proteins are purified, cleaved into peptides using proteolytic enzymes, and the peptides are then analysed and/or fragmented using MS (Fig. 3.1). Although variations may occur in experimental protocols or the techniques used, the general 'bottom-up' framework remains the same. The first step is sample preparation. Cells are lysed and cellular compartments of interest are isolated. Lipids, DNA and other impurities are discarded. The proteins are then dissolved and denatured, and one or more separation techniques are used to reduce the sample complexity. The separation procedure may occur before or after digestion. Where two-dimensional polyacrylamide gel electrophoresis is used, protein digestion follows separation (Fig. 3.1), and the resulting peptide mixture corresponds to one or a small number of proteins. The peptides are finally analysed by single-stage MS or MS/MS. In the 'shotgun' proteomics approach, digestion is performed on a complex mixture of proteins. The resulting peptides are then resolved by one or several dimensions of chromatography, and analysed by MS/MS. This technique is often preferred to two-dimensional gel electrophoresis because

of the possibility to directly couple the separation system with the mass spectrometer, thus automating data acquisition and protein identification.

3.3.2 Introduction to Mass Spectrometry

MS is a technology that measures the mass-to-charge ratio (m/z) of molecules. The first experiments carried out with MS date back to the end of the nineteenth century (Scripps Center for Mass Spectrometry 2005). Since then, technology for MS has improved dramatically and there are now more than 20 different mass spectrometers commercially available for proteomics. Whilst these instruments have different designs and functionality, they have a number of common constituents. All mass spectrometers carry out two distinct analytical functions: ionisation and mass analysis. Single-stage MS requires one mass analyser. MS/MS conceptually requires two mass analysers in series. The first analyser separates the peptides according to their m/z values; selected peptides (called precursors) then undergo fragmentation, and the second analyser measures the m/z ratios of the resulting fragments.

3.3.2.1 Ionisation

Various ionisation methods are used in MS. These include electron impact, chemical ionisation, fast atom bombardment, field desorption, electrospray ionisation (ESI) and laser desorption. In proteomics, mainly ESI and matrix-assisted laser desorption/ionisation (MALDI) are used. For MALDI analysis (Fig. 3.2) analytes are first embedded into a crystalline matrix on a metal 'target' plate. When this target is placed into the vacuum of a MALDI source, pulses of laser light (typically a nitrogen laser at 337 nm) are directed at the matrix, causing vibrational excitation and ejection of the analyte molecules and the matrix components. As the matrix evaporates, analytes are liberated and ionised. Generally, the observed ions of proteins and peptides are protonated and carry a single charge.

In an ESI source (Fig. 3.3), the sample is presented to the mass spectrometer in a liquid form at atmospheric pressure. It flows into a needle that is subject to a high voltage (1–6 kV). This electrical potential, applied between the needle tip and the inlet of the mass spectrometer, leads to an accumulation of the same type of charges on drops that exit the needle tip. Owing to electrostatic repulsion, these solvent drops spontaneously dissociate to form a fine spray of highly charged droplets. The flow of droplets is directed through a countercurrent flow of heated gas, causing the solvent to evaporate and the droplets to shrink. This causes the charge concentration on the surface of the droplets to increase. As the electrical charge reaches a critical state, known as the Rayleigh limit, the droplets explode into smaller and lower-charged particles. This process of shrinking and explosion is repeated until

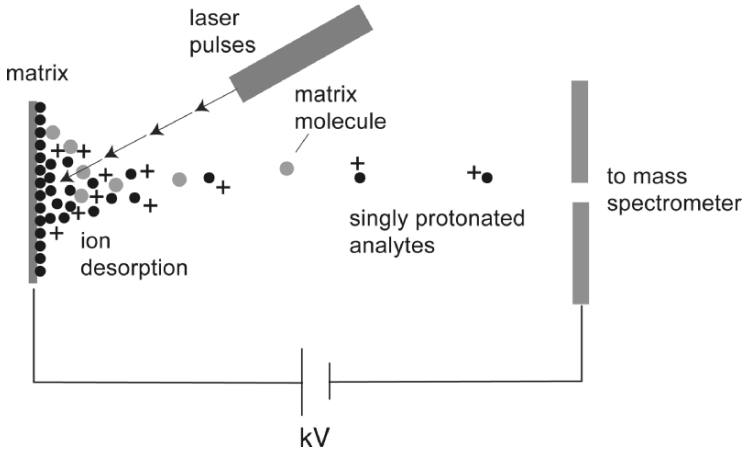


Fig. 3.2 Fundamentals of ionisation by the matrix-assisted laser desorption/ionisation (MALDI) process. Microlitre quantities of liquid samples are mixed with a matrix molecule, such as α-cyano-4hydroxycinnamic acid, and dried onto a stainless steel or gold-plated target. A pulsing laser, most commonly a N_2 laser emitting at 337 nm, is then used to irradiate the matrix-embedded sample. This creates molecular ions that are accelerated by an electric field

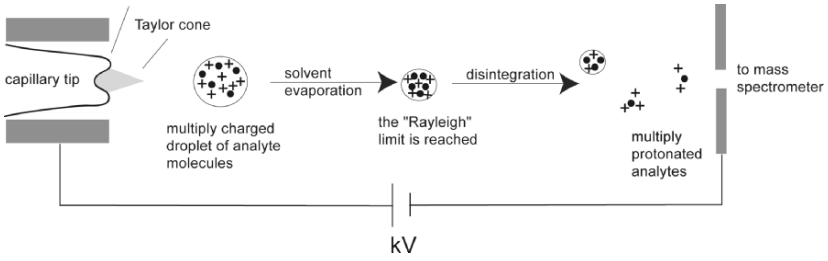


Fig. 3.3 The process of peptide ionisation by electrospray. See the text for more details

individually charged analyte molecules remain. Generally, a mixture of singly and multiply charged ions is generated. Since the sample is introduced in a liquid state at atmospheric pressure, ESI sources can easily be associated with online liquid-phase separation techniques, such as liquid chromatography.

3.3.2.2 Mass Analysis

Mass analysers separate charged molecules according to their mass-to-charge ratio m/z, where m is the mass of the ion and z is the number of its elementary charges. Four basic types of mass analysers are currently in use in proteomic research: quadrupole, ion-trap, time-of-flight (TOF) and FT ion cyclotron resonance (ICR) analysers. Below we describe each type of mass analyser.

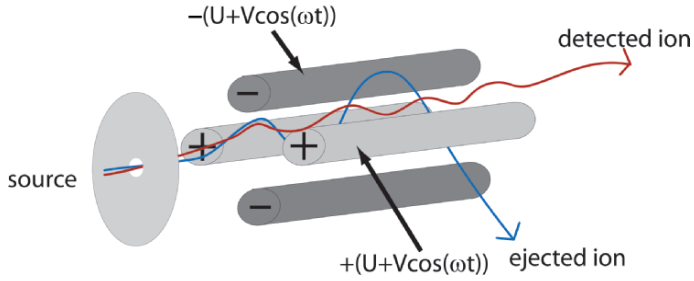


Fig. 3.4 A quadrupole. Depending on the voltages applied, ions are either ejected from the quadrupole or sent to the detector

Quadrupole mass analysers (Fig. 3.4) consist of four parallel and symmetrically arranged metallic rods. One couple of opposite rods have a positive electrical potential $+[U + V\cos(\omega t)]$, while the other couple of opposite rods have a negative potential $-[U + V\cos(\omega t)]$. Ions oscillate while traversing the field along the central axis of the rods. For each U, V and ω value, only ions of a certain m/z enter in resonance and follow a stable trajectory through the quadrupole to the detector. All others are not stabilised and deviate out of the mass analyser. The quadrupole is therefore considered as a mass filter. To obtain a complete spectrum, one continuously varies or scans the electromagnetic field in the quadrupole while the sample passes through the analyser.

Ion-trap mass analysers are devices that can store or trap charged molecules for a long time. Ions are trapped in time by electric potentials of approximately $U + V\cos(\omega t)$ produced by a ring-shaped electrode and two end-cap electrodes. This occurs in a space of 2–3 cm^3 that is filled with a dilute inert gas. Ions of different m/z values enter the trap at one of the end-cap electrodes and remain trapped, oscillating at frequencies that are related to their m/z values. By changing U, V and ω, ions of certain m/z become turn-to-turn excited and are ejected from the opposite end cap. Another geometry of ion traps is also gaining interest; the so-called linear ion trap. These traps are similar to quadrupoles, however electromagnetic signals are designed to trap the appropriate ions in a rectangular-shaped space. Current instruments have the capacity to scan m/z values at a very high speed.

TOF mass analysers measure ions that are accelerated in an electrical field, then travel down a field-free vacuum tube towards an ion detector (Fig. 3.5). All ions in the source are given the same initial amount of kinetic energy, but their velocity is a function of their mass and charge. The time needed to travel the distance between the source and the detector is therefore dependent on their m/z values. This can be calculated using the kinetic energy equation, given the tube length and the measured times of flight. The flight tubes used in high-resolution instruments usually include a so-called reflectron.

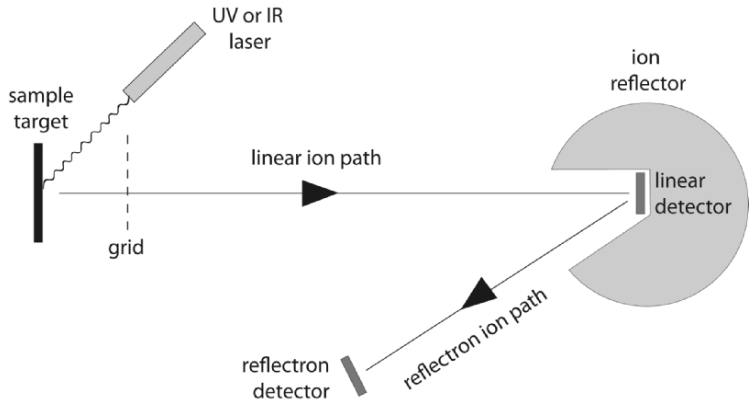


Fig. 3.5 A MALDI time-of-flight (TOF) mass spectrometer. Ions generated in a MALDI source are accelerated through a grid and enter a field-free flight tube with a velocity that is function of their mass. Their time of flight in this tube is measured by a detector, for instance directly in the linear detector. To increase mass resolution, recent MALDI-TOF instruments use an ion reflector. The reflector acts as an 'electronic mirror' that allows ions to be turned around and sent to a second detector in a refocused manner

The reflectron uses an electric field to reverse the path of ions as they approach the end of the first flight tube. The ions are then sent along a slightly different trajectory from their initial path towards a second detector. Highest resolution can be obtained by combining the use of a reflectron with a delayed extraction device. Delayed extraction allows the initial velocity of ions to be standardised at the entrance of the TOF analyser.

FT-ICR mass analysers allow ions to be accumulated and stored for periods as long as minutes. FT mass spectrometers consist of a cubic cell inside a strong magnetic field. Injected ions rotate around the magnetic field with a frequency typical for their *m*/*z*. By varying the electric fields, changes in the ion frequency of rotation can be measured and converted into *m*/*z* using a Fourier transformation.

3.3.2.3 Instrumentation

There are many ways of combining various ionisation sources, analysers and fragmentation devices. The most frequently used combinations for proteomic research are MALDI-TOF, MALDI-TOF/TOF, ESI–triple quadrupole, ESI–ion trap, ESI–quadrupole–TOF (ESI-Q-TOF), ESI–quadrupole–ion trap, ESI-MALDI–linear trap and FT-ICR. The MALDI-TOF/TOF combination consists of a MALDI source coupled with a linear TOF, followed by

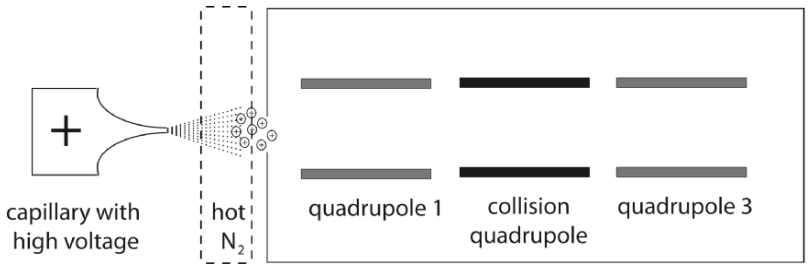


Fig. 3.6 A triple quadrupole mass spectrometer equipped with an electrospray ion source. Peptide ions introduced via the electrospray source are scanned in the first quadrupole; selected peptides are fragmented in the second; and the masses of the resulting fragments are scanned in the third and sent to the detector

a collision-induced dissociation fragmentation cell, and then a TOF analyser equipped with a reflectron. By contrast, the ESI–triple quadrupole is composed of three quadrupoles arranged in series (Fig. 3.6). A possible variation consists in replacing the last quadrupole by a TOF analyser (ESI-Q-TOF). Some mass spectrometers (such as ion traps) allow the two analysis steps and the fragmentation to occur at the same location. A FT cell might be added in series to a linear ion trap to measure selected or fragmented ions from the trap with the high resolution and accuracy of the FT cell. This allows the parallel analysis of ions in the trap and those in the FT cell. MALDI-TOF instruments are the least expensive and are widespread, with FT-ICR machines being the most expensive and typically found only in specialist MS facilities. A comprehensive list of instrumentation is regularly updated online (hyphen MassSpec Consultancy 2006).

3.3.3 Protein Identification by Peptide Mass Fingerprinting

3.3.3.1 Principle

Identification of proteins by peptide mass fingerprinting (PMF; Fig. 3.7) was independently described in 1993 by a number of groups (Henzel et al. 1993; James et al. 1993; Mann et al. 1993; Pappin et al. 1993; Yates et al. 1993). Proteins are first digested using a site-specific proteolytic enzyme. The masses of the resulting peptides are then determined by MS. Since each protein has a different sequence, the masses obtained for each protein are a unique 'fingerprint' (Fig. 3.8). Protein identification is performed by comparing the experimentally determined peptide masses with theoretically determined peptide masses generated from protein sequences in databases. Scoring systems are used to rank proteins, whereby high-ranking proteins from the database have the largest numbers of peptides in common with the protein that has been analysed.

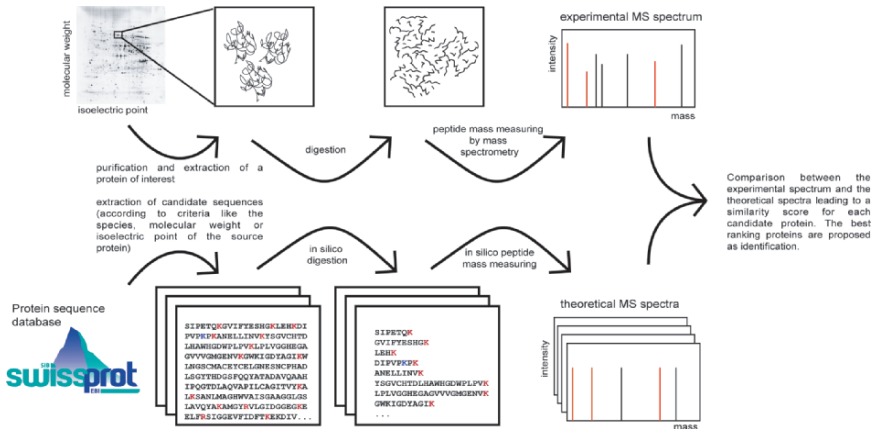


Fig. 3.7 Protein identification by peptide mass fingerprinting. Parallel processes occur experimentally and *in silico*. Finally, experimental peptide masses are compared with theoretical peptide masses to achieve protein identification

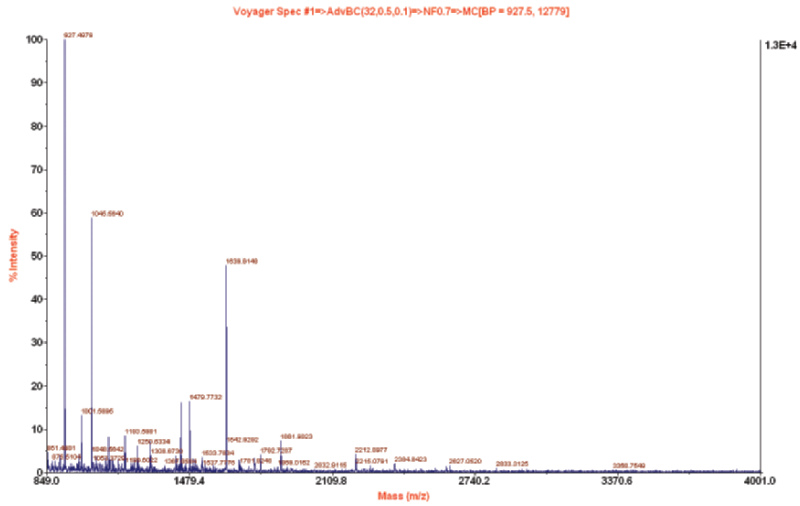


Fig. 3.8 Peptide mass fingerprinting spectrum resulting from the tryptic digestion of bovine serum albumin. The protein comes from a two-dimensional electrophoresis gel, and the spectrum was measured with a MALDI-TOF mass spectrometer in reflectron mode. Raw mass spectrometry spectra need to be processed to transform the signal into lists of monoisotopic masses. These are then inputted to protein-identification algorithms. Peak detection includes centroiding, noise filtering, calibration (shift of the m/z scale), deconvolution (determining the charge state of a peak) and deisotoping (removal of isotopic peaks)

Candidate proteins can be selected from the list of best matches on the basis of one or more attributes outlined in Sect. 3.2. Table 3.1 gives the names and Internet addresses of some current PMF tools.

Scoring systems for PMF are critical, and should take into account many factors to produce a robust score. These factors include dissimilarities in the peptide masses due to calibration errors, expected peak intensities, noise, contaminant or missing peaks, presence of post-translational modifications, and so on. A variety of different scoring functions have been implemented in various algorithms. Older tools such as FragFit (Henzel et al. 1993), PeptideSearch (Mann and Wilm 1994) and PepFrag (Fenyo et al. 1998) use simple scores based on the number of common masses between the experimental and theoretical spectra. MOWSE (Pappin et al. 1993) exploits a scoring scheme that accounts for the non-uniform distribution of protein and peptide molecular weights in databases. Similar schemes are exploited in MS-Fit (Clauser et al. 1999), Mascot (Perkins et al. 1999) and ProFound (Zhang and Chait 2000). SmartIdent (Gras et al. 1999) uses genetic algorithms to learn the scoring parameters and Aldente (Fig. 3.9) exploits Hough transform to determine the mass spectrometer deviation, to realign the experimental masses and to exclude any outliers.

An example of protein identification by PMF is shown in Fig. 3.9. Here, peptide mass values from the spectrum in Fig. 3.8 were used to search against *Bos taurus* proteins from the Swiss-Prot database. The identification tool was Aldente. Peptide mass tolerance was set to ± 25 ppm and proteins were not filtered out by pI or mass. Up to one missed cleavage was allowed. Two types of modifications were taken into account: cysteine carbamidomethylation and methionine oxidation. It can be seen that the protein was successfully identified as bovine serum albumin, with high confidence.

3.3.3.2 Identification and Characterisation of Modified Peptides by Peptide Mass Fingerprinting

Post-translational modifications are the covalent attachment of chemical groups to amino acid residues (see Chap. 5). Chemical modifications to amino acids may also be introduced deliberately (e.g. cysteine carbamidomethylation) or may happen as an artefact during sample preparation (e.g. oxidation of methionine). Most post-translational modifications are site-specific and/or position-specific. For example, oxidation may be observed on methionine, histidine and tryptophan; phosphorylation, a much studied post-translational modification, mostly occurs on serine, threonine and tyrosine, and methylation occurs at protein N-termini and on internal amino acids cysteine, histidine, lysine, asparagine, glutamine and arginine (Unimod 2006). Modifications change the masses of amino acids, and consequently the mass of any peptide that carries a modified amino acid. According to the modification type, the mass difference varies from a one to a few thousand daltons. For example, deamidation of asparagine and glutamine is characterised by a mass shift of 0.984 Da (loss of one hydrogen and a nitrogen, and gain of an oxygen), while myristoylation of glycine and lysine is characterised by a mass shift of

1) P02769 ALBU_BOVIN (C_1) Up

Serum albumin.
UniProtKB/Swiss-Prot : 9913

Score	Z-score	Mw	pI	Hits	Coverage	Shift (Da)	Slope (ppm)
82.26	396.98	66432	5.60	17	37%	0.032	-30

	Exp	Theo	Intensity			Delta		Dev	MC	CAM	MSO	PTM	Position		Sequence
	Da	Da	UI	%	rank	Da	ppm	ppm					start	end	
	1193.6	1193.602132	1	100	1	-0.00	-1	1	1	-	-		25	34	DTHKSEIAHR
	1249.62	1249.62114	1	100	1	-0.00	0	3	1	-	-		35	44	FKDLGEEHFK
	1419.69	1419.693648	1	100	1	-0.00	-2	5	-	1/1	-		89	100	SLHTLFGDELCK
	2019.97	2019.96914	1	100	1	0.00	0	15	1	1/1	-		139	155	LKPDPNTLCDEFKADEK
	2247.91	2247.942674	1	100	1	-0.03	-14	1	1	3/3	-		267	285	ECCHGDLLECADDRADLAK
	1532.77	1532.78108	1	100	1	-0.01	-6	2	1	2/2	-		298	309	LKECCDKPLLEK
	1567.73	1567.742706	1	100	1	-0.01	-7	1	-	-	-		347	359	DAFLGSFLYEYSR
	1439.8	1439.811728	1	100	1	-0.01	-7	0	1	-	-		360	371	RHPEYAVSVLLR
	1305.71	1305.7161	1	100	1	-0.01	-4	1	-	-	-		402	412	HLVDEPQNLIK
	1068.43	1068.441462	1	100	1	-0.01	-10	-10	-	1/1	-		413	420	QNCDQFEK
	1479.78	1479.79541	1	100	1	-0.02	-9	-1	-	-	-		421	433	LGEYGFQNALIVR
	1639.92	1639.937712	1	100	1	-0.02	-10	0	1	-	-		437	451	KVPQVSTPTLVEVSR
	898.48	898.481482	1	100	1	-0.00	-1	-6	-	1/1	-		483	489	LCVLHEK
*	1138.49	1138.497928	1	100	1	-0.01	-6	-4	-	2/2	-		499	507	CCTESLVNR
	1880.9	1880.921072	1	100	1	-0.02	-10	2	-	1/1	-		508	523	RPCFSALTPDETYVPK
	1907.92	1907.920738	1	100	1	-0.00	0	13	-	1/1	-		529	544	LFTFHADICTLPDTEK
	1927.79	1927.798252	1	100	1	-0.01	-3	9	1	3/3	-		581	597	CCAADDKEACFAVEGPK

View all couples (2 outliers)

```
  1  ---------- ---------- ----DTHKSE IAHRFKDLGE EHFKglvlia fsqylqqcpf dehvklvnel tefaktcvad
 81  eshagcekSL HTLFGDELCK vaslretygd madccekqep ernecflshk ddspdlpkLK PDPNTLCDEF KADEKkfwgk
161  ylyeiarrhp yfyapellyy ankyngvfqe ccqaedkgac llpkietmre kvlassarqr lrcasiqkfg eralkawsva
241  rlsqkfpkae fvevtklvtd ltkvhkECCH GDLLECADDR ADLAKyicdn qdtisskLKE CCDKPLLEKs hciaevekda
321  ipenlpplta dfaedkdvck nyqeakDAFL GSFLYEYSRR HPEYAVSVLL Rlakeyeatl eeccakddph acystvfdkl
401  kHLVDEPQNL IKQNCDQFEK LGEYGFQNAL IVRytrKVPQ VSTPTLVEVS Rslgkvgtrc ctkpesermp ctedylslil
481  nrLCVLHEKt pvsekvtkCC TESLVNRRPC FSALTPDETY VPKafdekLF TFHADICTLP DTEKqikkqt alvellkhkp
561  kateeqlktv menfvafvdk CCAADDKEAC FAVEGPKlvv stqtala
```

Link to other ExPASy tools				
Original spectrum	FindMod	GlycoMod	FindPept	PeptideMass
Recalibrated spectrum	FindMod	GlycoMod	FindPept	

Fig. 3.9 Peptide mass fingerprinting identification of bovine serum albumin with the Aldente program. Bovine serum albumin was the highest-scoring protein, with 34 peptide matches. The table contains the peptides that matched peaks of the input spectrum. The sequence coverage is displayed *below the table*. The amino acid sequences of the matched peptides are displayed in *uppercase letters* in the complete amino acid sequence of the protein. Links to additional tools on the ExPASy server are also provided to facilitate further analysis

210.198 Da (gain of 26 hydrogens, 14 carbons and one oxygen). PMF tools handle modifications by generating modified theoretical peptides according to a list of possible modifications. Some tools, like Aldente, also consider post-translational modification annotations from Swiss-Prot entries. MS spectra, as well as being an attribute for protein identification, can represent a starting point for the examination of unexpected protein modifications or processing. Peptides from an identified protein that do not match those from the database may carry post-translational or artefactual modifications, or may be mutated.

The tool FindMod has been developed to interpret mass shifts between non-matched peaks (Wilkins et al. 1999). It uses as input the MS spectrum and the identified protein sequence (or accession number) and searches possible post-translational modifications or mutations that would explain shifts between the experimentally measured peptide masses and the theoretical peptide masses calculated for the protein. If a mass difference corresponds to a known post-translational modification, rules are applied that examine the sequence of the peptide of interest and make predictions as to what amino acid in the peptide is likely to carry the modification. Typically, predictions are then confirmed using MS/MS techniques.

3.3.3.3 Limitations of Peptide Mass Fingerprinting

The PMF approach is rapid and efficient, particularly when used in automated MALDI-TOF mass spectrometers. However, it has a number of limitations, notably when:

1. Samples contain a mixture of proteins. The complexity of such spectra can result in false-positive identifications. In addition, suppression effects can interfere with the effective ionisation of certain peptides, therefore decreasing the chance of obtaining the number of peptides required for confident protein identification.
2. MS spectra are searched against large sequence databases. As the specificity of the method is based on statistics, the larger the database, the higher the number of randomly matched peptide masses.
3. The proteins carry unexpected modifications, therefore reducing the number of matching peptide masses. Similarly, sequence errors in databases, non-annotated alternative splicing products, protein-processing events and sequence mutations also reduce the number of matches.
4. The proteins under analysis are very small or very large. Very small proteins produce a very small number of peptides to be analysed. It is possible that these few peptides might not be present in a MS spectrum. In this case, the required minimum number of matching peptides might not be reached. In the case of large proteins (above about 150 kDa), the number of theoretical peptides is so large that a portion of them are likely to randomly match nearly every spectrum.
5. The protein sequence under investigation is not represented in the protein sequence database; here (a) random matches yield a low score and no significant hit is obtained, which means identification is not successful, (b) there is a protein match that corresponds to a protein with high sequence similarity, or (c) there is a random match that displays a significant score (false-positive identification). In the latter case, the user should carefully evaluate the quality of the hit, according to additional biological knowledge of the sample.

3.3.4 Tandem Mass Spectrometry Based Identification

3.3.4.1 Tandem Mass Spectrometry Spectra

MS/MS is used to fragment peptide precursor ions or parent ions into a series of smaller fragments. This yields tandem mass spectra composed of a precursor peptide and of a series of fragment peaks. The precursor peptide may be multiply charged, but the peptide fragments are singly charged in most cases. The number of peaks comprising a MS/MS spectrum varies from about ten to several hundred depending on factors like the precursor peptide length, the fragmentation quality, the mass spectrometer type and the parameters used to detect the peaks in the raw spectrum. Interpreting a MS/MS spectrum is not a straightforward procedure. Various ions can be produced during fragmentation (Fig. 3.10). Mass errors, isotopic peaks and noise also significantly complicate the interpretation.

MS/MS spectra usually contain a series of peaks that come from the successive fragmentation of amino acids in the peptide sequence. This is a key property, since small or large stretches of peptide amino acid sequence can be inferred from the mass differences between peptide fragments (Fig. 3.11).

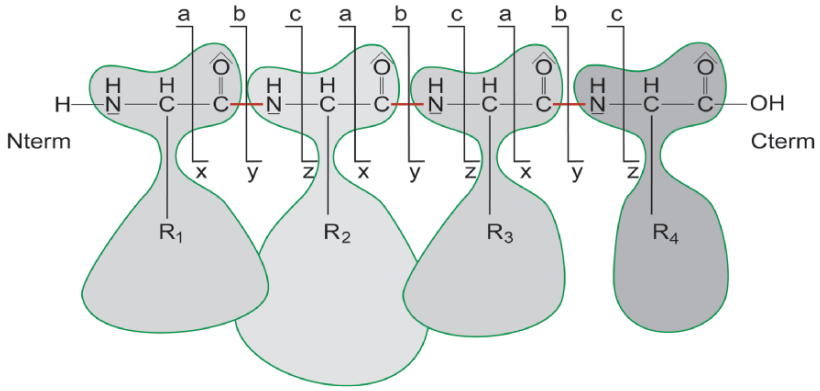


Fig. 3.10 Fragmentation of a tetrapeptide, with corresponding notation (according to Biemann 1990). When cleavage occurs at peptide amide bonds, the fragment ions are of b-ion types if the charge remains at the peptide N-terminus, or y-ion types if the charge remains at the peptide C-terminus. Only the charged portion of the peptide will be detected after fragmentation. The b-ion and y-ion fragmentation can produce ladders of fragments whose mass differences correspond to entire residues and therefore give peptide sequence information. Other fragmentation can also occur along the peptide backbone, to generate a and c ions, or x and z ions, depending on whether the ion constitutes an N-terminal or a C-terminal fragment. Additionally, when the collision energy is high (around 1 keV), additional fragment ions may also be generated, including internal fragments formed by breakage of two peptide bonds, as well as side-chain-specific ions (denoted d, v and w, not shown) formed by the loss of all or parts of amino acid side chains (Johnson 1988)

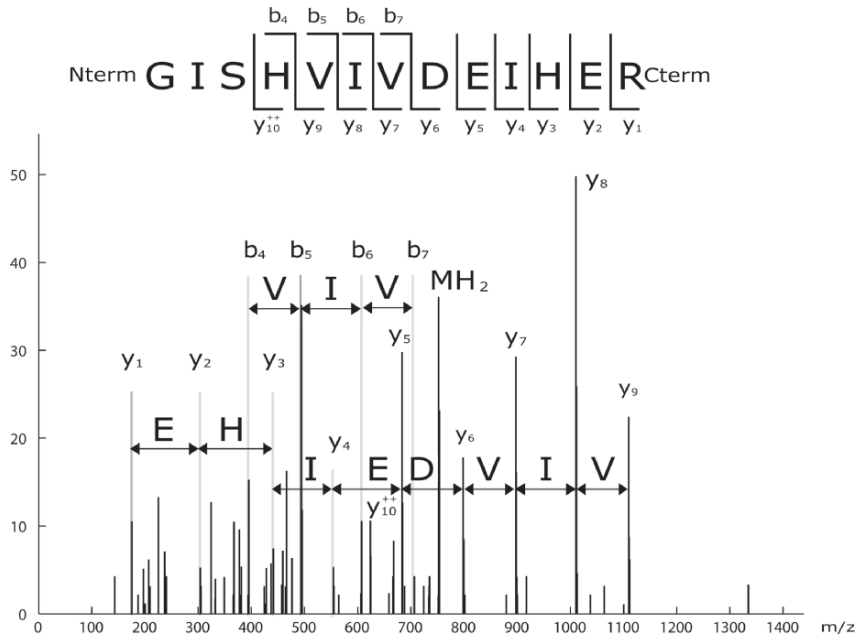


Fig. 3.11 Peptide fragment fingerprint corresponding to the peptide GISHVIVDEIHER. Peptide fragmentation is most useful when it produces a ladder of fragments, where the mass difference between each fragment (*arrows*) corresponds to that of a certain residue. In this manner, a partial sequence of a peptide can be read forward (from b ions) or backward (from y ions)

However, very high mass accuracy is required to obtain unambiguous sequence, especially to distinguish between lysine (mass 128.09 Da) and glutamine (mass 128.06 Da) residues. Note also that isoleucine and leucine are of identical mass so are not distinguishable without spectral information from specific fragmentation of their side chains.

3.3.4.2 The 'Peptide Fragment Fingerprinting' Approach

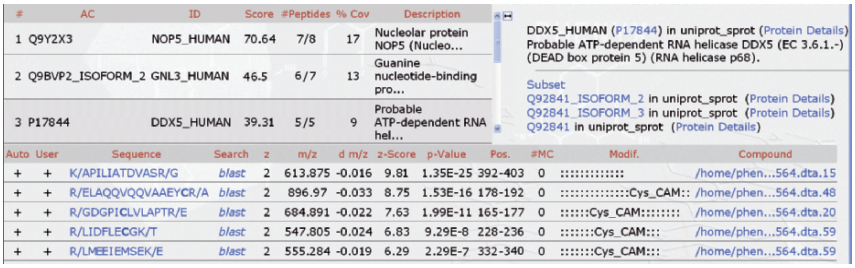
MS/MS is a widely used alternative to PMF. Firstly, as a number of peptides are usually fragmented for each protein, the identification is more robust and less equivocal. It is more robust because there is no need to identify all peptides of a given protein to achieve confident identification. It is less equivocal because the identification of several peptides for a given protein confirms its presence in the sample. In addition, MS/MS provides the opportunity to work with complex peptide mixtures. In such a case, a supplementary analysis step is necessary to produce a list of identified proteins from the set of peptides analysed. This is not as simple as one would expect, because peptide coverage

is rarely near 100% and different proteins can produce identical peptides. This is a particular issue in protein families, isoforms and splice variants. However, software tools now automatically lists all proteins carrying identical peptides, which greatly assists in resolving this problem. Last but not least, when enough coverage is obtained, MS/MS spectra can provide detailed information about peptide sequences and about possible post-translational modifications and sequence mutations.

The principle of protein identification by peptide fragment fingerprinting (PFF) is similar to that of PMF. The aim is to correlate an experimental MS/MS spectrum with virtual MS/MS spectra constructed from the theoretical digestion of proteins to peptides and an *in silico* fragmentation of these peptides. A matching score is calculated that depends on the correlation between the experimental spectrum and the virtual spectrum of the peptide being compared. The sequence of the best-matching virtual spectrum is likely to be the one that represents the sequence of the experimental peptide. As for PMF tools, the user may restrict a search to a given species or taxonomy range, thus avoiding parsing the whole taxonomic range of the database. The mass of the precursor peptide (which is produced by the mass spectrometer with the spectrum) is also used as a filter in the matching process. If available, partial sequence information, also known as a sequence tag, may be an efficient filter. As for PMF, one can consider amino acid modifications that are known to be experimentally induced, such as alkylation of cysteine residues; those that are artefactual, such as oxidation of methionine or tryptophan residues; or those that are potentially present in the native peptides, such as N-terminal formylation or acetylation or the phosphorylation of serine or threonine. Importantly, some tools can consider non-specifically cleaved peptides or amino acid mutations in a search. Users have to handle such searches with care, as these approaches dramatically increase the search space and might artificially result in an increased number of false-positive peptide matches for a number of protein entries (Sect. 3.3.4.3).

An example of protein identification with PFF is shown in Figs. 3.12 and 3.13. This was performed with the identification tool Phenyx (2004) whereby 60 MS/MS spectra obtained from the analysis of a human protein from a 2-DE gel were matched against the Swiss-Prot database. Proteins were filtered according to the protein source species (*Homo sapiens*) and the precursor mass (with error tolerance set to ±0.4 Da). Two modifications were taken into account: carbamidomethyl cysteine as a fixed modification and methionine sulfoxide as a variable modification. Enzyme specificity was trypsin, with one missed cleavage allowed. Phenyx confidently identified a total of 18 spectra and a couple of additional spectra with lower confidence. These belong to three human proteins: nucleolar protein NOP5 (seven spectra), guanine nucleotide binding protein-like 3 (six spectra) and probable ATP-dependent RNA helicase DDX5 (five spectra).

The question arising from the above example is the following: What happened with the other 43 spectra? Having such a proportion of non-identified



#	AC	ID	Score	#Peptides	% Cov	Description
1	Q9Y2X3	NOP5_HUMAN	70.64	7/8	17	Nucleolar protein NOP5 (Nucleo...
2	Q9BVP2_ISOFORM_2	GNL3_HUMAN	46.5	6/7	13	Guanine nucleotide-binding pro...
3	P17844	DDX5_HUMAN	39.31	5/5	9	Probable ATP-dependent RNA hel...

DDX5_HUMAN (P17844) in uniprot_sprot (Protein Details)
Probable ATP-dependent RNA helicase DDX5 (EC 3.6.1.-) (DEAD box protein 5) (RNA helicase p68).

Subset
Q92841_ISOFORM_2 in uniprot_sprot (Protein Details)
Q92841_ISOFORM_3 in uniprot_sprot (Protein Details)
Q92841 in uniprot_sprot (Protein Details)

Auto	User	Sequence	Search	z	m/z	d m/z	z-Score	p-Value	Pos.	#MC	Modif.	Compound
+	+	K/APILIATDVASR/G	blast	2	613.875	-0.016	9.81	1.35E-25	392-403	0	::::::::::::	/home/phen...564.dta.15
+	+	R/ELAQQVQQVAAEYCR/A	blast	2	896.97	-0.033	8.75	1.53E-16	178-192	0	::::::::::::Cys_CAM::	/home/phen...564.dta.48
+	+	R/GDGPICLVLAPTR/E	blast	2	684.891	-0.022	7.63	1.99E-11	165-177	0	::::::Cys_CAM:::::::	/home/phen...564.dta.20
+	+	R/LIDFLECGK/T	blast	2	547.805	-0.024	6.83	9.29E-8	228-236	0	:::::::Cys_CAM:::	/home/phen...564.dta.59
+	+	R/LMEEIEMSEK/E	blast	2	555.284	-0.019	6.29	2.29E-7	332-340	0	:::::::Cys_CAM:::	/home/phen...564.dta.59

Fig. 3.12 Peptide fragment fingerprinting identification result from the Phenyx program. A total of 60 spectra were analysed. The *upper table* shows the three proteins identified from the MS/MS spectra. The *lower table* shows five peptides corresponding to five spectra identified by Phenyx and belonging to the third protein P17844 (DDX5_HUMAN)

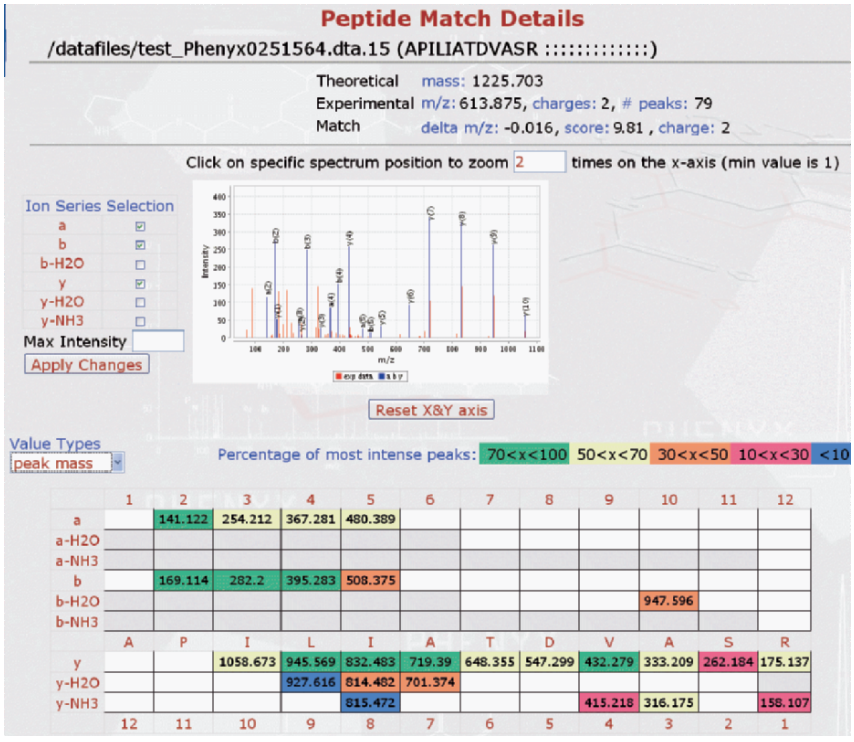


Fig. 3.13 Automatic interpretation of MS/MS spectra by the Phenyx program, showing a spectrum corresponding to peptide APILIATDVASR. *Blue peaks* are from matched ions. A selection of ion types to display can be made via the *checkboxes on the left of the spectrum*. The table shows a graphical representation of the matched peptide fragments. The peptide sequence is represented horizontally and the types vertically. Peak *m/z* values that were matched against theoretical fragments are *boxed* in various colours. The colour code is function of the relative intensity of the peaks in the spectrum. *Green* means high intensity. This example shows a high-confidence identification: the *p* value of the peptide match obtained and indicated in the protein overview is very small (1.35×10^{-25}), most of the intense peaks of the spectrum were matched and the peak matches show a well-defined y-ion series

spectra is very common in proteomics experiments. The possible reasons for a spectrum not being confidently matched to any peptide in a database are:

1. It may originate from a non-peptide contaminant.
2. It may be too noisy, or may undergo poor fragmentation, owing, for example, to the non-availability of a mobile proton to trigger fragmentation (Wysocki et al. 2000).
3. Inaccurate or incorrect precursor mass estimation may cause the identification to fail, particularly if the identification algorithm applies a precursor mass filter to select candidate peptides from the database.
4. It may be from a novel protein that is not present in the database.
5. The peptide that produced the spectrum may have resulted from an unexpected digestion event (missed cleavage or non-specific cleavage).
6. The sequence of the peptide that produced the spectrum may diverge from its corresponding sequence in the database owing to polymorphisms, mutation or transpeptidation events (Schaefer et al. 2005).

3.3.4.3 De Novo Sequencing

Another application of MS/MS is for the de novo sequencing of peptides. In this approach, the peptide sequence is inferred directly from the spectrum, independently of any information extracted from a pre-existing protein or DNA sequence database. Inference of sequences is done using specifically developed search algorithms. Since they do not use database information during spectrum interpretation, de novo sequencing algorithms work in a search space composed of the set of all possible sequences that can be represented by the spectrum. The only restriction is the mass of the parent ion. Owing to the large size of this search space, de novo sequencing methods are disadvantaged compared with PFF methods. They require spectra of higher quality with smaller fragment errors and a more or less continuous fragmentation pattern, or at least high-quality fragmentation for several adjacent amino acids. Spectra with poor fragmentation are very hard or even impossible to interpret and thus to generate a sequence. Despite these disadvantages, de novo methods may surpass PFF methods when searching genome databases that are subject to sequencing errors, when searching databases composed of non-identical homologous sequences and when analysing a spectrum that originates from a mutated protein. This is because de novo algorithms extract sequences from the spectrum that include the amino acid replacements, which are then handled by the similarity search algorithm by allowing mismatches between the de novo sequences and the database sequences.

De novo sequencing algorithms try to infer a complete sequence for a peptide of interest. However, low-quality fragmentation data can make it difficult for all amino acids in a sequence to be assigned. In such cases, missing

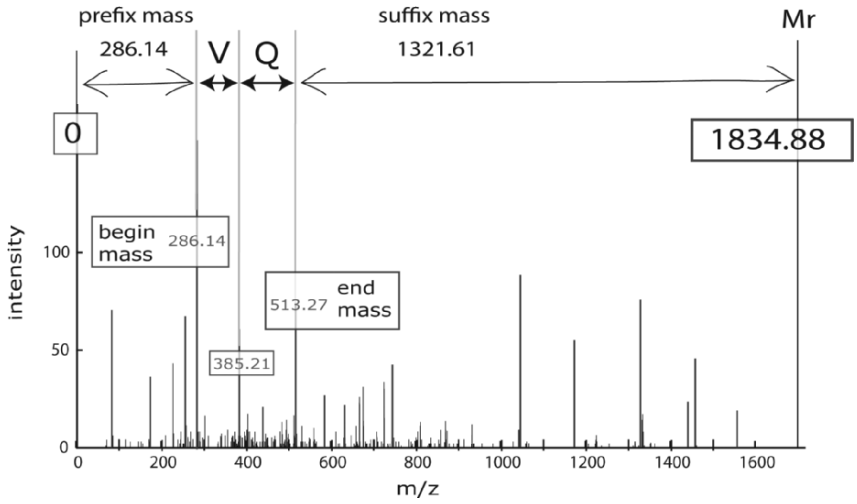


Fig. 3.14 A 'sequence tag' inferred from a MS/MS spectrum. The tag is composed of two flanking masses, called prefix and suffix masses, and of a short sequence VQ. Masses of the prefix and suffix regions are easily computed from the first and last masses composing the tag (denoted as *begin mass* and *end mass*) and from the precursor ion mass

fragmentation positions in the spectrum are handled by using combinations of two or three amino acids. When more than a few consecutive fragmentation positions are missing in the spectrum, or in the case of unexpected modifications, the algorithm may not be able to propose a solution, or the produced sequence may contain wrong amino acids. But for the purpose of protein identification, complete interpretation of the spectra is in most cases unnecessary. A few residues of the sequence obtained from a spectrum, called a 'peptide sequence tag' (Mann and Wilm 1994; Fig. 3.14), combined with the precursor ion mass are usually sufficiently specific to confidently identify a peptide. The reason why peptide sequence tags are not always used as a first step in identification is that automated tag extraction is not a straightforward procedure because of the huge number of false-positive tags that can be inferred from the peak patterns. Frank et al. (2005) state that the set of sequence tags for any peptide should be as small as possible but should provide adequate sequence coverage (at least one of the tags has to be correct, so that the true peptide is not filtered out).

3.3.4.4 Identification and Characterisation of Peptides with Unexpected Modifications

In a MS/MS experiment, an appreciable number of spectra are usually not confidently identified, even though they are of high quality. Many of these

spectra are likely to be from peptides that carry post-translationally modified amino acids. The presence of a post-translational modification in a peptide sequence alters the fragmentation pattern and therefore renders the interpretation of the corresponding MS/MS spectrum more difficult. The modification may, for example, alter the peptide fragmentation because of differing biochemical or physical properties (e.g. steric hindrance). Modifications will also alter the mass of peptide fragments compared with the mass of a theoretical unmodified peptide of the same sequence. Figure 3.15 illustrates how a modification on an amino acid may affect a fragmentation spectrum.

The identification of modifications using peptide fragmentation requires special strategies. One consideration is that the parent ion masses of modified peptides will be different from those of their unmodified counterparts in sequence databases. A second consideration is that fragmentation spectra from a modified and an unmodified peptide, when compared, have a number of fragments that will not match (on average, 50% of the masses are shifted). Consequently, and similar to PMF methods, PFF methods have to generate virtual spectra of all post-translational modification variants computed from the theoretical sequences. By this means, the shifted peaks can be correctly matched during the comparison process. This method requires the user to specify a list of modifications that are anticipated to appear on the peptide. The greater the number of possible modifications, the higher the number of potential peptides. This will affect the computing time, and will increase the number of false-positive identifications for a peptide.

But what happens when a peptide carries an unexpected modification or a modification which has yet to be chemically defined? Popitam (Hernandez et al. 2003; Fig. 3.16) is a software tool designed to perform 'open-modification searches'. In other words, it can take into account any type and number of differences between a MS/MS spectrum and theoretical peptides from a database.

3.3.4.5 Spectral Library Searches

In shotgun proteomics experiments, for example the analysis of human plasma, the same type of sample has often been analysed many times in either the same or a different laboratory. Spectra and peaklists from these experiments, when submitted to sequence search tools, result in the same peptides being identified again and again. Nothing is wrong with this in principle, except that sequence search algorithms must scrutinise a search space that is much bigger than the set of peptides that are likely to be identified. Each search spends most of its time reidentifying the same molecules. In order to make this step more efficient, and thus free computing capacity for other tasks, spectral library search tools have recently been adopted (Yates et al. 1998; Craig et al. 2006; Frewen et al. 2006; Lam et al. 2007). The principle is to identify peptides by matching new experimental spectra with those stored in

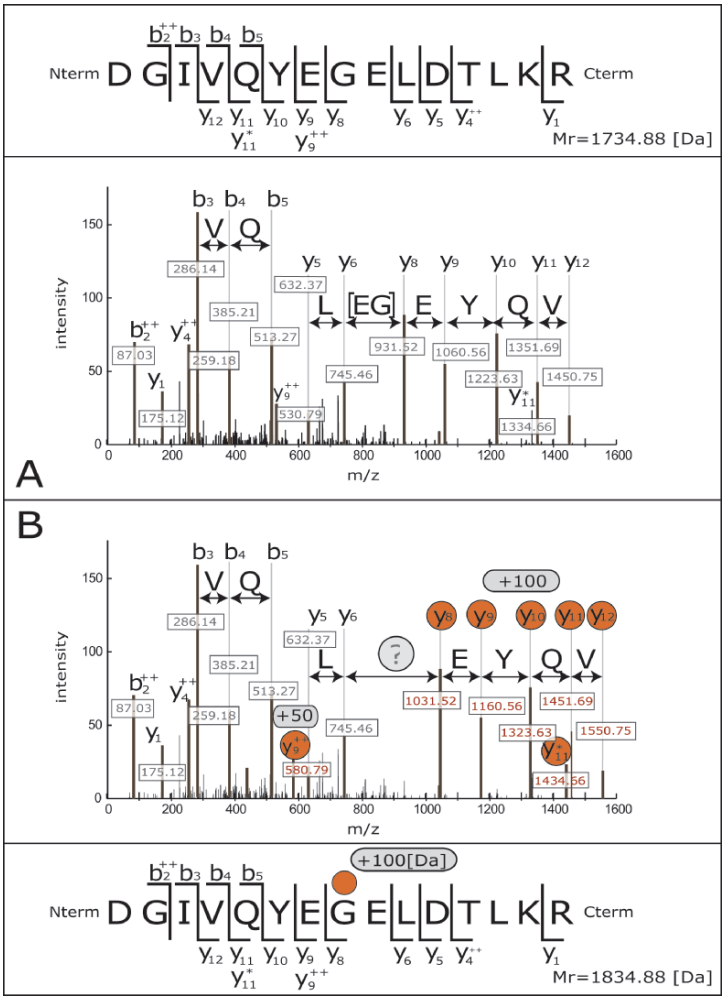
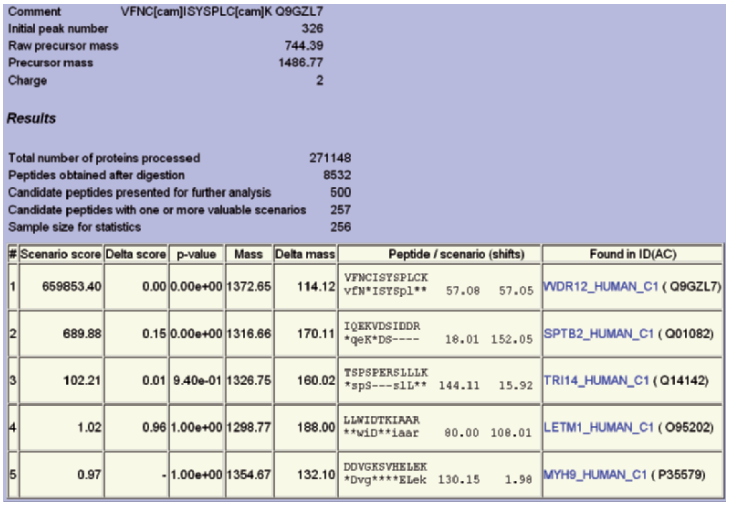


Fig. 3.15 Fragmentation mass spectra of modified and unmodified peptides. *A* An annotated spectrum of peptide DGIVQYEGELDTLKR. Major ion assignments are indicated, as well as the corresponding peak masses and sequence information that can be inferred from peak differences. *B* The same spectrum, in which a 100-Da modification has been simulated on the second glycine of peptide DGIVQYEGELDTLKR. *Peaks marked with a circle* represent fragments that carry the modification and are thus shifted by a δ value from their expected position. It should be noted that the shifted peaks are not necessarily grouped in the spectrum and that the shift in the mass will depend on the number of charges on each fragment. Moreover, according to the type of modification (gain or loss of atoms), peaks may be shifted either to the left or to the right. In *B*, sequence information is partly lost, as the amino acids GE can no longer be read from the mass difference between ions y6 and y8



Comment	VFNC[cam]ISYSPLC[cam]K Q9GZL7
Initial peak number	326
Raw precursor mass	744.39
Precursor mass	1486.77
Charge	2

Results

Total number of proteins processed	271148
Peptides obtained after digestion	8532
Candidate peptides presented for further analysis	500
Candidate peptides with one or more valuable scenarios	257
Sample size for statistics	256

#	Scenario score	Delta score	p-value	Mass	Delta mass	Peptide / scenario (shifts)			Found in ID(AC)
1	659853.40	0.00	0.00e+00	1372.65	114.12	VFNCISYSPLCK vfN*ISYSpl**	57.08	57.05	WDR12_HUMAN_C1 (Q9GZL7)
2	689.88	0.15	0.00e+00	1316.66	170.11	IQEKVDSIDDR *qeK*DS----	18.01	152.05	SPTB2_HUMAN_C1 (Q01082)
3	102.21	0.01	9.40e-01	1326.75	160.02	TSPSPERSLLLK *spS---slL**	144.11	15.92	TRI14_HUMAN_C1 (Q14142)
4	1.02	0.96	1.00e+00	1298.77	188.00	LLWIDTKIAAR **wiD**iaar	80.00	108.01	LETM1_HUMAN_C1 (O95202)
5	0.97	-	1.00e+00	1354.67	132.10	DDVGKSVHELEK *Dvg****ELek	130.15	1.98	MYH9_HUMAN_C1 (P35579)

Fig. 3.16 Output of Popitam, a program to find protein post-translational modifications in MS/MS data. A spectrum with two carbamidomethylated cysteine residues was presented to Popitam. No information was given according to the modification type, except an allowed mass range of 0–200 Da. The best-scoring scenario (*first row*) indicates that a mass of 57.05 Da should be added to the first cysteine and a mass of 57.04 Da should be added to the second cysteine, the terminal lysine or the C-terminal group

a library of well-characterised spectra. Such libraries and search tools have been used for many years to identify small molecules measured with gas chromatography–MS, where the electron impact ionisation method delivers very reproducible spectra from high-energy fragmentation. Advantages of spectral library searches over sequence searches (such as PFF) include:

1. A greatly reduced search space and a quantum gain in speed. The number of peptides searched corresponds to the number of spectra in the library. There is no need to use computationally costly parameters such as a high number of missed cleavages, partial or non-specific cleavage rules or consider variable modifications.
2. An easier scoring model. As the library of spectra contains experimental spectra, a certain level of spectra conservation is assumed for each peptide (as a function of its charge state and instrument settings). A similarity measure between spectra might therefore be sufficient to score a match.

Potential pitfalls of the method, however, include:

1. The library itself may be of inconsistent quality and be incomplete. The library has to be compiled with high-confidence identifications, and this may require manual curation. There is no chemical library of all possible peptides

available and current libraries are built from the result of experimental data and searches. Libraries will, therefore, always be incomplete.
2. Issues with instrument- and experiment-specific libraries. Each library is made of spectra generated by specific instruments, acquired under specific conditions. As fragmentation patterns are different from one instrument to another and under different acquisition settings, care has to be taken regarding the choice and use of each library.
3. The algorithms available in existing tools are relatively simple and are accurate for the current small size of the libraries (up to a few tens of thousands of spectra per species, where available). However, with increasing size of the libraries, care will be required to avoid an increased number of false positives.

The spectral library search method is well suited to rapidly identify and filter already known peptides and proteins. It is therefore primarily dedicated to targeted proteomics experiments, and therefore cannot, in its current form, be used to discover new peptides and proteins.

3.4 List of Tools and URLs

A number of protein-identification and protein-characterisation programs available on the Internet are listed in Table 3.1. These programs are grouped according to the protein attributes they use to query the database. Tools written in bold characters are developed by the Swiss Institute of Bioinformatics (SIB) and GeneBio in Geneva. Those developed by the SIB are made available through the ExPASy server (ExPASy Proteomics Server 1993; Gasteiger et al. 2005). This server is provided as a service to the life-science community by a multidisciplinary team at the SIB. It gives access to a variety of databases and analytical tools dedicated to proteins and proteomics, including similarity searches, pattern and profile searches, post-translational modification prediction, topology prediction, primary, secondary and tertiary structure analysis and sequence alignment.

3.5 Concluding Remarks

MS/MS identification has become a mature and robust analytical proteomic technique. To achieve this, it has required impressive research on several fronts: separation techniques, MS, computer science and database development. Different groups have tackled the issue of protein identification from different angles, resulting in specialised techniques that can efficiently address the numerous difficulties arising during protein identification.

Nowadays, particular attention is given to the identification of modified peptides, to the correlation of de novo peptide sequences with homologous proteins and to spectrum modelling. Recently, new tools have combined several strategies, and multistep identification procedures are starting to appear. This is undoubtedly a useful strategy to help achieve high-performance identification. Future work has to focus on improving scoring schemes even more, in order to reduce the number of false-negative and false-positive identifications. The availability of newly developed tools, the emergence of open-source projects and the unification of MS/MS spectrum and database formats will hopefully boost the development of global identification systems.

References

Biemann K (1990) Nomenclature for peptide fragment ions (positive ions). Methods Enzymol 193:886–887

Clauser KR, Baker P, Burlingame AL (1999) Role of accurate mass measurement (+/– 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem 71:2871–2882

Compute pI/MW (1994) Swiss Institute of Bioinformatics, Geneva. http://www.expasy.org/tools/pi_tool.html. Cited 22 Mar 2007

Craig R, Cortens JP, Fenyo D, Beavis RC (2006) Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 5:1843–1849

Eckerskorn C, Jungblut P, Mewes W, Klose J, Lottspeich F (1988) Identification of mouse brain proteins after two-dimensional electrophoresis and electroblotting by microsequence analysis and amino acid composition analysis. Electrophoresis 9:830–838

ExPASy Proteomics Server (1993) Swiss Institute of Bioinformatics, Geneva. http://www.expasy.org. Cited 22 Mar 2007

Fenyo D, Qin J, Chait BT (1998) Protein identification using mass spectrometric information. Electrophoresis 19:998–1005

Frank A, Tanner S, Bafna V, Pevzner P (2005) Peptide sequence tags for fast database search in mass-spectrometry. J Proteome Res 4:1287–1295

Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ (2006) Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal Chem 78:5678–5684

Gasteiger E, Hoogland C, Gattiger A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed) The proteomics protocols handbook. Humana, Totowa, pp 571–607

Gras R, Muller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanchez JC, Bairoch A, Hochstrasser DF, Appel RD (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. Electrophoresis 20:3535–3550

Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc Natl Acad Sci USA 90:5011–5015

Hernandez P, Gras R, Frey J, Appel RD (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics 3:870–878

hyphen MassSpec Consultancy (2006) LC-MS, a solution to all analytical problems? http://www.hyphenms.nl/. Cited 22 Mar 2007

James P, Quadroni M, Carafoli E, Gonnet G (1993) Protein identification by mass profile fingerprinting. Biochem Biophys Res Commun 195:58–64

Johnson RS, Martin SA, Biemann K (1988) Collision-induced fragmentation of (M+H)+ ions of peptides. Side chain specific sequence ions. Int J Mass Spectrom Ion Process 86:137–154

Jungblut P, Dzionara M, Klose J, Wittmann-Leibold B (1992) Identification of tissue proteins by amino acid analysis after purification by two-dimensional electrophoresis. J Protein Chem 11:603–612

Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R (2007) Development and validation of a spectral library searching method for peptide identification from tandem mass spectrometry. Proteomics (in press)

Latter GI, Burbeck S, Fleming J, Leavitt J (1984) Identification of polypeptides on two-dimensional electrophoresis gels by amino acid composition. Clin Chem 30:1925–1932

Mann M, Wilm M (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 66:4390–4399

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

Pappin DDJ, Hojrup P, Bleasby AJ (1993) Rapid identification of proteins by peptide-mass finger printing. Curr Biol 3:327–332

Perkins DN, Pappin DDJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567

Phenyx (2004) Geneva Bioinformatics (GeneBio), Geneva. http://www.phenyx-ms.com. Cited 22 Mar 2007

Schaefer H, Chamrad DC, Marcus K, Reidegeld KA, Bluggel M, Meyer HE (2005) Tryptic transpeptidation products observed in proteome analysis by liquid chromatography-tandem mass spectrometry. Proteomics 5:846–852

Scripps Center for Mass Spectrometry (2005) A history of mass spectrometry. http://masspec.scripps.edu/MSHistory/mshisto.php. Cited 22 Mar 2007

Unimod (2006) http://www.unimod.org/. Cited 22 Mar 2007

Wilkins MR, Ou K, Appel RD, Sanchez JC, Yan JX, Golaz O, Farnsworth V, Cartier P, Hochstrasser DF, Williams KL, Gooley AA (1996a) Rapid protein identification using N-terminal "sequence tag" and amino acid analysis. Biochem Biophys Res Commun 221:609–613

Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, Williams KL, Hochstrasser DF (1996b) From proteins to proteomes: large-scale protein identification by two-dimensional electrophoresis and amino acid analysis. Biotechnology 14:61–65

Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF (1999) Protein identification and analysis tools in ExPASy server. Methods Mol Biol 112:531–552

Wysocki VH, Tsaprailis G, Smith LL, Breci LA (2000). Mobile and localized protons: a framework for understanding peptide dissociation. J Mass Spectrom 35:1399–1406

Yates JR III, Speicher S, Griffin PR, Hunkapiller T (1993) Peptide mass maps: a highly informative approach to protein identification. Anal. Biochem 214:397–408

Yates JR III, Morgan SF, Gatlin CL, Griffin PR, Eng JK (1998) Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. Anal Chem 70:3557–3565

Zhang W, Chait BT (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. Anal Chem 72:2482–2489

4 Quantitation in Proteomics

Garry L. Corthals and Keith Rose

Abstract

The large-scale systematic quantitation of proteins is an important component of proteomics and has contributed to the emergence of the new fields of systems biology and molecular medicine. Quantitative proteomics is anticipated to provide new insights into biological function, facilitate the identification of diagnostic or prognostic disease markers, and contribute to the discovery of proteins as therapeutic targets. This field is growing rapidly and new applications and protocol design are driving technological innovations. Due to rapidly evolving innovations the field of quantitation is wide, and deep! So many topics cannot be discussed in such a short chapter. We advise the reader to gain an understanding of the computational procedures required for the processing of quantitative MS data, and ideally information on the various types of instruments that are best used. In this section we will provide a modest overview of the types of strategies employed in proteomics that rely on mass spectrometry for protein identification and concurrent quantitation. Throughout this review, comparative analysis is a recurring theme and one which is common to most quantitative techniques in proteomics.

4.1 Introduction

The systematic identification of the proteins in complex samples and the determination of their quantity or quantitative change are an important component of proteomics and the emerging field of systems biology. Quantitative protein profiling is anticipated to provide new insights into the function of biological processes, facilitate the identification of diagnostic or prognostic disease markers, and contribute to the discovery of proteins as therapeutic targets. In this chapter we discuss qualitative and quantitative approaches in proteomics, emphasising those that rely on mass spectrometry (MS) for protein identification and concurrent quantitation. The chapter is divided into three sections. Firstly, non-mass-spectrometric approaches to quantitation are reviewed. Secondly, relative and absolute quantitation is

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

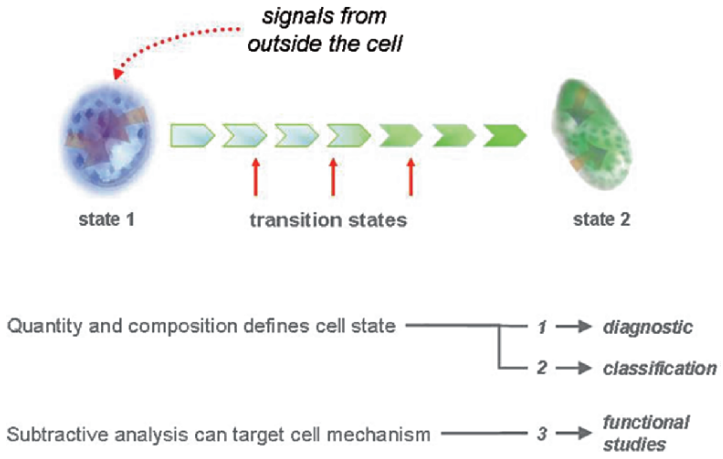


Fig. 4.1 Large-scale quantitative methods will have an impact in areas such as rapid and accurate diagnosis of patients and classification of diseases such as cancer. Subtractive analysis can produce leads to focus hypothesis-driven functional studies

discussed and how chemical, enzymatic, and metabolic labelling of proteins, and even label-fee approaches can facilitate quantitative MS measurements. Finally, we briefly look at our capacity to perform qualitative and quantitative analysis of the protein modifications glycosylation, phosphorylation, and ubiquitinylation.

A major application of quantitation is comparative analysis. Comparative analysis has three main components. Firstly, by defining the quantity and composition of a cell, one can elucidate diagnostic information (Fig. 4.1). Secondly, in larger-scale studies, quantity and composition also serve as a means for classification of cells or patient samples. Thirdly, quantitative measurements provide a means for subtractive analysis where one can target a cell mechanism, which in turn provides information about the function of those proteins that are upregulated or downregulated. Throughout this review, comparative analysis is a recurring theme and one which is common to most quantitative techniques in proteomics.

4.2 Non-mass-spectrometric Approaches to Quantitation

For approximately 25 years, quantitative expression of proteins has been undertaken by comparing two-dimensional electrophoresis (2-DE) results from two or more different cell populations (e.g. normal versus diseased). 2-DE is still successfully used for this purpose. Recently, Görg et al. (2004) comprehensively reviewed the 2-DE and MS workflow for proteomics. Their

review included the topics of sample preparation, protein solubilisation, and prefractionation; protein separation by 2-DE with immobilised pH gradients; protein detection and quantitation; computer-assisted analysis of 2-DE patterns; protein identification and characterisation by MS; and finally 2-DE protein databases. To successfully use 2-DE technology, one must pay careful attention to detail and, importantly, must define the biological and analytical replicates required for each experiment (Hunt et al. 2005). With 2-DE, semi-quantitative differences in expression can be revealed and the target proteins readily identified using MS.

To make quantitation by 2-DE more accurate, protein-reactive cyanine dyes have been developed and are used to undertake difference gel electrophoresis (DIGE). These have been recently reviewed by Van den Bergh and Arckens (2004) and Marouga et al. (2005). These reactive dyes for proteomics were initially developed at Carnegie Mellon University and then commercialised by Amersham Biosciences (now GE Healthcare). The first set of dyes react with amino groups (essentially side chains of lysine) and are used to 'minimally' label 1–5% of the proteins present in a pair of samples (Fig. 4.2). A different dye is used for each sample. After mixing the two

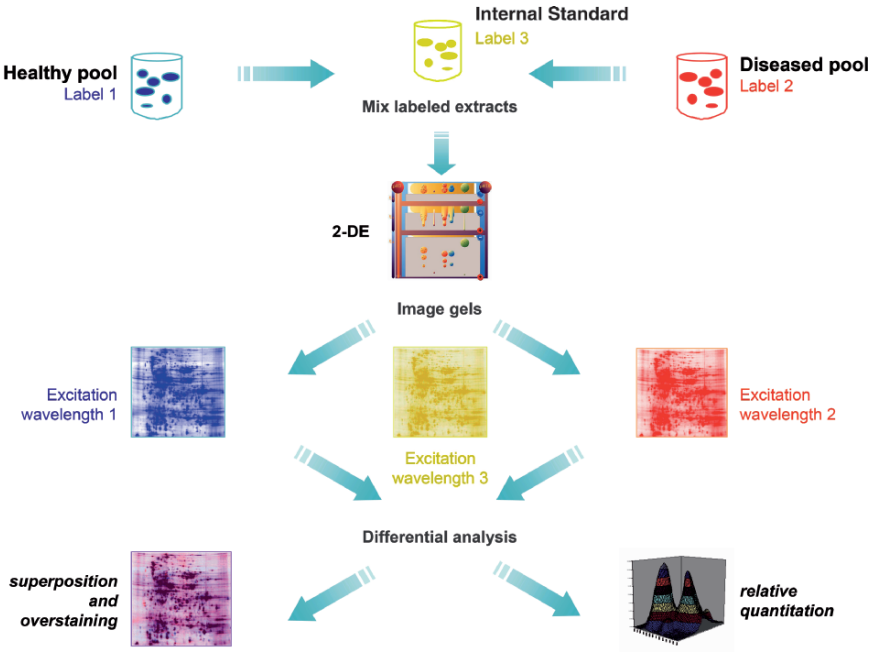


Fig. 4.2 Currently the simplest means for 'large-scale' (up to thousands) relative quantitative analysis: difference gel electrophoresis (DIGE). *2-DE* two-dimensional electrophoresis. (Courtesy of GE Healthcare)

labelled samples, the pooled sample is then run on a single gel. As the two dyes have different fluorophors, proteins present in one sample can be distinguished from the same proteins present in a second sample. Small irreproducibilities which occur when a gel is run and compared with another gel are thus compensated for, as both samples are run on the same gel. Scanning the gels according the absorption and emission characteristics of each dye, followed by image analysis, allows relative quantitation to be performed. In order to identify the proteins, the corresponding spots are generally excised for MS analysis, and this is most accurately performed after overstaining with Sypro stains or Coomassie blue. This will locate the majority of a protein, which is unlabelled and thus of lower mass, and does not migrate to exactly the same position as the dye-labelled form.

The structures of two of the initial Cy dyes are shown in Fig. 4.3. They react via their *N*-hydroxysuccinimide ester groups with amino groups on proteins, particularly on the amino acid lysine. This means that the amount of lysine in a protein will affect the degree to which it stains and can be detected. Note that the increased mass of the longer conjugated linker on Cy5 (+26 Da) is almost compensated by the change of a propyl to a methyl group (−28 Da) on one of the nitrogen atoms. If chromatographic separations are employed, in the absence of the detergent sodium dodecyl sulfate (SDS), loss of the protonatable side chain of lysine due to acylation with the reagent is partially compensated by the basic nature of the dye. Nevertheless, approximately 500 Da is added to the mass of a protein with the addition of each dye molecule. This is of significance for smaller proteins, as the dye-labelled portion (only 1–5% of the total protein) will migrate notably more slowly than the non-labelled portion. While the DIGE technique does offer clear advantages over multiple gels and conventional staining, and can be multiplexed, there can be a problem if prefractionation of proteins is required in order to increase dynamic range. Labelling should be performed as soon as possible in the process to avoid preferential losses prior to pooling of the labelled samples (Fig. 4.2);

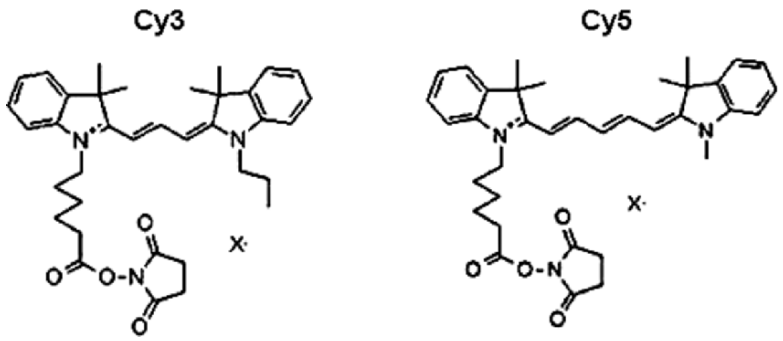


Fig. 4.3 Structures of the Cy3 and Cy5 DIGE reagents. (Figure supplied by GE Healthcare)

however, this means that abundant proteins, such as albumin and immunoglobulins in mammalian plasma, are present and labelled with the expensive reagents. If fractionation of proteins is performed prior to 2-DE, such as depletion of abundant proteins, oxidative or photolytic damage to the fluorophors can occur in a selective manner. The Cy5 dye is more susceptible to this damage than the Cy3 dye.

A set of 'saturation' dyes have recently become available (GE Healthcare) which react with the thiols of cysteine in reduced proteins. This reaction may be forced essentially to completion so that one labelled species of each protein, with all cysteine residues alkylated, is formed. This is in contrast with minimal labelling (1–5%) of proteins via lysine residues which yields zero-derivative protein as a major species, several isomers of mono-derivative protein, small amounts of various di-derivatives and so on. These will not all migrate to exactly the same position on the 2-DE gel. Overstaining with SYPRO ruby or Coomassie blue to locate the bulk of unlabelled protein, necessary with the minimal labelling, is not necessary when saturation labelling is used. The issues of when to label in an experiment, the presence of abundant proteins, and possible preferential damage to dyes during fractionation, however, remain. The lysine-reactive minimal-labelling procedure has been compared with the cysteine-reactive saturation labelling (Greengauz-Roberts et al. 2005; Shaw et al. 2003) and it was shown that saturation labelling was superior and has a 5–50-fold lower sample requirement. It does, however, require the presence of cysteine in a protein, which is a less common amino acid than lysine. While the DIGE approach is normally used in a different context from metabolic stable-isotope labelling and 'chemical tagging' (discussed later), these techniques are starting to be compared. Choe et al. (2005) compared isobaric tagging with non-DIGE 2-DE and found the mass-spectrometric method to give slightly a lower coefficient of variation, while Kolkmann et al. (2005) compared DIGE with metabolic isotope labelling and liquid chromatography (LC)–MS. In these initial comparisons, all methods give low coefficients of variation and are complementary. However, DIGE has the advantage that it is applicable to samples from large organisms, which is not the case with metabolic labelling. It is also more applicable to proteins which exist in several isoforms – these protein forms are often separable on a gel, whereas the purely peptide based LC-MS/MS approach identifies the proteins through peptides alone (Rose 2005).

With relative-abundance measurements, one typically adopts a discovery-based approach where unknown proteins, as well as known proteins, might be identified in the course of a proteomics project. By contrast, hypothesis-driven or systems-oriented proteomics often aims to systematically analyse a subset of proteins, such as a family of proteins related by sequence or a collection of proteins related by function. In this area, protein arrays could prove to be very powerful. Protein microarrays now exist that are chemically robust and stable, have a high binding efficiency and specificity, and are compact (Stoll et al. 2005). These arrays include antibody, phage displayed

antibody or polypeptide recognition moieties and allow the global analysis of distinct samples or subproteomes under carefully defined and controllable conditions, such as pH, temperature, ionic strength, and the presence or absence of cofactors. The ability to test the binding of substrates and define optimal experimental conditions is very attractive and facilitates reproducible large-scale screening of biochemical activity and protein interactions (Stoll et al. 2005; Zhu et al. 2000). They therefore offer a means for the functional analysis of native proteins. A good example of the potential of arrays is the characterisation of the protein binding partners of calmodulin; a conserved membrane protein and secondary messenger for growth, differentiation, and membrane trafficking. Zhu et al. (2001) arrayed 5,800 different proteins comprising 93% of yeast open reading frames and found six known and 33 novel calmodulin binding partners. While array-based proteomics is still in its infancy, it has incorporated concepts from microarray technology in the construction of what are most accurately called 'protein-detecting microarrays' (Kodadek 2001 2002; MacBeath 2002). A protein-detecting microarray comprises many different affinity reagents (frequently antibodies) arrayed at high spatial density on a solid support. Each reagent captures its target protein from a complex mixture, such as serum or a cell lysate, and the captured proteins are subsequently detected and quantified (MacBeath 2002). As this topic exceeds the scope of this chapter, we refer the reader to other excellent reviews of antibody arrays (Barry and Soloviev 2004) or protein-binding aptamer arrays. The latter appear to offer certain advantages over antibodies, although there is little proteomics experience with these arrays to date (Stadtherr et al. 2005).

4.3 Relative Quantitation by Mass Spectrometry

MS-based quantitation is an important addition to quantitation by 2-DE and protein chips. It has gained considerable popularity in the last 10 years. MS-based techniques are automated, can be high-throughput, and are thus potentially attractive for comprehensive and precise quantitative studies. They provide additional information to that provided by the 2-DE methods mentioned in the previous section, notably the identity of each protein or peptide which is under measurement. MS-based quantitative methods are essentially an application of stable isotope labelling, originally used for absolute measurements (De Leenheer and Thienpont 1992). An important breakthrough was made in relative quantitation in 1999 when Aebersold and colleagues (Gygi et al. 1999) first described isotope-coded affinity tagging (ICAT). This captured the attention of many researchers and initiated the quest for similar methods. Important contributions made over the past few years for quantitative proteomics have relied on the integration of specialised LC with electrospray

ionisation–MS, or matrix-assisted laser desorption/ionisation (MALDI)–MS workflows, although no industry-wide approach currently exists. As a result, there still is a need for generally applicable quantitative methods that can be employed and integrated with various separation techniques, such as 1-DE, 2-DE and LC. After about 5 years of active developments, we can now divide the MS-based strategies for relative quantitative into three conceptual approaches (Fig. 4.4). These are:

1. Chemical incorporation or 'tagging', where chemical modification of proteins in a site-specific manner is performed using a derivatisation reagent (Cagney and Emili 2002; Chakraborty and Regnier 2002; Gehanne et al. 2002; Goodlett et al. 2001; Guo et al. 2002; Gygi et al. 1999; Hamdan and Righetti 2002; Munchbach et al. 2000; Niwayama et al. 2001; Peters et al. 2001; Qiu et al. 2002; Wang and Regnier 2001).
2. Biological or metabolic incorporation, where labelling of the peptide/protein is achieved by growing cells in media enriched in stable isotope-containing amino acids (Conrads et al. 2001; Oda et al. 2001; Ong et al. 2002).
3. Enzymatic incorporation, where labelling of proteins is achieved during their enzymatic digestion to peptides. The digestion incorporates an ^{18}O from

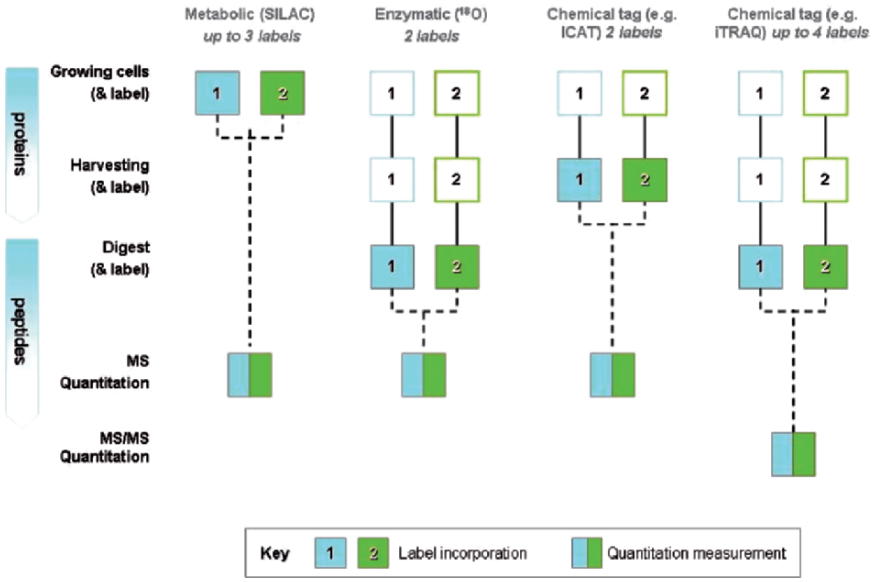


Fig. 4.4 Incorporation of isotopes into proteins and their use in relative quantitation. *Single-colour boxes* show at which step labelling takes place and *mixed-colour boxes* indicate where quantitative measurements occur. Different labelling strategies have areas where procedural errors could occur, or where the capacities of current instruments are overwhelmed by the number of peptides that need to be measured

isotopically heavy water into the C-terminus of peptides at the cleavage site (Heller et al. 2003; Mirgorodskaya et al. 2000; Rose et al. 1983; Yao et al. 2001).

Much of this work has previously been discussed and we refer the reader to recent reviews of MS-based quantitation in proteomics by Hamdan and Righetti (2002) or Sechi and Oda (2003). Developments in recent years that have not yet received widespread appreciation involve the use of methods for extracting changes in protein expression without the use of stable isotope labelling. This approach is discussed in Sect. 4.3.5.

4.3.1 Absolute or Relative Quantitation?

There is an important distinction between relative and absolute quantitation. With relative quantitation, such as the DIGE technique or isotope labelling (Sect. 4.3.2), semiquantitative information is obtained about the relative abundance of one or more proteins from samples sourced from two or more experimental conditions. However, information is not obtained about the actual amount of protein present in gram or molar terms, or its precise concentration. With absolute quantitation, first reported by De Leenheer and Thienpont (1992) and more recently by Gerber et al. (2003), actual quantities or concentrations of proteins or peptides can be determined for unknown samples by determining the exact ratio of the unknown to a known amount of a chemically synthesised control peptide. Later, a brief outline of absolute quantitation will be given. First, however, we will discuss mass-spectrometric approaches for relative quantitation.

To discover biologically relevant differences between two samples by MS, one sample is typically enriched with a stable heavy isotope in a specific way. The other sample is either not enriched or is labelled specifically with a light isotope. The non-radioactive heavier isotopes (^{2}H, ^{15}N, ^{13}C) may be introduced as chemical tags during cell metabolism, may be incorporated enzymatically (most often as ^{18}O introduced during proteolysis), or may be due to differences in isotopic composition of a covalently coupled ‘tag’. In all cases, the light-labelled and the heavy-labelled samples are finally mixed in equal proportion, prior to MS analysis. The ratio of light-labelled to heavy-labelled peptides is then measured with the mass spectrometer, yielding data for relative quantitation between the samples. Whatever incorporation method is used, it is possible to repeat the experiment with the heavy/light reagents reversed (Wang et al. 2002), thus confirming any observed differences.

4.3.2 Introduction of Stable Isotopes Using Chemical Tags

There are various approaches for relative quantitation in proteomics through use of reagents containing stable isotopes. As described above, these permit

mass differences to be observed between peptides from two different cell states. Of course, a mass difference may be generated by tagging with chemically different species, such as acetyl and propanoyl groups, but differences in physicochemical properties (in this case hydrophobicity) may lead to unwanted separations of the differentially labelled peptides during LC-MS. The analysis is thus limited to direct MALDI MS.

The general strategy for relative quantitation with stable isotopes, regardless of the chemistry used for incorporation, is as follows. Proteins or peptides present from one biological state are chemically tagged with a reagent that has a normal or 'light' isotopic distribution of ^{1}H, ^{12}C, ^{14}N, or ^{16}O. In parallel, proteins or peptides from a second biological state are tagged with the same reagent that contains the 'heavy' form the same reagent, where 'heavy' indicates the isotopes ^{2}H, ^{13}C, ^{15}N, or ^{18}O. Mixing of the tagged proteins or peptides then ensures that any subsequent losses of a given peptide will be paralleled by a proportionate loss of both light and heavy versions, thus preserving the ratio between the two. It is thus similar to the DIGE technique, which also involves mixing of two differently labelled samples, and is also similar to the use of internal standards for absolute quantitation discussed in Sect. 4.3.5. Generally, a mass difference of at least 4 Da between 'light' and 'heavy' peptides is preferred to clearly separate the two forms of the peptide, which are often doubly or triply charged. Owing to the chromatographic differences which may occur when deuterium labelling is used, carbon-13 is a preferred heavy isotope, although deuterium is often used in proof-of-concept experiments as it is less expensive. Prior to final separation of the peptides and measurement of the ratios of light and heavy versions of peptides via MS, there exists the opportunity, if appropriate tags have been employed, to selectively extract the tagged peptides in order to reduce the complexity of the mixture of peptides to be analysed.

Introduction of light and heavy isotopic tags may be effected by a number of reagents and techniques, and it is interesting to see that protein and peptide modification chemistry from the 1960s and 1970s is finding new applications in modern proteomics. It should be noted that this classical chemistry, which was originally described for the modification of purified proteins, may perform well in proof-of-concept experiments with standard proteins and relatively simple mixtures. However, this does not mean it will always have the absolute specificity desired when working with very complex mixtures of proteins of widely differing abundance. This is particularly true when conditions are employed to drive reactions to completion. One possible exception is the 2-methoxy-4,5-dihydro-1*H*-imidazole reagent described by Peters et al. (2001), a reagent which is more specific than *O*-methylisourea for the ε-amino group of lysine residues, even though a small amount of reaction with an α-amine can occur – especially in the case of glycine (Peters, private communication). The failure to use many of the classical chemistries is probably due to a combination of factors. For example, labelled reagents are not always commercially available or they are difficult to synthesise, reagents

may not be specific, appropriate software enabling automated analysis of hundreds to thousands of peptides may not be available, fears of patent issues may exist, and there may be disbelief that a successful poof-of-concept experiment could be made to work in a 'real' experimental case. This has led most work in proteomics to be concentrated on ICAT and derivatives such as iTRAQ (both which are commercially available) and the enzymic incorporation of ^{18}O. The latter case incorporates ^{18}O during protein digestion, and is discussed separately in Sect. 4.3.3. Isotopic tagging has been reviewed recently by Leitner and Lindner (2004). The breadth of this review is reflected in Table 4.1, which lists the tagging reagents used for quantitation by MS. Whilst not comprehensive, it serves as an overview of what is possible by incorporating chemical tags.

Table 4.1 Reagents used for relative quantitation in mass spectrometry based proteomics. Others include methyl and ethyl esterification; alkylation with methyl iodide, *N*-methylmaleimide and *N*-ethylmaleimide, *N*-ethyliodoacetamide and *N*-butyliodoacetamide; *S*-methyl thioacetimidate (and propionimidate); labelled *O*-methyl isourea, various halogenated reagents. As reviewed by Leitner and Lindner (2004)

Reagent	Reaction	Labels incorporated[a]	Comments	References
Acetyl *N*-hydroxy succinimide ester	Acylation of amino	H_3/D_3	Not completely specific and quantitative at the same time	–
Acetic anhydride	Acylation of amino	H_3/D_3	As above; immunoaffinity used	Warren et al. (2004)
Propionic anhydride	Acylation of amino	H_{10}/D_{10}	Used to acylate N-terminal residues after digestion	Zappacosta and Annan (2004)
Propionic *N*-hydroxy succinimide ester	Acylation of amino	$^{12}C_3/^{13}C_3$	Not completely specific and quantitative at the same time	Zhang et al. (2005)
Succinic anhydride	Acylation of amino	H_4/D_4	Converts amino to carboxyl	
Nicotinyl *N*-hydroxy-succinimide ester	Acylation of amino	H_4/D_4 or $^{12}C_6/^{13}C_6$	Isotope-coded protein label; charge helps MS/MS	Schmidt et al. (2005)
Phenyl-isothio-cyanate	Phenylthio-carbamoyl-ation of amino	H_5/D_5	Not completely specific for N-termini	–
2-Methoxy-4,5-dihydro-1*H*-imidazole	Guanidinyl-ation of Lys		Commercialised by Agilent as 4H reagent: directs MS/MS fragmentation	Peters et al. (2001)
ICAT	Alkylation of SH	H_8/D_8	Decomplexification; reagents available from ABI; not ideal	Gygi et al. (1999)

(continued)

Table 4.1 (continued)

Reagent	Reaction	Labels incorporated[a]	Comments	References
ICAT on solid phase	Alkylation of SH	H_8/D_8	Photochemical cleavage is delicate	Zhou et al. (2002)
Cleavable ICAT	Alkylation of SH	$^{12}C_n/^{13}C_n$	Decomplexification; reagent available from ABI; costly	Molloy et al. (2005)
iTRAQ		$^{12}C_n/^{13}C_n$	Isobaric, differences seen by MS/MS, sold by ABI	Choe et al. (2005)
N-t-Butyliodoacetamide and iodoacetanilide	Alkylation of SH	H_5/D_5 and H_9/D_9	Quantitative alkylation of cysteine residues	Pasquarello et al. (2004)
2-Vinylpyridine	Alkylation of SH	H_4/D_4	Cleaner and more quantitative alkylation	–
Acrylamide	Addition to SH	H_3/D_3	Quantitative alkylation of cysteine residues	Sechi (2002)
Formaldehyde	Reductive alkylation at amino	Two formaldehydes: H_4/D_4	Dimethyl derivatives formed from HCHO and DCDO	Hsu et al. (2003)
Nitrobenzylsulfenyl chloride	Alkylation of Trp	$^{12}C_6/^{13}C_6$		Kuyama et al. (2003)
Acryloyl quaternary amine compound	Alkylation of SH	None so far	Cation-exchange enrichment; APTA/QAT	Ren et al. (2004)
His tag, *S*-pyridyl	SS formation	H_4/D_4	Hys-Tag; enrich the alkylated Cys peptides	Olsen et al. (2004)
ECAT	Alkylation of SH	Metal ions in chelate	Element coded affinity tag , can multiplex, immunoenrichment	Whetstone et al. (2004)
Iodoacetyl VICAT	Substitution by SH	^{13}C, ^{15}N, ^{14}C tracer	Visible ICAT, very complex: ^{14}C, biotin, photolysis	Lu et al. (2004)
Covalent capture	Cys SS on solid phase	^{18}O in digestion		Liu et al. (2004)
Covalent capture	Cys alk on solid using iodoacetyl	$^{13}C_9$		Shi et al. (2004)
Covalent capture	Cys alk on solid using maleimide	H_{10}/D_{10}	Acid-labile isotope-coded extractant	Qiu et al. (2002)

D represents deuterium.
MS/MS tandem mass spectrometry.
[a] The numbers after the atoms refer to the number of times an atom is found in a labelling reagent.

Even though a great deal of ingenuity has been employed in the use of stable isotopic labelling, there remains a need for more specific, accurate, and easily applied techniques. Some of the reagents described react mainly with amino groups, some react fairly specifically with thiol residues. Some of the reagents are equipped with an affinity or other group which assists the selective extraction of labelled peptides and reduces the complexity of the mixture for final MS analysis. This extraction may be based on charge, hydrophobicity, affinity, or immunoaffinity. Proteomic data are very complex even without any labelling. When sophisticated separation techniques are employed, advanced bioinformatics techniques are needed to sift through the data. These are currently under development. For an appreciation of this data-sifting task, see a recent report that describes automated methods for the analysis of MS measurements of isotopically labelled samples (Zhang et al. 2005).

4.3.3 Enzyme-Mediated Incorporation of Stable Isotopes

Digestion of proteins in the presence of ^{18}O can isotopically label the resulting peptides. The most common approach involves incorporation of ^{18}O during proteolysis with trypsin, but other endoproteinases, such as Glu-C and Lys-C, which also form a covalent acyl-enzyme intermediate may be used. The enzyme peptide-*N*-glycosidase F (PNGaseF), used in analyses of N-glycosylation, also results in isotopic labelling through incorporation of one oxygen atom during the cleavage reaction. The mechanism of trypsin hydrolysis is shown in Fig. 4.5. After initial acyl-enzyme hydrolysis, which leads to incorporation of the label from water in the medium, the peptides

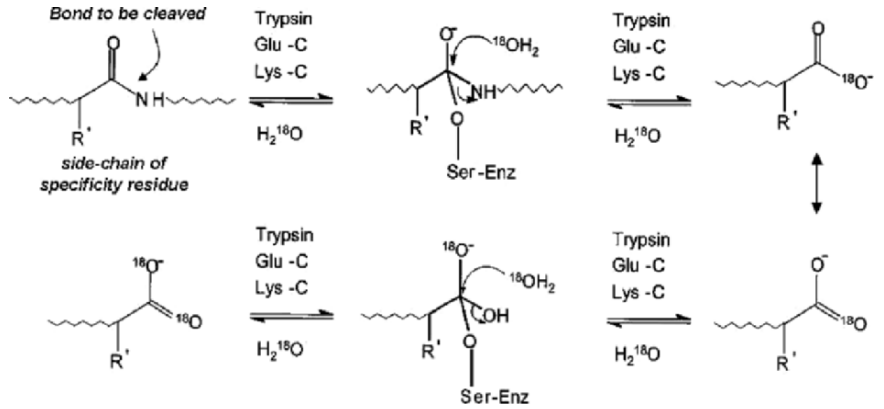


Fig. 4.5 Incorporation of oxygen during hydrolysis when an acyl-enzyme intermediate is formed. Rebinding of cleaved substrate to the enzyme followed by rehydrolysis leads to equilibration of oxygen isotopes with the bulk medium. (Adapted from Yao et al. 2003)

possessing the C-terminal specificity residue (e.g. lysine or arginine in the case of trypsin) may rebind to the enzyme. This reforms an acyl-enzyme complex, which can be rehydrolysed. Multiple rounds of rebinding and hydrolysis replace the unlabelled peptides with ^{18}O until equilibration with the bulk medium has occurred.

Incorporation of labelled oxygen during proteolysis has been exploited for decades to identify C-terminal peptides (Rose et al. 1983). While the technique has undergone improvements, including predigestion in unlabelled water followed by redigestion in labelled water for the incorporation step (Yao et al. 2001, 2003), the finer features determining incorporation are still not completely understood. For example, Hicks et al. (2005) showed that incorporation efficiency can be below 50% for some peptides. Nevertheless, a comparison of ^{18}O incorporation with ICAT labelling using an ion-trap mass spectrometer (Sakai et al. 2005) found ^{18}O to be superior.

As is evident from the discussion of the mechanism above, an endoproteinase that cleaves on the N-terminal side of its specific residue (e.g. Asp N) normally incorporates only one oxygen atom into the liberated carboxyl group (Rao et al. 2005). This is because rebinding of the labelled carboxyl group is difficult as the specific cleavage residue is no longer present (e.g. Asp). A high pH is also used for digestion (8.5) which favours the carboxylate over the carboxylic acid form. Accordingly, some extra incorporation can be observed at a lower pH of 6.

4.3.4 Biological Incorporation of Stable Isotopes by Metabolic Labelling

When cultured cells or small organisms are used, stable isotopes can be incorporated by using media containing ^{13}C-glucose, $^{15}NH_3$, or ^{13}C-labelled amino acids (Gu et al. 2004). This approach has been named SILAC, an abbreviation for stable isotope labelling by amino acids in cell culture (Blagoev et al. 2003). The SILAC method allows essential 'heavy' nutrients in the cell culture media to be incorporated into growing cells as new proteins are synthesised. Protein extracts from labelled and unlabelled cells are mixed before coanalysis by MS. Snijders et al. (2005) used $^{15}NH_3$, ^{13}C labelling, or an unlabelled form, giving three versions of each peptide. This permitted relative quantitation to be calculated, and the atomic compositions of carbon and nitrogen to be estimated for each peptide. As mentioned in Sect. 4.2, metabolic labelling has been compared with the DIGE technique (Kolkman et al. 2005).

With SILAC, no chemical labelling or affinity purification steps are required. This contrasts with the ICAT procedure. The SILAC method is compatible with many cell culture conditions, including primary cells. Ong et al. (2002) have shown that incorporation is complete, and that cells remain normal in the presence of labelled media. In their report they applied SILAC to the study of mouse C2C12 cells and followed the differentiation from

myoblasts into myotubes. This process of muscle differentiation involves dramatic changes in the expression levels of proteins as the cells differentiate from one cell type to another. Several proteins were found to be upregulated during this process.

4.3.5 Relative Quantitation Without Use of Stable Isotope Labelling

A relatively new and alternative computational method for detecting differential protein abundance by LC-MS without the need for isotopic labelling was reported by Wang et al. (2003). They found that the intensity of ion signal versus the molecular concentration was linear, and the correlation increased by reproducing LC-MS runs. Thus, it can be used to track quantitative differences between multiple samples, without the need to incorporate isotopic or other labels (as mentioned earlier). The approach has since rapidly been adopted by various groups (Nunn et al. 2006; Prakash et al. 2006; Radulovic et al. 2004), but has yet to become mainstream. As a result, limitations of the approach are not yet well known. There are multiple advantages for this approach compared with the existing isotopic labelling approaches. The most obvious advantage is that one can move to larger experimental designs as is the case in clinical studies. By comparison, the use of differential isotopic labelling (ICAT, SILAC, iTRAQ) limits experiments to pairwise or at best four-way comparisons. Other disadvantages related to labelling approaches are the increase of sample complexity and subsequent decrease of protein coverage. Every label at least doubles the number of peptides, or MS/MS events that must occur. Labelling experiments are also quite expensive.

4.3.6 Absolute Quantitation by Mass Spectrometry

Absolute quantitation can also be used in proteomics. First reported by De Leenheer and Thienpont (1992) and more recently by Gerber (2003), the various methodologies have recently been reviewed by Lill (2003). As a general strategy, a peptide of interest is synthesised to contain a certain isotope. It is possible to deplete an isotope (usually ^{13}C) rather than enrich it (Stocklin et al. 2000). The advantages of this include no difference in chromatographic behaviour, increased sensitivity due to a narrower isotopic distribution, easy recognition of a depleted peptide owing to its unnatural isotope distribution, and a relatively low cost as depleted material is usually cheaper than enriched material (Nelson et al. 2004). Subsequently, a known quantity of the labelled peptide of interest is spiked into a sample, and the sample is analysed by LC-MS/MS. The precise quantity of a peptide present in a sample can be calculated by comparing the quantity of labelled peptide with that from its corresponding, unlabelled partner.

4.4 Analysis of Known Post-translational Modifications

Quantitation by labelling one or more specific residues on peptides is an efficient means of whole proteome or subproteome analysis. In such analyses, however, only a small number of peptides are usually seen for each protein. Whilst this is appropriate for protein identification and quantitation, this is not useful for the analysis and quantitation of protein post-translational modifications. However, strategies that selectively target and enrich for post-translationally modified peptides can be used, borrowing from the same concept for selection of peptides for proteome analysis. We highlight three areas where large-scale quantitation has been attempted; glycosylation, phosphorylation, and ubiquitinylation analysis. While these methods are yet to be widely applied, they have the potential to be used on a large scale for semiquantitative analyses of protein post-translational modifications.

4.4.1 Glycosylation

The glycosylation of proteins enables them to fulfil important functions. Glycosylation structures can vary enormously, as can the biological consequences of glycosylation (see Chap. 5). For example, glycosylation affects the correct folding of a protein, and changes in glycosylation are also known to be associated with malignant cell transformation and tumorigenesis. Two known forms of glycosylation exist in humans: N-linked and O-linked. N-linked glycosylation is found on asparagine residues, usually at an N-x-S/T motif (where x is any amino acid except proline) (Bause 1983). N-linked glycosylation is present on proteins secreted into body fluids such as plasma, cerebrospinal fluid, urine, and others, and is found on extracellular domains of membrane proteins. O-linked glycosylation occurs on the side chains of serine or threonine residues.

The quest to measure protein glycosylation in a quantitative manner has gained momentum owing to the combination of advances in the analytical power of MS and selective enrichment procedures. Two new methods for the identification of N-linked glycopeptides will be discussed (Kaji et al. 2003; Zhang et al. 2003). The work of Kaji et al. (2003) describes MS analysis of protein glycosylation using a combination of lectin-affinity purification followed by isotope-coded tagging. The second approach by Zhang et al. (2003) describes a method to directly identify and quantify N-linked glycoproteins, and applies the method for the enrichment of cell-surface proteins and improved profiling of serum proteins by removing albumin from serum samples. There are similarities and differences in these methods. Both are similar in that they first generate peptides and capture the glycopeptides from the peptide mixture onto a solid support and release the N-linked glycans by

PNGaseF. MS is used to identify the N-linked glycosylation sites on the peptides and quantify the relative abundance of glycopeptides using isotope-coded tags. Their differences lie in the mechanisms by which the glycopeptides are captured and the isotopic labelling occurs.

In the lectin-affinity approach, glycoproteins are purified from complex mixtures using lectin-affinity chromatography (Kaji et al. 2003). The glycoprotein-enriched fraction is digested with trypsin, and glycopeptides are captured a second time using the same lectin column. The N-linked glycopeptides are then cleaved with PNGaseF in ^{18}O water, which yields a glycosylation site-specific tag. Quantitation then occurs in the same manner as described earlier for ^{18}O labelling. In the approach described by Zhang et al. (2003), glycoproteins are first selectively, covalently conjugated to a solid support via hydrazide chemistry. Periodate oxidation converts the *cis*-diol groups of carbohydrates to aldehydes, which then form covalent hydrazone bonds with hydrazide groups immobilised on the solid support. Non-glycosylated proteins are not bound and are washed away. Immobilised glycoproteins are then digested with trypsin, and non-glycosylated peptides are removed by washing, leaving only glycosylated peptides attached to the solid support. In a final chemistry step, the α-amino groups of the immobilised glycopeptides are labelled with isotopically light or heavy forms of succinic anhydride (d_0/d_4, deuterium) after the ε-amino groups of lysine have been converted to homoarginine. The glycosylated peptides are then released from the support using PNGaseF. Besides releasing the glycosylated peptides, PNGaseF treatment also results in the conversion of glycosylated asparagines to aspartic acid, generating a 1-Da mass shift at the formerly glycosylated asparagine residues. With a high-resolution mass spectrometer, a single analysis can identify N-glycosylated proteins, the site(s) of N-glycosylation, and the relative quantity of the identified glycopeptides if stable isotope labelling is employed. The methods have been used for the analysis of serum proteome profiling and serum cell-surface analysis. It is worth pointing out that albumin, the most abundant serum protein, does not contain any N-linked glycosylation sites and is therefore transparent to the method. A unique feature of the approach of Zhang et al., which has yet to be performed in a large-scale study, is the ability to conjugate and release N-linked and O-linked glycoproteins. Because PNGaseF only releases the N-glycosylated peptides, O-linked peptide release must utilise a different strategy. A panel of exoglycosidases must be employed to sequentially remove monosaccharides until only the Galβ1,3GalNAc core remains attached to the serine or threonine residue. The core can then be released by *O*-glycosidase. More details regarding this approach have been discussed by Zhang et al. (2004). A complication exists for peptides or proteins possessing an N-terminal serine or threonine residue, since these are also converted to aldehyde (glyoxylyl) groups by periodate oxidation and may also form hydrazone bonds.

4.4.2 Phosphorylation

Receptor-mediated intracellular signalling regulates most aspects of cellular homeostasis, including cell proliferation, differentiation, and apoptosis. It is achieved by reversible protein phosphorylation (Cohen 2000; Hunter 2000; Pawson 1995; Pawson and Scott 1997), which is widespread. It is estimated that up to one third of mammalian proteins are phosphorylated, many of these subject to regulation by multisite phosphorylation (Cohen 2002; Mann et al. 2002). The most commonly phosphorylated amino acids are serine, threonine, and tyrosine. The phosphoramidates of arginine, histidine, and lysine also occur, as do acyl derivatives of aspartic acid and glutamic acid, although they are less abundant. Some of these modifications are not typically observed unless specific precautions are taken to prevent their loss during protein isolation and analysis (Krishna and Wold 1993).

For the analysis of protein phosphorylation, three general strategies have emerged. The first is MS-based where the detection and identification of phosphorylated peptides in simple mixtures is achieved by employing various ionisation and scanning procedures in particular types of mass spectrometers (Corthals et al. 2005). The other two are based on selective enrichment of phosphoproteins or phosphopeptides, followed by MS. With all these methods, isolation of the phosphoprotein or phosphopeptide is critical to their success. This is because many phosphorylated proteins are present in extremely small amounts and often with low stoichiometry of phosphorylation. Furthermore, multiple differentially phosphorylated protein isoforms may exist, which further complicates site-specific analyses. Thus, we are presented with a formidable challenge, which is to isolate quantities of phosphorylated proteins or peptides in sufficient quantities to permit identification and quantitation.

Enrichment of phosphorylated proteins can be achieved through the use of immobilised metal affinity chromatography (IMAC) (Andersson and Porath 1986; Posewitz and Tempst 1999) or through the use of antibodies (Gronborg et al. 2002). Antibodies have successfully been employed in a wide range of proteomic studies, and are specific for phosphorylated amino acids or for specific amino acid sequences containing phosphorylated amino acids. There are a myriad of examples, although the full potential of their employment has not yet been explored. A case in point is the strategic application of phosphotyrosine immunoprecipitation in tandem with methyl esterification and IMAC enrichment of tryptic phosphopeptides followed by sequence analysis by MS (Salomon et al. 2003). In a subsequent study from another group, methyl-esterification was found not to decrease the actual binding of unphosphorylated peptides to the IMAC resin; rather, it increased the specificity of elution of phosphopeptides with phosphate buffer (Haydon et al. 2003). Nevertheless, there remains a need for generalised methods that target selective isolation of all phosphorylated proteins or phosphorylated peptides. Greater than 90% of protein phosphorylation occurs at serine or threonine residues

yet immunoaffinity purification of phosphoserine or phosphothreonine proteins has been less successful than for phosphotyrosine owing to the lack of high-affinity, broadly reactive phosphoserine-specific or phosphothreonine-specific antibodies.

Quantitative phosphoprotein measurements have traditionally been performed by measuring protein phosphorylation after metabolic radiolabelling with inorganic [^{32}P]phosphate in combination with 2-DE or SDS polyacrylamide gel elecrophoresis, or two-dimensional phosphopeptide mapping (Gallis et al. 1999). Protein identification and discovery of phosphorylated amino acids is accomplished by MS (Becker et al. 1998; Gallis et al. 1999). 2-DE can also be used to detect quantitative differences in phosphorylation if suitable antibodies are available. Following western blotting and realignment of the gel with the detected phosphoprotein, subsequent site-specific identification is then achieved by MS (Tremolada et al. 2005). The speed and specificity of these methods are low and new methods have been developed to replace these approaches.

One approach is to directly couple quantitation with peptide identification, by isotope labelling two samples with differentially coded tags so that the samples can be mixed and analysed simultaneously. This approach gives increased speed of analysis and is analogous to the methods described earlier for protein identification. Each phosphopeptide then appears as two peaks in a mass spectrum, and the relative abundances of the peaks reflect the amount of the phosphopeptide in each sample. Labelling can be achieved by metabolic labelling of proteins, or chemical labelling of functional groups such as peptide N-termini or C-termini (Bonenfant et al. 2003; Liu and Regnier 2002; Yao et al. 2001), or various other methods reported in Table 4.1.

Several new procedures deserve particular mention as they were specifically designed for quantitative phosphoprotein analysis. These procedures were made feasible through chemical derivatisation of phosphoamino acids involving phosphoamidate chemistry (Zhou et al. 2001), or through β-elimination and Michael addition reaction of the phosphorylation site where the phosphorylated residue is chemically converted to a covalent affinity tag (Goshe et al. 2002; Oda et al. 2001; Weckwerth et al. 2000). While the direct mapping of serine, threonine, and tyrosine phosphorylation sites is attractive (discussed above), the procedures are relatively complicated and the yields are low. The β-elimination and Michael addition reaction have two additional limitations. First, they are limited to serine and threonine residues and cannot analyse other phosphoamino acids. Second, intracellular serine and threonine may also be dynamically glycosylated through an O-linked *N*-acetylglucosamine (Wells et al. 2001). An excellent extension to these methods was developed by Knight and colleagues. Here, β-elimination and Michael addition reactions were also used to create lysine analogues (aminoethylcysteine and *β*-methylaminoethylcysteine, respectively) that can be cleaved with a lysine-specific protease to map sites of phosphorylation (Knight et al. 2003; Rusnak et al. 2002). Selective capture and modification of phosphopeptides,

using cysteamine on a solid support, could facilitate phosphopeptide enrichment and identification in a single step. This could even be combined with ^{18}O labelling for quantitative analysis.

4.4.3 Ubiquitinylation

Ubiquitin, a small protein of 76 amino acids, covalently attaches to proteins as monomers or lysine-linked chains. Proteins can become dynamically modified with ubiquitin by various enzymes, producing a number of different outcomes. Ubiquitin modifications are typically associated with protein degradation through targeting to proteasomes. However, ubiquitin has other cellular roles not associated with proteasomal degradation and it is also evident that the type of ubiquitin linkage to a protein may influence its fate. The defective regulation of ubiquitin has now been found in diseases ranging from developmental abnormalities and autoimmunity, to neurodegenerative diseases and cancer (Layfield et al. 2001; Weissman 2001).

During the ubiquitinylation process, ubiquitin is covalently conjugated to the ε-amino group of lysine residues of target proteins. Tryptic digestion of ubiquitinylated proteins produces a signature peptide at the ubiquitinylation site containing diglycine remnants, a mass shift at the lysine residue of 114.1 Da, and missed proteolytic cleavage because trypsin proteolysis cannot occur at modified lysines. Quantitative analysis of ubiquitinylation is possible through the use of ICAT (Gygi et al. 1999), however this will miss many ubiquitin sites as only cysteine-containing residues are targeted. Alternatively, stable metabolic labelling that incorporates isotopes into living cells before harvesting could be of use (Blagoev et al. 2003; Ong et al. 2002). The obvious benefit of metabolic labelling is the quantification of non-cysteine-containing peptides, which would include most diglycine-containing signature peptides.

4.5 Conclusions

For quantitative protein analysis there are a bewildering number of procedures that one could follow. However, before considering methods for quantitative proteomics it is important to note that MS is integrated into a workflow that incorporates various sciences, including biochemical and chromatographic procedures, and heavily relies on sophisticated instrument control procedures and data-analysis software. Current MS instruments may cycle through a number of events, including automatic measurement of ratios and selection of specific ions from a mixture of ions, fragment selected ions, and record the precise masses and ratios of the resulting fragment ions. Thus, a desired analytical strategy must be suitable for the analytical capacity

that is available. A limiting step in the analytical capabilities of a project might well define the eventual strategy employed.

Any analytical strategy must be capable of delivering results that relate to the goals of a scientific project. As mentioned at the outset of this chapter, quantitative analysis covers three important areas. For diagnostic purposes, one needs to define the quantity and composition of a cell or biofluid. For classification of disease states and drug efficacy, similar quantitative and compositional analyses are needed. Finally, relative quantitative measurements provide a means for subtractive analysis where one can target a particular cell mechanism, which in turn provides information about the function of those proteins that are upregulated or downregulated. Careful selection of a quantitative proteomics strategy must always consider these issues.

Acknowledgements. We thank the University of Geneva and the Swiss National Science Foundation for financial support.

References

Andersson L, Porath J (1986) Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. Anal Biochem 1:250–254

Barry R, Soloviev M (2004) Quantitative protein profiling using antibody arrays. Proteomics 4:3717–3726

Bause E (1983) Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. Biochem J 209:331–336

Becker S, Corthals GL, Aebersold R, Groner B, Muller CW (1998) Expression of a tyrosine phosphorylated DNA binding Stat3beta dimer in bacteria. FEBS Lett 441:141–147

Blagoev B, Kratchmarova I, Ong SE, Nielsen M, Foster LJ, Mann M (2003) A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. Nat Biotechnol 21:315–318

Bonenfant D, Schmelzle T, Jacinto E, Crespo JL, Mini T, Hall MN, Jenoe P (2003) Quantitation of changes in protein phosphorylation: a simple method based on stable isotope labeling and mass spectrometry. Proc Natl Acad Sci USA 100:880–885

Cagney G, Emili A (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. Nat Biotechnol 20:163–170

Chakraborty A, Regnier FE (2002) Global internal standard technology for comparative proteomics. J. Chromatogr A 949:173–184

Choe LH, Aggarwal K, Franck Z, Lee KH (2005) A comparison of the consistency of proteome quantitation using two-dimensional electrophoresis and shotgun isobaric tagging in *Escherichia coli* cells. Electrophoresis 26:2437–2449

Cohen P (2000) The regulation of protein function by multisite phosphorylation – a 25 year update. Trends Biochem Sci 25:596–601

Cohen P (2002) Protein kinases – the major drug targets of the twenty-first century? Nat Rev Drug Discov 1:309–315

Conrads TP, Alving K, Veenstra TD, Belov ME, Anderson GA, Anderson DJ, Lipton MS, Pasa-Tolic L, Udseth HR, Chrisler WB, Thrall BD, Smith RD (2001) Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling: Anal Chem 73:2132–2139

Corthals GL, Aebersold R, Goodlett DR (2005) Identification of phosphorylation sites using microimmobilized metal affinity chromatography. Methods Enzymol 405:66–81

De Leenheer AP, Thienpont LM (1992) Application of isotope dilution-mass spectrometry in clinical chemistry pharmacokinetics and toxicology. Mass Spectrom Rev 11:249–307

Gallis B, Corthals GL, Goodlett DR, Ueba H, Kim F, Presnell SR, Figeys D, Harrison DG, Berk BC, Aebersold R, Corson MA (1999) Identification of flow-dependent endothelial nitric-oxide synthase phosphorylation sites by mass spectrometry and regulation of phosphorylation and nitric oxide production by the phosphatidylinositol 3-kinase inhibitor LY294002. J Biol Chem 274:30101–30108

Gehanne S, Cecconi D, Carboni L, Righetti PG, Domenici E, Hamdan M (2002) Quantitative analysis of two-dimensional gel-separated proteins using isotopically marked alkylating agents and matrix-assisted laser desorption/ionization mass spectrometry. Rapid Commun Mass Spectrom 16:1692–1698

Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci USA 100:6940–6945

Goodlett DR, Keller A, Watts JD, Newitt R, Yi EC, Purvine S, Eng JK, von Haller P, Aebersold R, Kolker E (2001) Differential stable isotope labeling of peptides for quantitation and *de novo* sequence derivation. Rapid Commun Mass Spectrom 15:1214–1221

Gorg A, Weiss W, Dunn MJ (2004) Current two-dimensional electrophoresis technology for proteomics. Proteomics 4:3665–3685

Goshe MB, Veenstra TD, Panisko EA, Conrads TP, Angell NH, Smith RD (2002) Phosphoprotein isotope-coded affinity tags application to the enrichment and identification of low-abundance phosphoproteins. Anal Chem 74:607–616

Greengauz-Roberts O, Stoppler H, Nomura S, Yamaguchi H, Goldenring JR, Podolsky RH, Lee JR, Dynan WS (2005) Saturation labeling with cysteine-reactive cyanine fluorescent dyes provides increased sensitivity for protein expression profiling of laser-microdissected clinical specimens. Proteomics 5:1746–1757

Gronborg M, Kristiansen TZ, Stensballe A, Andersen JS, Ohara O, Mann M, Jensen ON, Pandey A (2002) A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies identification of a novel protein, Frigg, as a protein kinase A substrate. Mol Cell Proteomics 1:517–527

Gu S, Du Y, Chen J, Liu Z, Bradbury EM, Hu CA, Chen X (2004) Large-scale quantitative proteomic study of PUMA-induced apoptosis using two-dimensional liquid chromatography-mass spectrometry coupled with amino acid-coded mass tagging. J Proteome Res 3:1191–1200

Guo L, Eisenman JR, Mahimkar RM, Peschon JJ, Paxton RJ, Black RA, Johnson RS (2002) A proteomic approach for the identification of cell-surface proteins shed by metalloproteases. Mol Cell Proteomics 1:30–36

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17:994–999

Hamdan M, Righetti PG (2002) Modern strategies for protein quantification in proteome analysis: advantages and limitations. Mass Spectrom Rev 4:287–302

Haydon CE, Eyers PA, Aveline-Wolf LD, Resing KA, Maller JL, Ahn NG (2003) Identification of novel phosphorylation sites on *Xenopus laevis* aurora A and analysis of phosphopeptide enrichment by immobilized metal-affinity chromatography. Mol Cell Proteomics 2:1055–1067

Heller M, Mattou H, Menzel C, Yao X (2003) Trypsin catalyzed ^{16}O to ^{18}O exchange for comparative proteomics tandem mass spectrometry comparison using MALDI-TOF ESI-QTOF and ESI-ion trap mass spectrometers. J Am Soc Mass Spectrom 14:704–718

Hicks WA, Halligan BD, Slyper RY, Twigger SN, Greene AS, Olivier M (2005) Simultaneous quantification and identification using ^{18}O labeling with an ion trap mass spectrometer and the analysis software application "ZoomQuant". J Am Soc Mass Spectrom 16:916–925

Hsu JL, Huang SY, Chow NH, Chen SH (2003) Stable-isotope dimethyl labeling for quantitative proteomics. Anal Chem 75:6843–6852

Hunt SM, Thomas MR, Sebastian LT, Pedersen SK, Harcourt RL, Sloane AJ, Wilkins MR (2005) Optimal replication and the importance of experimental design for gel-based quantitative proteomics. J Proteome Res 4:809–819

Hunter T (2000) Signaling – 2000 and beyond. Cell 100:113–127

Kaji H, Saito H, Yamauchi Y, Shinkawa T, Taoka M, Hirabayashi J, Kasai K, Takahashi N, Isobe T (2003) Lectin affinity capture isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. Nat Biotechnol 21:667–672

Knight ZA, Schilling B, Row RH, Kenski DM, Gibson BW, Shokat KM (2003) Phosphospecific proteolysis for mapping sites of protein phosphorylation. Nat Biotechnol 21:1047–1054

Kodadek T (2001) Protein microarrays prospects and problems. Chem Biol 8:105–115

Kodadek T (2002) Development of protein-detecting microarrays and related devices Trends Biochem Sci 27:295–300

Kolkman A, Dirksen EH, Slijper M, Heck AJ (2005) Double standards in quantitative proteomics direct comparative assessment of difference in gel electrophoresis and metabolic stable isotope labeling. Mol Cell Proteomics 4:255–266

Krishna RG, Wold F (1993) Post-translational modification of proteins. Adv Enzymol Relat Areas Mol Biol 67:265–298

Kuyama H, Watanabe M, Toda C, Ando E, Tanaka K, Nishimura O (2003) An approach to quantitative proteome analysis by labeling tryptophan residues. Rapid Commun Mass Spectrom 17:1642–1650

Layfield R, Alban A, Mayer RJ, Lowe J (2001) The ubiquitin protein catabolic disorders. Neuropathol Appl Neurobiol 27:171–179

Leitner A, Lindner W (2004) Current chemical tagging strategies for proteome analysis by mass spectrometry. J Chromatogr B 813:1–26

Lill J (2003) Proteomic tools for quantitation by mass spectrometry. Mass Spectrom Rev 22:182–194

Liu P, Regnier FE (2002) An isotope coding strategy for proteomics involving both amine and carboxyl group labeling. J Proteome Res 1:443–450

Liu T, Qian WJ, Strittmatter EF, Cam DG 2nd, Anderson GA, Thrall BD, Smith RD (2004) High-throughput comparative proteome analysis using a quantitative cysteinyl-peptide enrichment technology. Anal Chem 76:5345–5353

Lu Y, Bottari P, Turecek F, Aebersold R, Gelb MH (2004) Absolute quantification of specific proteins in complex mixtures using visible isotope-coded affinity tags. Anal Chem 76:4104–4011

MacBeath G (2002) Protein microarrays and proteomics. Nat Genet 32(Suppl): 526–532

Mann M, Ong SE, Gronborg M, Steen H, Jensen ON, Pandey A (2002) Analysis of protein phosphorylation using mass spectrometry deciphering the phosphoproteome. Trends Biotechnol. 20:261–268.

Marouga R, David S, Hawkins E (2005) The development of the DIGE system: 2D fluorescence difference gel analysis technology. Anal Bioanal Chem 382:669–678

Mirgorodskaya OA, Kozmin YP, TitoMI, Korner R, Sonksen CP, Roepstorff P (2000) Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using ^{18}O-labeled internal standards. Rapid Commun Mass Spectrom 14:1226–1232

Molloy MP, Donohoe S, Brzezinski EE, Kilby GW, Stevenson TI, Baker JD, Goodlett DR, Gage DA (2005) Large-scale evaluation of quantitative reproducibility and proteome coverage using acid cleavable isotope coded affinity tag mass spectrometry for proteomic profiling. Proteomics 5:1204–1208

Munchbach M, Quadroni M, Miotto G, James P (2000) Quantitation and facilitated *de novo* sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. Anal Chem 72:4047–4057

Nelson RW, Nedelko D, Tubbs KA, Kiernan UA (2004) Quantitative mass spectrometric immunoassay of insulin like growth factor. J Proteome Res 3:851–855

Niwayama S, Kurono S, Matsumoto H (2001) Synthesis of d-labeled N-alkylmaleimides and application to quantitative peptide analysis by isotope differential mass spectrometry. Bioorg Med Chem Lett 11:2257–2261

Nunn BL, Shaffer SA, Scherl A, Gallis B, Wu M, Miller SI, Goodlett DR (2006) Comparison of a Salmonella typhimurium proteome defined by shotgun proteomics directly on an LTQ-FT and by proteome pre-fractionation on an LCQ-DUO. Brief Funct Genomic Proteomic 5:154–168

Oda Y, Nagasu T, Chait BT (2001) Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. Nat Biotechnol 19:379–382

Olsen JV, Andersen JR, Nielsen PA, Nielsen ML, Figeys D, Mann M, Wisniewski JR (2004) HysTag – a novel proteomic quantification tool applied to differential display analysis of membrane proteins from distinct areas of mouse brain. Mol Cell Proteomics 3:82–92

Ong SE, BlagoeB, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002) Stable isotope labeling by amino acids in cell culture SILAC as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1:376–386

Pasquarello C, Sanchez JC, Hochstrasser DF, Corthals GL (2004) N-t-Butyliodoacetamide and iodoacetanilide two new cysteine alkylating reagents for relative quantitation of proteins. Rapid Commun Mass Spectrom 18:117–127

Pawson T (1995) Protein modules and signalling networks. Nature 373:573–580

Pawson T, Scott JD (1997) Signaling through scaffold anchoring and adaptor proteins. Science 278:2075–2080

Peters EC, Horn DM, Tully DC, Brock A (2001) A novel multifunctional labeling reagent for enhanced protein characterization with mass spectrometry. Rapid Commun Mass Spectrom 15:2387–2392

Posewitz MC, Tempst P (1999) Immobilized gallium(III) affinity chromatography of phosphopeptides. Anal Chem 71:2883–2392

Prakash A, Mallick P, Whiteaker J, Zhang H, Paulovich A, Flory M, Lee H, Aebersold R, Schwikowski B (2006) Signal maps for mass spectrometry-based comparative proteomics. Mol Cell Proteomics 5:423–432

Qiu Y, Sousa EA, Hewick RM, Wang JH (2002) Acid-labile isotope-coded extractants a class of reagents for quantitative mass spectrometric analysis of complex protein mixtures. Anal Chem 74:4969–4979

Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, Mao Y, Emili A (2004) Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics 3:984–997

Ren D, Julka S, Inerowicz HD, Regnier FE (2004) Enrichment of cysteine-containing peptides from tryptic digests using a quaternary amine tag. Anal Chem 76:4522–4230

Rao KC, Carruth RT, Miyagi M (2005) Proteolytic ^{18}O labeling by peptidyl-Lys metallo-endopeptidase for comparative proteomics. J Proteome Res 4:507–514

Rose K (2005) Industrial-scale proteomics analysis of human plasma. In: Figeys D (ed) Industrial proteomics: applications for biotechnology and pharmaceuticals. Wiley, New York. pp 217–229

Rose K, Simona MG, Offord RE, Prior CP, Otto B, Thatcher DR (1983) A new mass-spectrometric C-terminal sequencing technique finds a similarity between gamma-interferon and alpha 2-interferon and identifies a proteolytically clipped gamma-interferon that retains full antiviral activity. Biochem J 215:273–277

Rusnak F, Zhou J, Hathaway GM (2002) Identification of phosphorylated and glycosylated sites in peptides by chemically targeted proteolysis. J Biomol Tech 13:228–237

Sakai J, Kojima S, Yanagi K, Kanaoka M (2005) ^{18}O-labeling quantitative proteomics using an ion trap mass spectrometer. Proteomics 5:16–23

Salomon AR, Ficarro SB, Brill LM, Brinker A, Phung QT, Ericson C, Sauer K, Brock A, Horn DM, Schultz PG, Peters EC (2003) Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. Proc Natl Acad Sci USA 100:443–448

Schmidt A, Kellermann J, Lottspeich F (2005) A novel strategy for quantitative proteomics using isotope-coded protein labels. Proteomics 5:4–15

Sechi S (2002) A method to identify and simultaneously determine the relative quantities of proteins isolated by gel electrophoresis. Rapid Commun Mass Spectrom 16:1416–1424
Sechi S, Oda Y (2003) Quantitative proteomics using mass spectrometry. Curr Opin Chem Biol 7:70–77
Shaw J, Rowlinson R, Nickson J, Stone T, Sweet A, Williams K, Tonge R (2003) Evaluation of saturation labelling two-dimensional difference gel electrophoresis fluorescent dyes. Proteomics 3:1181–1195
Shi Y, Xiang R, Crawford JK, Colangelo CM, Horvath C, Wilkins JA (2004) A simple solid phase mass tagging approach for quantitative proteomics. J Proteome Res 3:104–111
Snijders AP, de Vos MG, de Koning B, Wright PC (2005) A fast method for quantitative proteomics based on a combination between two-dimensional electrophoresis and ^{15}N-metabolic labelling. Electrophoresis 26:3191–3199
Stadtherr K, Wolf H, Lindner P (2005) An aptamer-based protein biochip. Anal Chem 77:3437–3743
Stocklin R, Arrighi JF, Hoang-Van K, Vu L, Cerini F, Gilles N, Genet R, Markussen J, Offord RE, Rose K (2000) Positive and negative labeling of human proinsulin insulin and C-peptide with stable isotopes: new tools for *in vivo* pharmacokinetic and metabolic studies. Methods Mol Biol 146:293–315
Stoll D, Templin MF, Bachmann J, Joos TO (2005) Protein microarrays applications and future challenges. Curr Opin Drug Discov Dev 8 239–252
Tremolada L, Magni F, Valsecchi C, Sarto C, Mocarelli P, Perego R, Cordani N, Favini P, Galli Kienle M, Sanchez JC, Hochstrasser DF, Corthals GL (2005) Characterization of heat shock protein 27 phosphorylation sites in renal cell carcinoma. Proteomics 5 788–795
Van den Bergh G, Arckens L (2004) Fluorescent two-dimensional difference gel electrophoresis unveils the potential of gel-based proteomics. Curr Opin Biotechnol 15:38–43
Wang S, Regnier FE (2001) Proteomics based on selecting and quantifying cysteine containing peptides by covalent chromatography. J Chromatogr A 924:345–357
Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal Chem 75:4818–4826
Wang YK, Quinn DF, Ma Z, Fu EW (2002) Inverse labeling-mass spectrometry for the rapid identification of differentially expressed protein markers/targets. J Chromatogr B 782:291–306
Warren EN, Elms PJ, Parker CE, Borchers CH (2004) Development of a protein chi a MS-based method for quantitation of protein expression and modification levels using an immunoaffinity approach. Anal Chem 76:4082–4092
Weckwerth W, Willmitzer L, Fiehn O (2000) Comparative quantification and identification of phosphoproteins using stable isotope labeling and liquid chromatography/mass spectrometry. Rapid Commun Mass Spectrom 14:1677–1681
Weissman AM (2001) Themes and variations on ubiquitylation. Nat Rev Mol Cell Biol 2:169–178
Wells L, Vosseller K, Hart GW (2001) Glycosylation of nucleocytoplasmic proteins: signal transduction and O-GlcNAc. Science 291:2376–2378
Whetstone PA, Butlin NG, Corneillie TM, Meares CF (2004) Element-coded affinity tags for peptides and proteins. Bioconjugate Chem 15:3–6
Yao X, Freas A, Ramirez J, Demire PA, Fenselau C (2001) Proteolytic ^{18}O labeling for comparative proteomics model studies with two serotypes of adenovirus. Anal Chem 73:2836–2842
Yao X, Afonso C, Fenselau C (2003) Dissection of proteolytic ^{18}O labeling endoprotease-catalyzed ^{16}O to ^{18}O exchange of truncated peptide substrates. J Proteome Res 2:147–152
Zappacosta F, Annan RS (2004) N-terminal isotope tagging strategy for quantitative proteomics results-driven analysis of protein abundance changes. Anal Chem 76:6618–6627
Zhang H, Li XJ, Martin DB, Aebersold R (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry stable isotope labeling and mass spectrometry. Nat Biotechnol 21:660–666

Zhang H, Yan W, Aebersold R (2004) Chemical probes and tandem mass spectrometry a strategy for the quantitative analysis of proteomes and subproteomes. Curr Opin Chem Biol 8:66–75

Zhang X, Hines W, Adamec J, Asara JM, Naylor S, Regnier FE (2005) An automated method for the analysis of stable isotope labeling data in proteomics. J Am Soc Mass Spectrom 16:1181–1191

Zhou H, Watts JD, Aebersold R (2001) A systematic approach to the analysis of protein phosphorylation. Nat Biotechnol 19:375–378

Zhou H, Ranish JA, Watts JD, Aebersold R (2002) Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. Nat Biotechnol 20:512–515

Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, Smith D, Gerstein M, Reed MA, Snyder M (2000) Analysis of yeast protein kinases using protein chips. Nat Genet 26:283–289

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M, Snyder M (2001) Global analysis of protein activities using proteome chips. Science 293:2101–2105

5 One Gene, Many Proteins

Nicolle H. Packer, Andrew A. Gooley, and Marc R. Wilkins

Abstract

The release of the sequence of the human genome in early 2000 generated enormous excitement with the promise of rapid identification of the gene products responsible for cellular function. What has become apparent is that knowledge of the DNA sequence of the gene that is involved in a particular metabolic process is insufficient to predict its function, expression and activity. The very numbers illustrate this difference – it is estimated that there are 22,000 to 25,000 known genes in the human genome but there are probably greater than 1,000,000 proteins in the human proteome. This discrepancy has extended the focus of proteomics to the analysis and understanding of the modifications that occur to proteins both during and after translation of the gene. As we move further into understanding protein function it is becoming increasingly obvious that many of the changes associated with disease and differentiation are to do with the modifications to the proteins rather than only to do with the regulation of the expression of the gene. The main difficulty which has slowed the understanding of the biological role of these protein modifications has been the perception that the analysis of these alterations is difficult and is best left to the limited number of experts in each field. However, the reality is that the increasing availability of sample preparation, mass spectrometric and bioinformatic tools specifically designed for the analysis of post-translational modifications, is enabling the function of these instruments of biological diversity to be explored.

5.1 Introduction

Any protein in a proteome can be modified by co- or post-translational modifications. There are hundreds of different types of modifications, all of which can influence a protein's charge, hydrophobicity, conformation and/or stability and as a consequence, its function. Hence, the 'one gene–one-polypeptide' paradigm is now outdated as we known that in eukaryotes, prokaryotes and archaea the polypeptide translation of a single gene may be modified to create multiple gene products (Fig. 5.1). This concept has been

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

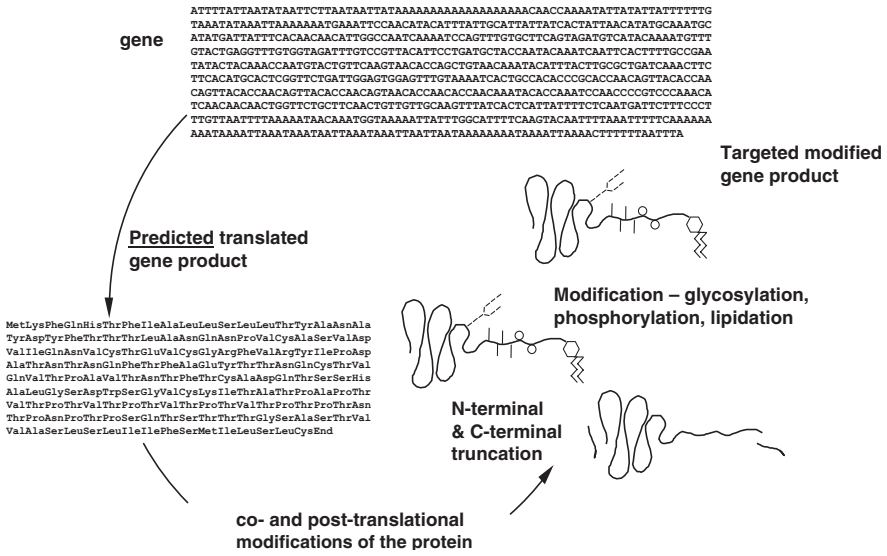


Fig. 5.1 Proteome diversity: from one gene sequence to one predicted protein sequence to many protein products

strongly validated by the sequencing of the human genome, whereby it became clear that there were far fewer genes than originally expected, only 22,000 to 25,000 or so known genes coding for what may be the million or so protein products expressed by humans. It does not take a mathematician to deduce from this that for each gene there are an average of 40 gene products produced by transcriptional, co- and post-translational modifications. With significant improvements in analytical technology, we now understand that proteins can carry many sites and types of modification – referred to as multisite modifications (Yang 2005). A great deal of research is now focused on detailed analysis of proteins so that we can understand how these modifications influence protein structure and function. Many examples are known where modifications play a clear role in a reversible mode – 'on' and 'off' states which either modulate the function of the protein or act as a true molecular switch.

From a definitional perspective, protein post-translational modifications are often considered to be:

1. The modification of an amino acid by the covalent linkage of simple chemicals, such as an acetyl or a phosphoryl group, through to the addition of complex structures such as lipids and carbohydrates.
2. The cleavage of a finished translated transcript to a mature form; for example the processing of signal or activation peptides as well as the autocatalytic protein processing associated with inteins.
3. The cross-linking of amino acids, such as cysteine and tyrosine.

Splice variants of proteins, derived from the alternative processing of messenger RNA, also produce multiple forms of proteins. However, these are not protein co- or post-translational modifications and will not be considered here.

Protein post-translational modifications, whilst ubiquitous, remain elusive. Even though massive amounts of nucleic acid sequences are available, it is still difficult to accurately predict the type of modification(s) that will occur on a protein from its DNA or amino acid sequence alone. Clearly, a protein sequence must influence whether a modification can occur at a specific residue, and there are sequence-specific motifs that serve as signals for certain post-translational modifications (Table 5.1). However, this is modulated by a protein's context, as a protein must be in a certain compartment of the cell to be modified in certain ways and even where a motif exists, its modification is often influenced by the features of the surrounding amino acid sequence. For example, N-linked glycosylation occurs only on proteins which are passing through the Golgi apparatus and is found exclusively in the consensus sequence N-X-S/T. Just as importantly, modifications on proteins should only be considered as 'real' when identified from *in vivo* sourced proteins. This creates a dilemma for the identification of labile modifications such as phosphorylation, and for the analysis of very low abundance proteins. It must be kept in mind that the presence or absence of modifications can have important biological implications and modifications may not be accurately reflected in proteins sourced from *in vitro* or recombinant systems.

Specific amino acid sequences, in a number of cases, have been identified which result in the processing of a polypeptide or the addition of a certain post-translational modification. Secretion signal peptides, mitochondrial and chloroplast target peptides can be localised accurately in protein sequences and all result in the processing of a polypeptide (Emanuelsson et al. 2000). Phosphorylation and glycosylation have also been localised on a sufficiently large number of proteins to allow the use of neural network algorithms to construct modification-specific predictive databases (Gupta et al. 1999; Blom et al. 2004). Sequence motifs for other post-translational modifications are also known, and have been used for large-scale prediction (Lee et al. 2006), but these are currently less accurate. Notwithstanding this, the experimental discovery of post-translational modifications on proteins continues to provide most of our knowledge on the structural and chemical features of the protein sequences which are likely to carry a particular post-translational modification. Table 5.1 lists some of the best studied examples of co- and post-translational modifications that are of known protein sequence specificity.

It is not our intention to undertake a comprehensive review of protein co- and post-translational modifications here as the field has previously been extensively covered elsewhere. There are several excellent reviews which discuss many of the functions of co- and post-translational modifications as well as the methods used for their identification (Jensen 2006; Sect. 5.5). The focus of this chapter will, instead, be more on the emerging

Table 5.1 Sequence motifs for some of the common co- and post-translational modifications. Amino acids in *italics* are those that will be modified in each case

Modification	Sequence	Comment
Phosphorylation	Aaa-Aaa-X-Aaa-*Tyr*	Sites recognised by Tyr kinase
	X-*Ser/Thr*-Neg-Neg-Neg	Sites recognised by casein kinase II
	Lys/Arg-X-X–*Ser/Thr*-X–X-X-Lys/Arg	Sites recognised by protein kinase C[a]
	X-Ser/Thr-Pro-X or Ser/Thr-Pro-X	Sites recognised by p45 casein kinase[b]
Sulfation	Neg-Neg-*Tyr*-Neg-Neg	Tyr–SO_4 is frequently found at clusters of negatively charged amino acids
Glycosylation	*Asn*-X-Ser/Thr/Cys where X≠P	N-linked glycosylation at Asn
N-myristoylation	*Gly*-X-X-X-Ser/Thr	Myristate attached to Gly[c]
Hydroxylation	Gly-X-*Pro*-Gly-X	Hydroxylation of Pro is influenced by X
N-methylation	N-methyl*Met*-X-X-P-X	Methylation in *Escherichia coli*[d]
Carboxymethylation	*Glu-Glu*-X-X-Aaa-Ser/Thr	For carboxymethylation at Glu[e]
Signal peptidase site	Aaa-X-Baa-↓-X	Cleavage of the N-terminal signal peptide[f]
Prenylation – farnesyl, geranyl-geranyl	*Cys*-Aaa-Aaa-X-COOH	Cys in the Cys-Aaa-Aaa-X box. Typical of Ras proteins[g]
SUMO	ψ-Lys-X-Glu (ψ is a hydrophobic amino acid)	Lysine, with clusters of acidic residues located downstream[h]. Note that 23% of sumoylation sites use a different motif

Amino acids are represented with the standard three-letter code. *Aaa* is any aliphatic residue (Ile, Leu, Val and can be Ala), *X* is any amino acid, *Neg* is Asp or Glu, *Baa* is Ala, Gly or Ser.
SUMO small ubiquitin-related modifier.
[a] House et al. (1987).
[b] Angelov (1994).
[c] Johnson et al. (1994).
[d] Apostol et al. (1995).
[e] Terwilliger et al. (1986).
[f] Perlman and Halvorson (1983).
[g] Zhang and Casey (1996).
[h] Yang et al. (2006).

imperative to consider post-translational modifications when looking for mechanisms, markers and medicines. In doing so, we hope to raise the profile of the numerous and varied modifications that occur to proteins during differentiation and disease, and to encourage researchers to actively look for these modifications as a means to obtain a rich and unique understanding of biology.

5.2 An Overview of Modifications: What Are They and Where Do They Occur?

The possible diversity of protein products from any gene is staggering when it is considered that there are hundreds of types of modifications (Jensen 2006), many which can be altered temporally (such as during development and differentiation) and spatially (according to the location within the cell). Unlike the genome, which can be seen as a static information repository, the proteome has a temporal dimension and much of the dynamism associated with this comes from protein post-translational modifications. For the sake of classification, we can spatially compartmentalise the proteome into different organelles. Similarly, we can separate proteins into essentially three different categories: intracellular, cell-surface-associated and extracellular (Table 5.2). By spatially separating the proteome into a number of compartments, we can help define the types of modification that can occur to any particular protein in the proteome. For example, the covalent attachment of asparagine-linked oligosaccharides is only found on secreted proteins and those associated with the endoplasmic reticulum, Golgi apparatus, cell surface and lysosomes. As an extension of this, since many of the modifications are enzyme-mediated, there is a requirement that a protein carrying a certain

Table 5.2 Compartmentalisation of protein modifications by subcellular location

Compartment	Subcompartment	Type of modification
Intracellular	Nucleus	Acetylation (histone acetylation at ε-amino), phosphorylation, SUMO, O-GlcNAc
	Lysosome	Mannose-6-phosphate labelled N-linked sugars
	Mitochondria	N-formyl acylation
	Chloroplast	N-formyl acylation, pigments and light harvesting groups (e.g. chlorophyll)
	Golgi apparatus	N- and O-linked oligosaccharide, sulfation, palmitoylation
	Endoplasmic reticulum	N-linked oligosaccharide, GPI-anchor
	Cytosol	Acetylation, methylation, phosphorylation, O-GlcNAc, SUMO
	Ribosome	Myristoylation
Cell surface	Plasma membrane	N- and O-glycosylation, GPI-anchor
Extracellular	Extracellular fluid	N- and O-glycosylation, acetylation, phosphorylation
	Extracellular matrix	Hydroxylation, phosphorylation, N- and O-glycosylation

modification be derived from (although not necessarily found in) a compartment that contains appropriate modifying enzymes. For example, acetylation of the histones occurs in two compartments: lysine residues are predominantly acetylated in the nucleus, while α-amino groups are acetylated in the cytosol (Lopez-Rodas et al. 1991; Yamada and Bradshaw 1991). In some cases, however, a modification arises in transit through a compartment: for example, most cell-surface proteins which transit through the Golgi apparatus contain one or two N-glycosylation sites and several O-linked glycosylation sites.

5.3 How Do We Find Post-translational Modifications?

5.3.1 Separation of Isoforms

Many proteins that are post-translationally modified exist as a number of isoforms. This is because proteins can exist in modified and unmodified states, they may carry different numbers of one or more types of modifications, and some modifications, such as the N-linked and O-linked carbohydrates, are heterogeneous by nature. This difference in the type and stoichiometry of modifications can lead to the presence of a large number of protein products for any one gene. This presents an enormous challenge for the researcher who wishes to separate and visualise isoforms. The technique that is most useful for the separation of protein isoforms is 2-D polyacrylamide gel electrophoresis (PAGE) (see Chap. 2), as it separates proteins by their charge and mass. 'Trains' of protein spots can be seen on many 2-D gels, which represent post-translationally modified versions of a single gene product. The difference gel electrophoresis technique, used in association with 2-D PAGE, can help the discovery of differences in these trains of isoforms between two or more samples (see Chap. 3). Note also that 2-D PAGE can also resolve protein isoforms that result from N-terminal or C-terminal processing of a protein or which are the result of alternate splicing. These events, however, have unpredictable effects on the charge and mass of proteins and are usually only identified by the analysis of discrete protein spots.

The modification of proteins with phosphate, sulfate or carboxyl groups (for example, sialic acid) are direct ways of introducing new negative charges onto proteins and may result in multiple isoforms after separation by 2-D PAGE (Hughes et al. 1992). These trains of spots can vary with disease states and in different tissues. For example, the pattern of spots for serotransferrin from human sera and cerebrospinal fluid are quite different (Wilkins et al. 1996), as are the patterns of transferrin spots in normal patients and those with alcohol-related disease (Gravel et al. 1996) and the isoforms of native and recombinant forms of erythropoietin (Khan et al. 2005). Table 5.3 lists some of the common modifications as well as the modified amino acids that can produce a charge-dependent change to a protein.

Table 5.3 Modifications which lead to a charge-dependent change to a protein

Modification	Amino acids affected and comments
Acylation	Loss of the α-amino positive charge. Predominantly the addition of an acetyl (typically to Ser or Ala) or a pyroglutamyl group (pyrilidone carboxylic acid modification of Gln). In addition formyl, pyruvyl, α-ketobutyryl, glucuronyl, α-aminoacyl groups are found on the α-amino group. Other modifications include fatty acylation, myristoylation, palmitoylation and isoprenylation
Alkylation	Alteration of the α-amino or ε-amino positive charge. The predominant alkyl group is a methyl, which can be monomethyl, dimethyl or trimethyl. Well-known examples are the methyl derivatives of Lys[a] in histones, and His in actin and the myosin light-chain kinase. Other methylated amino acids include Arg, Phe and carboxyl residues. Other modifications to the Lys ε-amino acid include biotinylation, sumoylation, neddylation and ubiquitination
Carboxyl-methylation	Esterification of specific protein carboxyl groups by methyltransferases
Phosphorylation	Predominantly modifications to Ser, Thr and Tyr, but less commonly also Cys and His. Direct linkage to these amino acids or indirect links via oligosaccharides increases the negative charge[b]
Sulfation	Predominantly on Tyr and on oligosaccharides attached to Asn, Thr and Ser (e.g. mucins in cystic fibrosis[c])
Carboxylation	γ-Carboxyglutamate and β-carboxyaspartate both have two neighbouring carboxyl groups and can have two negative charges but with different pK values
Sialylation	Predominantly on oligosaccharides attached to Asn, Thr and Ser
Proteolytic processing	Truncation of protein N- and C-terminal regions which contain charged amino acids will lead to an alteration in pI, while not necessarily resulting in a substantial shift in M_r

[a] The methyl derivatives of Lys can adopt different charges depending on the pH. Both monomethyllysine and dimethyllysine are alkyl amines that can become protonated, while trimethyllysine is a quaternary amine which is always positively charged.
[b] The phosphodiester linked oligosaccharides are labile to mild acid, so caution must be taken when handling this type of modification.
[c] Lo-Guidice et al. (1994).

Mass spectrometry (MS) techniques for protein identification, including peptide mass fingerprinting and 'shotgun' proteomics, are less well suited to the separation of protein isoforms as they analyse peptides rather than proteins. However, specific affinity chromatography resins for the enrichment and separation of specifically modified peptides, such as the use of titanium dioxide to enhance the discovery of phosphoamino acids (Larsen et al. 2005), can assist in the discovery of modifications *per se*. However, this peptide-based analysis does not usually map the modification to any particular isoform within what may be a complex mixture of proteins.

5.3.2 Detection of Co- and Post-translational Modifications

The initial step, once a 2-D separation has been achieved, is to identify which spots carry modifications of interest. Table 5.4 lists a number of methods that can be used for the detection of post-translational modifications and their sensitivity. In hypothesis-independent work, it would be desirable for specific stains to be used sequentially for identifying alterations in post-translational modifications induced by disease, development and other cellular processes. For example, a single 2-D gel or blot could be subjected to different stains to pinpoint glycosylation and phosphorylation, both of which may occur on the same protein. This vision has, in part, been realised with the development of the fluorescent stains ProQ emerald and ProQ diamond. ProQ emerald specifically detects the presence of glycoproteins in 2-D gels (Steinberg et al. 2001) and allows subsequent staining with the protein detection stain SYPRO ruby. As ProQ emerald and SYPRO ruby emit at different wavelengths, gels can be double-stained and scanned with two wavelengths at once. ProQ diamond detects the presence of phosphate moieties on a broad variety of amino acids, and is compatible with downstream MS analysis (Schulenberg et al. 2004).

In addition to the approaches described above to help identify modifications on all and any proteins, there is also a great deal of hypothesis-driven research requiring the direct detection of post-translational modifications. For this, the western blot combined with a variety of detection methods presents a sensitive means for the detection of specific modifications. Phosphorylated peptides, for example, are highly immunogenic and there are a vast number of anti-phosphoamino acid antibodies available. Of these, the anti-phosphotyrosine antibodies are quite generic, whereas most phosphoserine and phosphothreonine antibodies are amino acid sequence specific (Coba et al. 2003). For glycosylation, lectins allow the specific detection

Table 5.4 Methods used for detection of post-translational modifications on gels or blots

Detection method	Medium	Sensitivity[a]	Specificity
Monoclonal antibodies	Nitrocellulose, PVDF	10 ng	Specific epitopes, e.g. sugar, phosphate
Metabolic labelling	Gel, PVDF, nitrocellulose	50 ng	Specific precursors, e.g. ^{32}P, ^{3}H-GalNAc, ^{31}S
Lectins	Nitrocellulose, PVDF	0.1 µg	May be specific to one glycan structure
Fluorescent stains	Gel	1–15 ng	Phosphates, sugars, lipids
MS analysis	Gel, PVDF	5 µg	Modified peptides

[a] The sensitivity will depend on extent of modification of the protein.
PVDF poly(vinylidene difluoride).

of certain types of glycans or of certain monosaccharides. These are used in much the same manner as antibodies, by incubation with western-blotted membranes; however, they provide little or no structural information about any detected glycans. A note of caution should be raised regarding the transfer efficiency of modified proteins from gels to membranes, as small, large or highly charged proteins often do not electroblot efficiently and the standard general protein stains (e.g. Coomassie blue) often do not visualise glycoproteins well.

5.3.3 Strategy for the Analysis of Modifications: Top Down Versus Bottom Up

Since we wrote a similar chapter almost a decade ago (Gooley and Packer 1997), two events have significantly impacted our understanding of the proteome – the sequencing of the human genome and the rise of the mass spectrometer as the tool of choice for protein analysis. The now straightforward approach of digesting a protein to its constituent peptides and searching protein databases with the peptide parent and daughter ion masses resulting from liquid chromatography (LC)– MS-MS has allowed the identification of thousands of proteins from dozens of species and tissues. MS (see Chap. 3) also presents a powerful means to analyse the expression levels of hundreds of proteins at once. In the context of post-translational modifications, however, LC-MS-MS approaches have one major limitation. They do not, and cannot, facilitate the separation of protein isoforms. Instead, they usually present the proteome as a list of expressed proteins, with expression levels, if such experimental approaches have been used. Even if modified peptides are found on a protein, the analytical incompleteness of the LC-MS-MS approaches (Wilkins et al. 2006) mean that it remains impossible to understand the number and type of modified isoforms that are likely to be present for each protein. As most proteins exist as two or more differently modified forms, and multiple modifications on proteins are commonplace (Yang 2005), this is a serious issue. Because of this, the detailed characterisation of whole proteins, purified by 2-D PAGE techniques, is required to get a complete view of the presence and complexity of modifications in the proteome. Thus, when reviewing the field of post-translational modifications in proteomics it must be kept in mind that researchers approach the analysis of post-translational modifications from two different but complementary angles:

1. A top-down approach where single proteins or isoforms of proteins are analysed for their post-translational modifications.
2. A bottom-up approach where the MS-MS analysis of complex mixtures is used to gain a large-scale view on protein post-translational modifications in the proteome.

These two strategies can be used in synergy to help discover and confirm co- and post-translational modifications.

5.3.4 Mass Spectrometry for Analysis of Co- and Post-translational Modifications

The fundamental principle for MS analysis of post-translational modifications is that amino acids, when modified, have a greater mass than their unmodified counterparts. The precise mass of the modifications can be predicted from their atomic composition, allowing a mass difference between modified and unmodified amino acids to be precisely calculated. By extrapolation, the masses of modified and unmodified proteins or peptides can also be predicted. For example, methylation of an amino acid such as lysine will increase the mass of an amino acid by 14 mass units; thus, a peptide or protein carrying a methyllysine will be 14 mass units heavier than one that has an unmodified lysine. Detailed lists of mass differences due to the presence of modifications can be found elsewhere (Wilkins et al. 1999; Creasy and Cottrell 2004).

When applying MS to the analysis of post-translational modifications, there are five major means by which the modifications can be detected and analysed

1. *The precise measurement of the mass of the whole, modified protein*. Software is available to predict the modifications that are likely to be present from the difference between the modified and the unmodified protein (Meng et al. 2004). This approach relies on highly accurate mass measurements, the sequence of the protein being known and a small number of modifications being present. It is best undertaken on small proteins that are analysed with Fourier transform mass spectrometers.
2. *The discovery of mass differences between modified and unmodified peptides during peptide mass fingerprinting*. In this approach, proteins are first digested to peptides. The modified peptides can then be detected as part of the peptide mass-fingerprinting approach (see Chap. 3) owing to the mass differences of modified and unmodified amino acids. Where the identity of the protein is known, sophisticated tools can be used to better predict the presence of modifications (Wilkins et al. 1999). Such approaches first detect the presence of modifications by searching for relevant mass differences, and then consider if a peptide contains the appropriate amino acid(s) that can carry such a modification, if these amino acids are in a sequence that is known to cause the addition of the modification, and if the species itself is known to modify proteins in that way.
3. *The discovery of modifications during the MS-MS fragmentation of peptides*. During the analysis of single proteins or complex mixtures, MS-MS allows the detection of modifications on peptides by firstly determining if their mass is greater than expected and the peptide is thus likely to be modified, and secondly by determining if fragmentation is consistent with

one or more amino acids in the peptide carrying a certain modification. Similar to the approaches outlined above, this has been made possible through the linking of MS to powerful software tools (Creasy and Cottrell 2002; Hernandez et al. 2003; Matthiesen et al 2005). It presents a major avenue for the discovery of modifications on a small or large scale.

4. *The detection of modification-specific 'reporter' ions during peptide fragmentation*. Many modifications, during the course of parent ion mass analysis or peptide fragmentation, undergo specific cleavage events or spontaneous decay. This generates specific reporter ions that indicate the presence of certain modified amino acids. In some cases, the chemical substitution or alteration of a modified amino acid can be undertaken to generate new reporter ions for that modification.
5. *The analysis of modifications after their release from proteins*. This is specific to modifications involving glycosylation and fatty acids. In the case of N-linked glycosylation, the glycan can be released from a protein or peptide with the enzyme PNGaseF. Interestingly, this deglycosylation converts modified asparagine residues into aspartic acid, causing a 1-Da increase in mass, and hence this allows site identification. The released oligosaccharides can then be analysed by fragmentation (Sect. 5.4.4).

The role of MS in the analysis of modifications is central, but it must be kept in mind that it is not an analytical 'holy grail'. This is because the data a mass spectrometer provides are the mass of a compound, or in the case of most analyses for post-translational modifications, the difference between the mass of a modified peptide and the same peptide without modifications. This can be troublesome for the analysis of protein glycosylation. The various hexoses (glucose, mannose, galactose) and inositol all have the same monoisotopic residue mass (162.058 Da), as do glucosamine and galactosamine (161.069 Da). Similarly, there is no mass difference between the acetylated amino sugars GalNAc and GlcNAc (203.079 Da). Thus, if the five different trisialylated oligosaccharides of fetuin are collected and subjected to MS, only a single mass is seen, since the trisialics are isomeric structures of the same monosaccharide constituents. Compositional analyses by other techniques may thus be needed in these cases to corroborate MS-based assignments of glycans. The modifications of phosphate and sulfate also have very similar monoisotopic residue masses (phosphate 79.663 Da, sulfate 79.957 Da), but MS-MS fragmentation can be used to differentiate between them.

5.4 Analysis of Specific Modifications

As mentioned earlier, there are a number of strategies for the analysis of post-translational modifications. Whilst these strategies can, in the most part, be applied to many types of modifications, it may be useful to briefly explore

how four common types of modification have been analysed to date. This should illustrate the utility of different analytical approaches.

5.4.1 Acetylation

N-acetylation is a major eukaryotic co- and post-translational modification of proteins. It is found on the amino terminus of proteins and also on the ε-amino group of lysine residues, particularly on histones. As a modification, it is stable under conditions of MS and has been discovered on proteins using top-down and bottom-up approaches (Thomas et al. 2006; Wang et al. 2005). In addition to studies of protein acetylation using MS, a proteome-wide study has also been undertaken to determine the effect of knocking out an acetyltransferase in yeast. A *Saccharomyces cerevisiae* mutant was isolated that lacked this activity (Mullen et al. 1989). 2-D PAGE was used to compare the soluble proteins from the wild-type yeast with those from mutant cells that lacked the acetyltransferase (Lee et al. 1989). Among 855 proteins found by 2-D PAGE in wild-type and mutant yeast cells, 20% of the protein spots either disappeared or migrated to a different pI in the mutant. Of these, proteins of higher pI did not change in apparent mass but shifted uniformly by +0.1 to +0.2 pI units, a shift consistent with the protonation of the α-NH_2 group. Interestingly, another 12% of proteins were observed to have either increased or decreased expression levels, suggesting that acetylation is involved in a regulatory role with protein expression (Lee et al. 1989). While only 20% of proteins were altered owing to the deletion of an *N*-acetyltransferase gene, others have suggested figures much higher (50%) for soluble N-acetylated proteins (Driessen et al. 1985). The presence of a family of *N*-acetyltransferases may account for these differences (Lee et al. 1989).

5.4.2 Phosphorylation

Traditionally, phosphorylated proteins have been found either by radiolabelling, which cannot be applied in many situations, or by phosphoamino acid specific antibodies on blots from 1-D and 2-D gels (Wang 1988; Goldstein et al. 1995). More recently, MS has become the tool of choice for the identification of phosphoamino acids. A general strategy for the analysis of phosphorylation is for the affinity enrichment of phosphopeptides, followed by their analysis by MS-MS techniques. The affinity enrichment, which has been undertaken using immobilised metal affinity chromatography (Neville et al. 1997), affinity for titanium dioxide (Larsen et al. 2005) or with anti-phosphoamino acid antibodies (Rush et al. 2005) is necessary to ensure that the modified peptides (which are usually present as a minor species) are present in sufficient quantities to permit their analysis. These approaches have

been applied on a large scale, and proteome-wide analysis of phosphorylation, otherwise termed phosphoproteome analysis, has been undertaken in species including yeast (Ficarro et al. 2002), *Arabidopsis* (Jones et al. 2006) and in human HeLA cells (Beausoleil et al. 2004). There is also interest in the use of protein chips for phosphoproteomic research. Chips containing arrays of antibodies specific for a particular protein, used in conjunction with a second phosphoamino acid antibody, allow the capture and semiquantitative monitoring of the phosphorylation status of proteins of interest (Gembitsky et al. 2004).

5.4.3 Ubiquitination and Sumoylation

A rather different type of post-translational modification involves the covalent attachment of the whole proteins, ubiquitin or small ubiquitin-related modifier (SUMO), to other proteins. These modifications are involved in a diversity of processes and are some of the largest known post-translational modifications. Polyubiquitination targets proteins for degradation by the 26S proteasome in the eukaryotic cytosol or in the nucleus and thereby has a crucial role in regulating protein availability. Monoubiquitination serves as a sorting signal that directs membrane proteins for degradation within the lysosomal/vacuolar compartment (Kirkpatrick et al. 2005). SUMO appears to be involved in a larger variety of cellular processes, including nuclear transport, maintenance of genome integrity, transcriptional regulation and signal transduction (Johnson 2004). The functional purpose of the SUMO modification varies from protein to protein, and in most cases remains poorly understood.

A general approach has been described for the detection of ubiquitinated proteins (Peng et al. 2003). Ubiquitin in *S. cerevisiae* was made as a fusion with a multi-histidine tag, and affinity chromatography used to purify the ubiquitin–histidine–protein conjugates. Identification of proteins with MS yielded 1,075 different proteins that had been modified by covalent linkage with ubiquitin. The same research team extended this strategy to the identification of SUMO-modified proteins (Denison et al. 2005), whereby they identified about 250 sumoylated proteins. These and similar strategies have been adopted as a means to generate a proteome-scale view of such modifications (Rosas-Acosta et al. 2005).

5.4.4 Glycosylation

The characterisation of protein glycosylation has long been considered to be complex. This is because microheterogeneity can occur through the presence of different glycoforms at a single amino acid site, as can macroheterogeneity which results from the variable presence of glycoforms at

different amino acids in different molecules of the same protein. It has been calculated that recombinant tissue plasminogen activator, which has three sites of completely characterised N-glycosylation, has potentially 11,520 possible isoforms (Appfel et al. 1995). Fortunately, biology seems to be relatively well ordered so that glycoproteins, although they have a huge potential heterogeneity, are limited in their structure depending on the availability of tissue-specific glycosyltransferases. It is estimated that the glycosylation machinery, comprising glycosyltransferases and glycosidases, makes up at least 1% of the human genome. The apparent redundancy of these genes, for example the 24 unique serine/threonine *O*-GalNAc transferases carrying out the same single step of adding the first sugar to O-linked serines and threonines, may be explained by the tissue-specific expression of these genes. This specificity emphasises the importance of characterising the differences between these glycoforms. This characterisation is multidimensional if a detailed structure is required, and involves extensive analyses to determine (1) the monosaccharide composition, (2) the attachment site(s) to the protein, (3) the sequence, branching and linkage positions and (4) the anomeric configuration of the monosaccharides.

2-D electrophoresis often provides the first clue to the presence of protein glycosylation as seen by the trains of spots visible on most gel separations of proteins from eukaryotic cells. In fact, 68% of human proteins have the N-linked NXS/T glycosylation motif and 50% of the expressed genes are estimated actually be glycosylated (Apweiler et al. 1999; Zhang et al. 2006). Sample prefractionation based on mass, charge or affinity to the sugar-binding lectins is used, as in general proteomics, to enrich the sample for glycoproteins for further analysis. Releasing the oligosaccharides enzymatically or chemically results in the loss of site-specific information but is our preferred initial approach. It allows for the analysis of glycan heterogeneity independently of the peptide. Site information can be determined in parallel by the capture of those peptides which are glycosylated (Zhang and Aebersold 2006), by the appearance of new peptides in a mass profile after deglycosylation (Wilson et al. 2002) or by the search for specific glycopeptide masses after glycosylation structures present on the protein are known.

The analytical strategies have now been downscaled to successfully characterise the glycopeptides and oligosaccharides of glycoproteins separated at the level of 2-D PAGE (Wilson et al. 2002) and microLC (Karlsson et al. 2004). Characterisation has moved from the classical chemical derivatisation and chromatographic separation to the use of a range of MS techniques (reviewed by Morelle et al. 2006; Zaia 2004). Ion-trap MS offers the most information because of its ability to continue fragmenting (MSn) the monosaccharide substructures and obtain data which can differentiate between isomers (Ashline et al. 2005). The affinity of lectins for different glycan epitopes is being used for analysis and enrichment (Hirabayashi 2004) as well as in the production of microarray chips designed to profile the heterogeneity of glycosylation on a protein (Koopmann and Blackburn 2003).

An important new step in enabling the analysis of glycosylation has been the development of glycoinformatics. A number of glycan structure databases have been developed, such as GlycoSuiteDB (2004), Glycosciences (2006), KEGG GLYCAN (2006), and Glycan Database (2006). Note, however, that these are currently not integrated and thus contain different information formatted in different ways. A challenge has been in the formatting of the structures themselves; although there are only six monosaccharide residue masses found in mammalian glycans: Hex, HexNAc, DeoxyHex, Pent, NeuAc, NeuGc, these can be linked in five ways and be adorned by methyls, acetyls, phosphates and sulfates at different positions. The same oligosaccharide composition can exist as many different structural configurations. The possible variation increases if different species are included. Computer-readable and human-readable formats of structures are required.

An analogous process to peptide mass sequencing has been developed, whereby MS-MS fragmentation data are matched against glycan structure databases to allow the determination of a glycan structure (Ethier et al. 2003; Joshi et al. 2004). A limitation of this approach is that the databases remain relatively small and thus are limited to those structures previously reported. Alternatively generic "bottom up" approaches are being developed in which the fragmentation characteristics of oligosaccharide substructures are being used to piece together a probable parent structure (Tseng et al. 1999; Zhang et al. 2005; Tang et al. 2005). Yet another approach to the determination of glycan structures is the "top down" method being developed in which fragmentation trees based on a parent ion topology are predicted and matched (Gaucher et al. 2000; Lapadula et al. 2005). In the fullness of time, and with the development of these tools, the perceived difficulty of determining the structure and the attachment site of glycans to proteins will become facile.

5.5 The Function of Protein Post-translational Modifications: More Than Meets the Eye?

The type of analytical knowledge described in the previous sections is just a prelude to the question: What is the function of these modifications in a protein? More importantly, if we think in human terms, how can we use this knowledge to prevent or treat the manifestations of errors in the post-translational processing of the gene in human disease. After the Human Genome Project, there has been a focus on using the knowledge of the genetic complement of the human to identify and modify the genetic defects responsible for disease. However, it has become obvious that whilst some answers can be determined this way, and notwithstanding the absolute need for this knowledge, much cannot be explained by simply thinking in genetic terms. The presence or absence of a gene of course is critical to function, but

Table 5.5 Known functions of some post-translational modifications

Modification	Known function	Review
Phosphorylation	Reversible molecular switch, signal transduction	Cohen (2000)
Acetylation	Capping of protein amino termini, reversible molecular switch	Glozak et al. (2005)
SUMO	Implicated in nuclear import, transcriptional regulation and modulation of protein–protein interactions. Distinct from ubiquitination as it does not target proteins for degradation	Hay (2005)
Ubiquitination	Polyubiquitination targets proteins for degradation; monoubiquitination functions as a sorting signal	Kirkpatrick et al. (2005)
Methylation	Regulation of expression of genes	Martin and Zhang (2005)
Sulfation	Modulator of protein–protein interactions of secreted and membrane-bound proteins	Kehoe and Bertozzi (2000)
Glycosylation	Protein–protein interactions, protection from proteolysis, directs protein folding in the Golgi apparatus	Ohtsubo and Marth (2006)
O-GlcNAc	Reversible molecular switch, exists inter changeably with phosphorylation	Slawson and Hart (2003)

incorrect modification of the protein is increasingly being found to affect the correct functioning of the gene product.

The majority of the published literature, however, describes the function of only a single type of protein modification. There are a number of excellent reviews which summarise the known functions of a specific co- and post-translational modification to proteins (summarised in Table 5.5). The tendency is to analyse the specific modification that the laboratory is familiar with, and is able to analyse, but it is essential that scientists remain open to the presence and interaction of numerous modifications that may occur on a single protein. The analysis of nine isoforms of serum amyloid A protein by Ducret et al. (1996) is a good example of the heterogeneity of modifications that can affect a single gene product. The amyloid A protein exists in forms in which the N-terminal arginine is missing, some of the protein is glycosylated, there is oxidation at tryptophan residues and there is suspected modification in the form of dimethyl asparagine.

Interrelations of various post-translational control mechanisms have been found on many proteins. Numerous functional proteins are modified at multiple sites by additions such as phosphorylation, acetylation, methylation, ubiquitination, sumoylation and citrullination (reviewed by Yang 2005). For example, ubiquitination is used as both a degradation and a protein sorting signal. On the other hand, proteolysis-linked multiubiquitination can be

regulated by phosphorylation (Huang et al. 2006). Reversible multisite modifications of methylation, acetylation, ubiquitination and phosphorylation have all been described on the N-terminus of histones and are thought to affect gene expression and chromatin structure crucial to development and differentiation processes (Beck et al. 2006; Villar-Garea and Imhof 2006). An adjacent proline-directed phosphorylation site has been found to regulate the sumoylation of several transcription factors (Hietakangas et al. 2006). As another example, there is a complementary switching mechanism at specific amino acids (serine, threonine) between phosphorylation and *N*-acetylglucosamine addition. This is involved in many nuclear/cytoplasmic control mechanisms (Slawson and Hart 2003). In addition, the requirement of oligomeric complexes for many protein functions points at these protein–protein assembly processes as a general mechanism for regulation at the post-translational level.

The function of glycosylation is moving away from the view that 'all theories are correct' (Varki 1993) in which there are a plethora of seemingly contradictory activities affected by adding a sugar to a protein. The significance of glycosylation is now being defined in a range of critical cellular functions and diseases, predominantly at this stage focused on inflammation and cancer (Dube and Bertozzi 2005) and pathogenic infection (Sharon 2006). In addition, such dogmas as the notion that bacteria do not glycosylate their proteins have been overturned (Szymanski and Wren 2005). Importantly, in the context of human health, the types of oligosaccharide structures that are being added to recombinant therapeutic drugs are coming under the close scrutiny of the regulatory authorities in the USA and the EU as they are being seen to not only be critical to activity but also to vary during production in eukaryotic cell-expression systems (Bondaryk and Packer 2004).

5.6 Some Interesting Modification Stories

The literature is just beginning to better describe the importance of post-translational modifications. To illustrate this concept, we have chosen some 'stories' of specific proteins whose activity is affected by one or more modifications to the expressed amino acid sequence. Our focus is on some of those which are known to be important to humans in a medical context.

5.6.1 The Erythropoietin Story

Erythropoietin has become the most produced recombinant protein drug in the world. It increases patient red blood cell counts; hence, there is a very large market for counteracting anaemia associated with cancer treatment and kidney disease. It is also used illegally as a performance-enhancing drug in sports.

The protein exists as numerous isoforms as a result of the extensive heterogeneity of glycosylation on three N-linked sites and two O-linked sites. This is easily seen in a 2-D electrophoretic separation as differential sialylation confers different net charge on the glycoforms. The recombinant drug form, produced in CHO cell culture, has fewer sialylated sugars and causes the isoforms to focus at a more basic pI (Fig. 5.2). Whereas this forms the basis of drug detection in urinary sports drug testing (Khan et al. 2005), it presents a problem for the generic production of the drug as more companies start production in the wake of the expiry of the original patent. As seen (Fig. 5.3), the glycosylation profile of these products can vary and

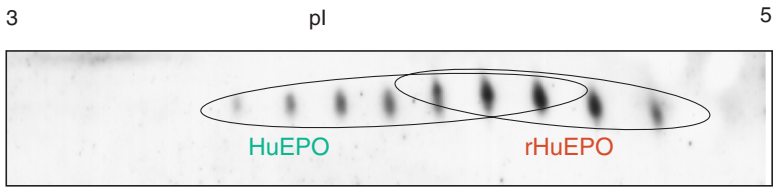


Fig. 5.2 Separation by 2-D gel electrophoresis of the isoforms of erythropoietin due to different glycosylation; HuEPO (native human isoforms), rHuEPO (recombinant isoforms expressed in CHO cells). (From Khan et al. 2005)

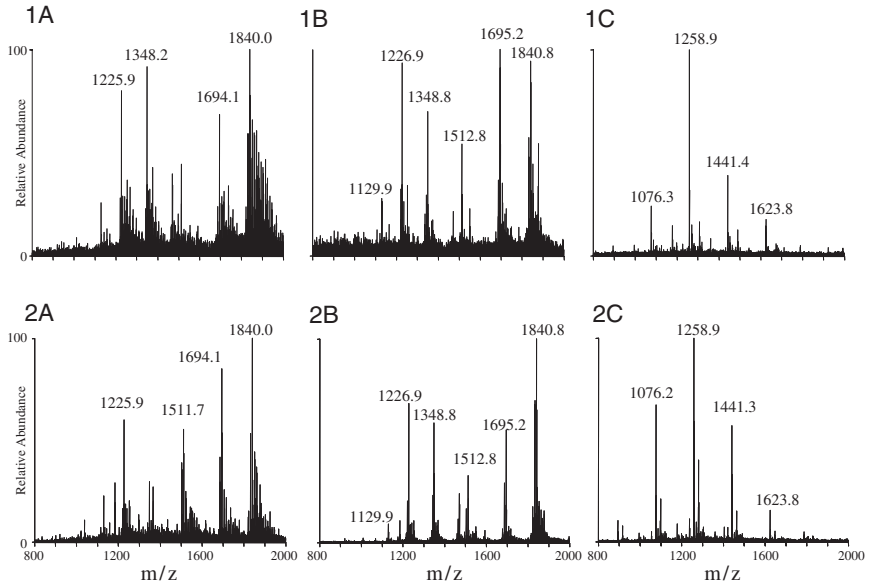


Fig. 5.3 Mass-spectrometric profiles of two different supplies (*1A–1C*, *2A–2C*) of recombinant EPO showing variation in glycan modifications. *1A*, *2A* Oligosaccharides released by PNGaseF. *1B*, *2B* Deacetylated and reduced oligosaccharides released by PNGaseF. *1C*, *2C* Deacetylated, desialylated and reduced oligosaccharides released by PNGaseF

is dependent upon the cell-expression system as well as the culture conditions. Although the different glycosylation of the recombinant drug compared with that found naturally in the blood appears not to affect the erythropoietic activity of the protein, adding two more sites of glycosylation on erythropoietin and expressing in CHO cells results in three times the half-life in the body and increased *in vivo* activity of the drug (Sinclair and Elliott 2005). Longer retention of the drug has the beneficial clinical effect of decreasing the number of injections required for patients.

5.6.2 The Apolipoprotein E Story

Apolipoprotein E (ApoE) is a plasma glycoprotein and is a component of several classes of lipoproteins involved in cholesterol transport and clearance. The ApoE protein appears as several isoforms on a 2-D gel. In this case the isoforms are due to glycosylation heterogeneity as well as to the allelic variants common to genetically complex organisms. There are three major isoforms of ApoE, referred to as E2, E3 and E4 coded for by three alleles (Table 5.6). The presence of the three alleles gives rise to six different genotypes. The inherited ApoE genotype of an individual is postulated to affect susceptibility to disease and the different isoforms have been associated with an increased risk of atherosclerosis and neurodegenerative disorders, including Alzheimer's disease (Hatters et al. 2006).

The common isoforms of ApoE thus differ owing to amino acid substitutions that result from single nucleotide substitutions. The amino acid substitutions are shown in Table 5.6 and as arginine is substituted for cysteine, new sites of tryptic digestion are created. Peptide mass fingerprinting of these isoforms can thus be used to easily type the allelic variation by the specific diagnostic ions associated with the variants (Table 5.7, Fig. 5.4). Currently nucleic acid polymerase chain reaction methods are used to identify the variants in individuals but proteomic approaches such as this can also provide a means to determine genetic variation by the direct analysis of the gene product.

Table 5.6 Amino acid substitutions in common isoforms of apolipoprotein E

Isoform	Amino acid 130	Amino acid 176
E2	Cys	Cys
E3	Cys	Arg
E4	Arg	Arg

Table 5.7 Diagnostic peptide masses for the common apolipoprotein E isoforms

	Diagnostic peptides			
Isoform	Position of peptide	Peptide mass (Da)	Position of peptide	Peptide mass (Da)
E2	176-185	1,051.5 (1,108.6[a])	122-132	1,165.5 (1,222.5[a])
E3	177-185	948.5	122-132	1,165.5 (1,222.5[a])
E4	177-185	948.5	122-130	1,005.5

[a]Peptide mass on alkylation of cysteines with iodoacetamide.

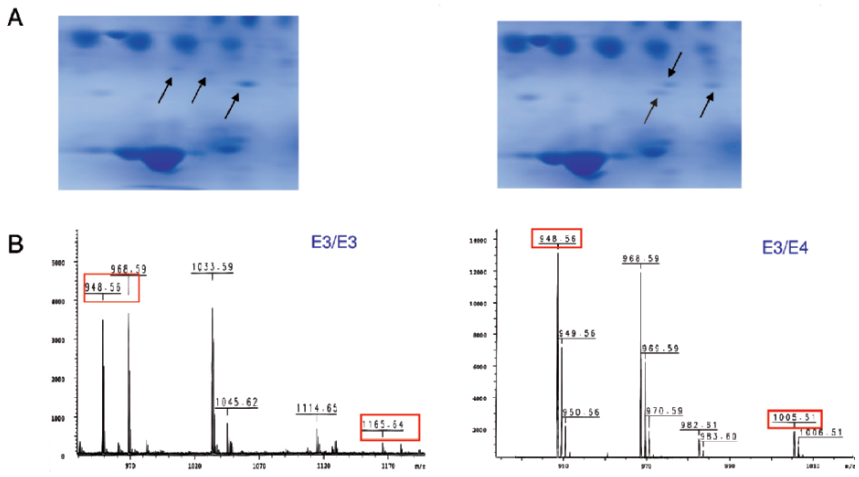


Fig. 5.4 Identification of different allelic (E3/E3 and E3/E4) variants of APOE. **a** 2-D gel electrophoresis. **b** Diagnostic peptide mass fingerprinting ions

5.6.3 The Progeria Story

Progeria, or Hutchinson's Gilford progeria, is a very rare, fatal syndrome of accelerated and premature aging. Typical characteristics include growth retardation, loss of subcutaneous tissue, loss of hair and prominent scalp veins, aged appearance of skin, prominent joints and osteoporosis and progressive atherosclerotic cardiovascular disease. It is considered to be a model system for understanding the metabolism of aging.

Alterations to the protein lamin A have been implicated as the cause of this disease. Lamin A is a major component of the scaffold of proteins just inside the nuclear membrane called the lamina. Mature lamin A is formed by post-translational processing of a pre-lamin A precursor. The pre-lamin A protein has a CAAX box at its carboxy terminus, which signals isoprenylation, the addition of a farnesyl group to the cysteine by the enzyme farnesyltransferase.

Addition of the farnesyl moiety allows pre-lamin A to localise to the nuclear periphery, where the final cleavage step of a portion of the protein, including the farnesyl group, frees lamin A to integrate properly into the nuclear lamina. The truncated lamin A protein, termed progerin, has an internal deletion of 50 amino acids which produces a protein that carries the farnesylation motif but lacks the cleavage motif, and it is thought that the farnesylated lamin A becomes permanently attached to the inner nuclear membrane and cannot be integrated into the nuclear scaffold.

Farnesyltransferase inhibitors are drugs which were originally developed to inhibit some cancer-causing proteins that require farnesylation for function, like the oncoprotein Ras, and are now being tested in phase III clinical trials of patients with myeloid leukemia. These drugs had the effect of reversing the dramatic nuclear structure abnormalities in cells grown from skin biopsies of progeria patients (Capell et al. 2005; Glynn and Glover 2005). The inhibition of this post-translational modification thus promises to be a possible target not only in the treatment of this rare disease but more generally in the metabolic defects involved in aging and cancer.

5.6.4 The Influenza Story

A topical story is the potential threat of a pandemic of influenza resulting from mutations in the highly pathogenic avian influenza virus. Influenza is a highly contagious, acute, viral infection of the respiratory tract. The causative agent of the disease is an immunologically diverse, single-strand RNA virus. Type A viruses are the most prevalent and are associated with most serious health risks and epidemics. Glycosylation of both the host cell receptor and the two main viral membrane proteins is intrinsically involved in many aspects of the pathogenicity of the virus.

The main determinants of infection are two viral envelope glycoproteins haemagglutinin and neuraminidase, which interact with the sugars on the membrane of the host cell. Haemagglutinin is a spiky projection from the envelope, the purpose of which is to bind to host cells, and then fuse with the envelope and plasma membrane. The conserved carbohydrate chains in the stem of the spike stabilise haemagglutinin in the form susceptible to the conformational change necessary for fusion (Ohuchi et al. 1997). The precursor haemagglutinin protein undergoes post-translational cleavage into two subunits by the host proteases which are essential for viral infectivity. There is evidence that a site-specific glycosylation affects this cleavage of the influenza virus haemagglutinin and thus virulence (Deshpande et al. 1987). The receptor for haemagglutinin on the host cell surface is the sugar *N*-acetylneuraminic acid (sialic acid). Haemagglutinin on avian viruses preferentially binds to α2-3-linked sialic acids on receptors of intestinal epithelial cells, whereas human viral haemagglutinin is specific for the α2-6 linkage on epithelial cells of the lungs and upper respiratory tract. This seemingly small

variation in sugar structure may confer a control point on the transmission of the influenza virus from birds to humans.

The action of the other major viral envelope protein, neuraminidase, is to cleave the sialic acid from the membrane glycolipid so the new virus particles can be released from host cells. An additional glycosylation site within the neuraminidase protein globular head has been reported to contribute to the high virulence of the H5N1 virus (Hulse et al. 2004). In addition, the specificity of this neuraminidase has been the target for the successful development of two influenza drug therapies (Relenza™, Tamiflu™) that are synthetic structural analogues of the neuraminidase active site, and which thus inhibit the transmission of the virus.

5.7 Future Directions

This chapter has outlined the importance that co- and post translational modifications have in distinguishing the proteome from the genome. We have touched on most of the technologies currently used to analyse these modified proteins. Many of these co- and post-translational modifications are closely regulated and change depending on the tissue, developmental stage, disease state or age of the cell and we now have some of the tools necessary to identify, locate and characterise the differences induced by these modifications. The challenge still remains for the scientist to move beyond protein identification to look deliberately for the type, position and function of the many modifications to the amino acid sequence. More importantly, now that the technologies are able to work at these levels, the challenge is to link these modifications to metabolic and functional changes in the cell. Moreover, we need to use the identified alterations of proteins to help us develop targeted diagnostics and drugs to control those protein modifications responsible for human disease.

Acknowledgements. This chapter contains some original unpublished data acquired as part of the work of Niclas Karlsson (Sect. 5.6.1) and Karen Morris (Sect. 5.6.2). The authors acknowledge the support of Proteome Systems Limited.

References

Angelov I (1994) Characterisation of a proline-directed casein kinase from bovine brain. Arch Biochem Biophys 310:97–107

Apostol I, Aitken J, Levine J, Lippincott J, Davidson, JS, Abbott-Brown D (1995) Recombinant protein sequences can trigger methylation of N-terminal amino acids in *Escherichia coli*. Protein Sci 4:2616–2618

Appfel A, Chakel J, Udiavar S, Hancock W, Souders C, Pungor E Jr (1995) Use of hyphenated liquid-phase analyses and mass spectrometric approaches for the characterisation of

glycoproteins derived from recombinant DNA. In: Snyder AP (ed) Biochemical and biotechnological applications of electrospray ionization mass spectrometry. American Chemical Society, Washington, pp 432–471

Apweiler R, Hermjakob H, Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochim Biophys Acta 1473:4–8

Ashline D, Singh S, Hanneman A, Reinhold V (2005) Congruent strategies for carbohydrate sequencing. 1. Mining structural details by MS(n). Anal Chem 77:6250–6262

Beausoleil SA, Jedrychowski M, Schwartz D, Elias JE, Villen J, Li J, Cohn MA, Cantley LC, Gygi SP (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc Natl Acad Sci USA 101:12130–12135

Beck HC, Nielsen EC, Matthiesen R, Jensen LH, Sehested M, Finn P, Grauslund M, Hansen AM, Jensen ON (2006) Quantitative proteomic analysis of post-translational modifications of human histones. Mol Cell Proteomics 5:1314–1325

Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 4:1633–1649

Bondaryk RP, Packer NH (2004) Microheterogeneity, standardization and characterization in glycoprotein drugs. Curr Drug Discov 4:31–32

Capell BC, Erdos MR, Madigan JP, Fiordalisi JJ, Varga R, Conneely KN, Gordon LB, Der CJ, Cox AD, Collins FS (2005) Inhibiting farnesylation of progerin prevents the characteristic nuclear blebbing of Hutchinson-Gilford progeria syndrome. Proc Natl Acad Sci USA 102:12879–12884

Coba MP, Turyn D, Pena C (2003) Synthesis and immunogenic properties of phosphopeptides related to the human insulin receptor. J Pept Res 61:17–23

Cohen P (2000) The regulation of protein function by multisite phosphorylation – a 25 year update. Trends Biochem Sci 25:596–601

Glycan Database (2006) Consortium for Functional Glycomics, La Jolla, USA. http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp. Cited 22 March 2007

Creasy DM, Cottrell JS (2002) Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics 2:1426–1434

Creasy DM, Cottrell JS (2004) Unimod: protein modifications for mass spectrometry. Proteomics 4:1534–1536

Denison C, Rudner AD, Gerber SA, Bakalarski CE, Moazed D, Gygi SP (2005) A proteomic strategy for gaining insights into protein sumoylation in yeast. Mol Cell Proteomics 4:246–254

Deshpande KL, Fried VA, Ando M, Webster RG (1987) Glycosylation affects cleavage of an H5N2 influenza virus hemagglutinin and regulates virulence. Proc Natl Acad Sci USA 84:36–40

Driessen HP, De Jong WW, Tesser GI, Bloemendal H (1985) The mechanism of N-terminal acetylation of proteins. Crit Rev Biochem 18:281–325

Dube DH, Bertozzi CR (2005) Glycans in cancer and inflammation–potential for therapeutics and diagnostics. Nat Rev Drug Discov 4:477–488

Ducret A, Bruun CF, Bures EJ, Marhaug G, Aebersold R (1996) Characterization of human serum amyloid A protein isoforms separated by two-dimensional electrophoresis by liquid chromatography/electrospray ionization tandem mass spectrometry. Electrophoresis 17:866–876

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

Ethier M, Saba JA, Spearman M, Krokhin O, Butler M, Ens W, Standing KG, Perreault H (2003) Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. Rapid Commun Mass Spectrom 17:2713–2720

Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, Hunt DF, White FM (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. Nat Biotechnol 20:301–305

Gaucher SP, Morrow J, Leary JA (2000) STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. Anal Chem 72:2331–2336
Gembitsky DS, Lawlor K, Jacovina A, Yaneva M, Tempst P (2004) A prototype antibody microarray platform to monitor changes in protein tyrosine phosphorylation. Mol Cell Proteomics 3:1102–1118
Glozak MA, Sengupta N, Zhang X, Seto E (2005) Acetylation and deacetylation of non-histone proteins. Gene 363:15–23
Glynn MW, Glover TW (2005) Incomplete processing of mutant lamin A in Hutchinson-Gilford progeria leads to nuclear abnormalities, which are reversed by farnesyltransferase inhibition. Hum Mol Genet 14:2959–2969
Goldstein M, Lee KY, Lew JY, Harada K, Wu J, Haycock JW, Hokfelt T, Deutch AY (1995) Antibodies to a segment of tyrosine hydroxylase phosphorylated at serine 40. J Neurochem 64:2281–2287
Gooley A, Packer N (1997) The importance of co- and post-translational modifications in proteome projects. In: Wilkins MR, Williams KL, Appel RD, Hochstrasser DF (eds) Proteome research: new frontiers in functional genomics. Springer, Berlin. pp 65–91
Gravel P, Walzer C, Aubry C, Balant LP, Yersin B, Hochstrasser DF, Guimon J (1996) New alterations of serum glycoproteins in alcoholic and cirrhotic patients revealed by high resolution two-dimensional gel electrophoresis. Biochem Biophys Res Commun 220:78–85
Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. Nucleic Acids Res 27:370–372
Glycosciences (2006) DKFZ, Heidelberg, Germany. http://www.glycosciences.de/sweetdb. Cited 22 March 2007
GlycoSuiteDB (2004) Proteome Systems, Sydney, Australia. http://www.glycosuite.com. Cited 22 March 2007
Hatters DM, Peters-Libeu CA, Weisgraber KH (2006) Apolipoprotein E structure: insights into function. Trends Biochem Sci 31:445–454
Hay RT (2005) SUMO: a history of modification. Mol Cell 18:1–12
Hernandez P, Gras R, Frey J, Appel RD (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics. 3:870–878
Hietakangas V, Anckar J, Blomster HA, Fujimoto M, Palvimo JJ, Nakai A, Sistonen L (2006) PDSM, a motif for phosphorylation-dependent SUMO modification. Proc Natl Acad Sci USA 103:45–50
Hirabayashi J (2004) Lectin-based structural glycomics: glycoproteomics and glycan profiling. Glycoconj J 21:35–40
House C, Wettenhall RE, Kemp BE (1987) The influence of basic residues on the substrate specificity of protein kinase C. J Biol Chem 262:772–777
Huang F, Kirkpatrick D, Jiang X, Gygi S, Sorkin A (2006) Differential regulation of EGF receptor internalization and degradation by multiubiquitination within the kinase domain. Mol Cell 21:737–748
Hughes GJ, Frutiger S, Paquet N, Ravier F, Pasquali C, Sanchez JC, James R, Tissot JD, Bjellqvist B, Hochstrasser D (1992) Plasma protein map: an update by microsequencing. Electrophoresis 13:707–714
Hulse DJ, Webster RG, Russell RJ, Perez DR (2004) Molecular determinants within the surface proteins involved in the pathogenicity of H5N1 influenza viruses in chickens. J Virol 7:9954–9964
Jensen ON (2006) Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol 7:391–403
Johnson ES (2004) Protein modification by SUMO. Annu Rev Biochem 73:355–382
Johnson RD, Bhatnagar RS, Knoll LJ, Gordon JI (1994) Genetic and biochemical studies of protein N-myristoylation. Annu Rev Biochem 63:869–914
Jones AM, Bennett MH, Mansfield JW, Grant M (2006) Analysis of the defence phosphoproteome of *Arabidopsis thaliana* using differential mass tagging. Proteomics 6:4155–4165

Joshi HJ, Harrison MJ, Schulz BL, Cooper CA, Packer NH, Karlsson, NG (2004) Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. Proteomics 4:1650–1664

Karlsson NG, Wilson NL, Wirth HJ, Dawes P, Joshi H, Packer NH. (2004) Negative ion graphitised carbon nano-liquid chromatography/mass spectrometry increases sensitivity for glycoprotein oligosaccharide analysis. Rapid Commun Mass Spectrom 18:2282–2292

KEGG GLYCAN (2006) Kanehisa Laboratories, Kyoto, Japan. http://www.genome.jp/kegg/glycan/. Cited 22 March 2007

Kehoe JW, Bertozzi CR (2000) Tyrosine sulfation: a modulator of extracellular protein-protein interactions. Chem Biol 7:R57–61

Khan A, Grinyer J, Truong ST, Breen EJ, Packer NH (2005) New urinary EPO drug testing method using two-dimensional gel electrophoresis. Clin Chim Acta 358:119–130

Kirkpatrick DS, Denison C, Gygi SP (2005) Weighing in on ubiquitin: the expanding role of mass-spectrometry-based proteomics. Nat Cell Biol 7:750–757

Koopmann JO, Blackburn J (2003) High affinity capture surface for matrix-assisted laser desorption/ionisation compatible protein microarrays. Rapid Commun Mass Spectrom 17:455–462

Lapadula AJ, Hatcher PJ, Hanneman AJ, Ashline DJ, Zhang H, Reinhold VN (2005) Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MS(n) data. Anal Chem 77:6271–6279

Larsen MR, Thingholm TE, Jensen ON, Roepstorff P, Jorgensen TJ (2005) Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. Mol Cell Proteomics 4 873–886

Lee FJ, Lin LW, Smith JA (1989) N alpha-acetyltransferase deficiency alters protein synthesis in *Saccharomyces cerevisiae*. FEBS Lett 256:139–142

Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, Wang TH (2006) dbPTM: an information repository of protein post-translational modification. Nucleic Acids Res 34:D622–627

Lo-Guidice JM, Wieruszeski JM, Lemoine J, Verbert A, Roussel P, Lamblin G (1994) Sialylation and sulfation of the carbohydrate chains in respiratory mucins from a patient with cystic fibrosis. J Biol Chem 269:18794–18813

Lopez-Rodas G, Georgieva EI, Sendra R, Loidl P (1991) Histone acetylation in *Zea mays* I. J Biol Chem 266:18745–18750

Martin C, Zhang Y (2005) The diverse functions of histone lysine methylation. Nat Rev Mol Cell Biol 6:838–849

Matthiesen R, Trelle MB, Hojrup P, Bunkenborg J, Jensen ON (2005) VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. J Proteome Res 4:2338–2347

Meng F, Du Y, Miller LM, Patrie SM, Robinson DE, Kelleher NL (2004) Molecular-level description of proteins from *Saccharomyces cerevisiae* using quadrupole FT hybrid mass spectrometry for top down proteomics. Anal Chem 76:2852–2858

Morelle W, Canis K, Chirat F, Faid V, Michalski JC (2006) The use of mass spectrometry for the proteomic analysis of glycosylation. Proteomics 6:3993–4015

Mullen JR, Kayne PS, Moerschell RP, Tsunasawa S, Gribskov M, Colavito-Shepanski M, Grunstein-Sherman F, Sternglanz R (1989) Identification and characterisation of genes and mutants in N-terminal acetyltransferase from yeast. EMBO J 8:2067–2075

Neville DC, Rozanas CR, Price EM, Gruis DB, Verkman AS, Townsend RR (1997) Evidence for phosphorylation of serine 753 in CFTR using a novel metal-ion affinity resin and matrix-assisted laser desorption mass spectrometry. Protein Sci 6:2436–2445

Ohtsubo K, Marth JD (2006) Glycosylation in cellular mechanisms of health and disease. Cell 126:855–867

Ohuchi R, Ohuchi M, Garten W, Klenk HD (1997) Oligosaccharides in the stem region maintain the influenza virus hemagglutinin in the metastable form required for fusion activity. J Virol 71:3719–3725

Peng J, Schwartz D, Elias JE, Thoreen CC, Cheng D, Marsischky G, Roelofs J, Finley D, Gygi SP (2003) A proteomics approach to understanding protein ubiquitination. Nat Biotechnol 21:921–926

Perlman D, Halvorson HO (1983) A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. J Mol Biol 167:391–409

Rosas-Acosta G, Russell WK, Deyrieux A, Russell DH, Wilson VG (2005) A universal strategy for proteomic studies of SUMO and other ubiquitin-like modifiers. Mol Cell Proteomics 4:56–72

Rush J, Moritz A, Lee KA, Guo A, Goss VL, Spek EJ, Zhang H, Zha XM, Polakiewicz RD, Comb MJ (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. Nat Biotechnol 23:94–101

Schulenberg B, Goodman TN, Aggeler R, Capaldi RA, Patton WF (2004) Characterization of dynamic and steady-state protein phosphorylation using a fluorescent phosphoprotein gel stain and mass spectrometry. Electrophoresis 25:2526–2532

Sharon N (2006) Carbohydrates as future anti-adhesion drugs for infectious diseases. Biochim Biophys Acta 1760:527–537

Sinclair AM, Elliott S (2005) Glycoengineering: the effect of glycosylation on the properties of therapeutic proteins. J Pharm Sci 94:1626–1635

Slawson C, Hart GW (2003) Dynamic interplay between O-GlcNAc and O-phosphate: the sweet side of protein regulation. Curr Opin Struct Biol 13:631–636

Steinberg TH, Pretty On Top K, Berggren KN, Kemper C, Jones L, Diwu Z, Haugland RP, Patton WF (2001) Rapid and simple single nanogram detection of glycoproteins in polyacrylamide gels on electroblots. Proteomics 1:841–855

Szymanski CM, Wren BW (2005) Protein glycosylation in bacterial mucosal pathogens. Nat Rev Microbiol 3:25–37

Tang H, Mechref Y, Novotny MV (2005) Automated interpretation of MS/MS spectra of oligosaccharides. Bioinformatics 21:i431–439

Terwilliger TC, Wang JY, Koshland DE (1986) Kinetics of receptor modification: the multiply methylated aspartate receptors involved in bacterial chemotaxis. J Biol Chem 261:10814–10820

Thomas CE, Kelleher NL, Mizzen CA (2006) Mass spectrometric characterization of human histone H3: a bird's eye view. J Proteome Res 5:240–247

Tseng, K, Hedrick JL, Lebrilla CB (1999) Catalog-library approach for the rapid and sensitive structural elucidation of oligosaccharides. Anal Chem 71:3747–3754

Varki A (1993) Biological roles of oligosaccharides: all of the theories are correct. Glycobiology 3:97–130

Villar-Garea A, Imhof A (2006) The analysis of histone modifications. Biochim Biophys Acta 1764:1932–1939

Wang D, Thompson P, Cole PA, Cotter RJ (2005) Structural analysis of a highly acetylated protein using a curved-field reflectron mass spectrometer. Proteomics 5:2288–2296

Wang JY (1988) Antibodies for phosphotyrosine: analytical and preparative tool for tyrosyl-phosphorylated proteins. Anal Biochem 172:1–7

Wilkins MR, Sanchez JC, Williams KL, Hochstrasser DF (1996) Current challenges and future applications for protein maps and post-translational vector maps in proteome projects. Electrophoresis 17:830–838

Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, Ou K, Sanchez JC, Bairoch A, Williams KL, Hochstrasser DF (1999) High-throughput mass spectrometric discovery of protein post-translational modifications. J Mol Biol 289:645–657

Wilkins MR, Appel RD, Van Eyk JE, Chung MC, Gorg A, Hecker M, Huber LA, Langen H, Link AJ, Paik YK, Patterson SD, Pennington SR, Rabilloud T, Simpson RJ, Weiss W, Dunn MJ (2006) Guidelines for the next 10 years of proteomics. Proteomics 6:4–8

Wilson NL, Schulz BL, Karlsson NG, Packer NH (2002) Sequential analysis of N- and O-linked glycosylation of 2D-PAGE separated glycoproteins. J Proteome Res 1:521–529

Yamada R, Bradshaw RA (1991) Rat liver polysome N alpha-acetyltransferase: substrate specificity. Biochemistry 30:1017–1021

Yang SH, Galanis A, Witty J, Sharrocks AD (2006) An extended consensus motif enhances the specificity of substrate modification by SUMO. EMBO J 25:5083–5093

Yang XJ (2005) Multisite protein modification and intramolecular signaling. Oncogene 24:1653–1662

Zaia J (2004) Mass spectrometry of oligosaccharides. Mass Spectrom Rev 23:161–227

Zhang FL, Casey PJ (1996) Protein prenylation: molecular mechanisms and functional consequences. Annu Rev Biochem 65:241–269

Zhang H, Aebersold R (2006). Isolation of glycoproteins and identification of their N-linked glycosylation sites. Methods Mol Biol 328:177-185

Zhang H, Singh S, Reinhold VN (2005) Congruent strategies for carbohydrate sequencing. 2. FragLib: an MS(n) spectral library. Anal Chem 77:6263–6270

Zhang H, Loriaux P, Eng J, Campbell D, Keller A, Moss P, Bonneau R, Zhang N, Zhou Y, Wollscheid B, Cooke K, Yi EC, Lee H, Peskind ER, Zhang J, Smith RD, Aebersold R (2006) UniPep, a database for human N-linked glycosites: a resource for biomarker discovery. Genome Biol 7:R73

6 Proteome Imaging

Patricia M. Palagi, Daniel Walther, Catherine G. Zimmermann-Ivol, and Ron D. Appel

Abstract

The field of proteomics has benefited enormously from the use of images, in particular those resulting from the analysis of samples by two-dimensional electrophoresis. Proteome imaging has the same function as a microscope: it allows the resolution of features and constituents of an object, in this case a proteome, which are otherwise invisible to the naked eye. Functionally, proteome imaging also allows differences between proteomic samples to be highlighted. The four major areas of proteome imaging, related to the underlying analytical proteomic technologies, are discussed in this chapter: two-dimensional electrophoresis gel images, images of the combination of liquid chromatography and mass spectrometry, images from the 'molecular scanner' and mass spectrometry tissue images. Their common goal is to display proteomes in a form that is amenable to human vision and computer analysis, which can subsequently facilitate the comparison of two or more samples and assist protein identification. The background of the imaging procedures both in terms of the experimental requirements and the available software and tools is described, as well as the applications and use of proteome imaging for real proteomic studies.

6.1 Introduction

Images are critical in the biomedical sciences. They can be the key to patient diagnosis and can assist in the understanding of voluminous and complex data. The field of proteomics has also benefited enormously from the use of images, in particular those resulting from the analysis of samples by two-dimensional electrophoresis (2-DE). This technique, one of the first employed in the field of proteomics (see Chap. 2), involves the separation of proteins according to their isoelectric point and molecular weight, and the staining of the resulting 2-DE gel to reveal constituent proteins. The results are captured and analysed as an image, revealing the subsample of a proteome. As a consequence, we can state that proteome imaging began with the 2-DE gel and can be defined as the visualisation and analysis of proteomic

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

data in the form of a multidimensional image or plot.[1] Proteome imaging has the same function as a microscope: it allows the resolution of features and constituents of an object, in this case a proteome, which are otherwise invisible to the naked eye. Functionally, proteome imaging also allows differences between proteomic samples to be highlighted.

New techniques for the study of proteomes, such as mass spectrometry (MS) and liquid chromatography (LC), have expanded the scope of proteome imaging. In fact, proteome imaging depends on the underlying analytical proteomic technologies, which explains the contents of this chapter. Four major areas of proteome imaging are discussed here. These are 2-DE gel images, images of the combination of LC and MS, images from the 'molecular scanner' and MS tissue images. Their common goal is to display proteomes in a form that is amenable to human vision and computer analysis, which can subsequently facilitate the comparison of two or more samples and assist protein identification. Each section describes the background of the imaging procedures, both in terms of the experimental requirements, and the available software and tools. Each section also describes the applications and use of proteome imaging for real proteomic studies.

6.2 Image Analysis of Two-Dimensional Electrophoresis Gels

Since 1975, 2-DE gels have been used to study protein-expression patterns in a large variety of biological and biomedical samples. The necessity of analysing images of 2-DE gels with the aid of computers was immediately recognised. Both the high resolution of the technique – thousands of protein spots can be resolved from one sample on each 2-D gel – and the capacity to analyse multiple samples at the same time makes proteome research difficult without the use of software tools.

A first generation of tools and software for the analysis of 2-DE gels was developed in the early 1980s. For the most part, they made use of mainframes, minicomputers and software environments that had poor programmable graphical interfaces or none at all. Pioneering representatives of this generation included GELLAB (Lemkin and Lipkin 1981), TYCHO (Anderson et al. 1981), LIPS (Skolnick 1982) and Elsie (Vo et al. 1981).

With the advent of graphical interfaces and windowing systems, linked to the availability of reasonably priced workstations, a new wave of software packages for 2-DE analysis was designed. These mostly ran under the UNIX operating system. The second-generation of software packages was composed of Elsie-4 (based on Elsie; Olson and Miller 1988), QUEST (Garrels 1989), Gellab-II (based on Gellab; Lemkin and Lester 1989) and Melanie (based on Elsie-4; Appel et al.

[1]It is worth noting that we use the term 'proteome imaging' when referring to the analysis of two-dimensional or greater than two-dimensional data. This excludes one-dimensional data such as single mass spectra.

1991). Their major drawback was a lack of a friendly interface and, as a consequence, the difficulties that non-computer scientists faced in using them. Some of this software evolved into third-generation packages which were turned into commercial products such as PDQuest (based on QUEST), Gellab-II+ (based on Gellab-II) and Melanie II (based on Melanie; Appel et al. 1997a). This third generation was facilitated by a generalised use of well-designed graphical interfaces, the availability of powerful low-cost personal workstations, the emergence of the World Wide Web and the wide adoption of object-oriented programming. They shared the properties of modularity, ease of use and availability on more than one platform, including for UNIX, the Macintosh and the PC. In the last few years, this generation of software has undergone continuous improvement to meet the needs of the small laboratories that might run a few gels per year as well as the needs of the high-throughput facilities. Their algorithms have been optimised to improve key steps in gel analysis, and their interfaces have been reshaped to be more user-friendly. Table 6.1 lists some of the third-generation commercially available software for 2-D image analysis, as well as other software that has been more recently developed and released. Although each software package listed in Table 6.1 has its own philosophy and approach, they all provide the basic functionalities and operations necessary to carry out a complete study of 2-D gel images. This includes protein spot detection and quantitation, gel-to-gel image matching to find spots that correspond between gels, and data analysis to find differentially expressed spots in a population of gels. These functions will be explained in the following paragraphs and illustrated with the package ImageMaster 2D Platinum release 6. This package is underpinned by Melanie, developed by the Proteome Informatics Group at the Swiss Institute of Bioinformatics.

6.2.1 First Steps in Gel Image Analysis

Before the computational analysis of 2-D gels can begin, a digital image must first be created from the gel itself. This is achieved by the use of flatbed

Table 6.1 Major commercial software for 2-D image analysis

Software	Company	Source Web site
DeCyder	GE Healthcare	http://www.gehealthcare.com
Delta2D[a]	Decodon	http://www.decodon.com
ImageMaster™ 2D Platinum	GE Healthcare	http://www.gehealthcare.com
PDQuest[a]	Bio-Rad	http://www.bio-rad.com
Progenesis/Phoretix	Nonlinear Dynamics	http://www.nonlinear.com
Proteomweaver	Definiens	http://www.definiens.com

[a] All packages run under Windows. Delta2D also runs under Linux, Sun Solaris and Mac OSX. PDQuest also runs under Mac OS.

document scanners, camera systems, densitometers, phosphor imagers or fluorescence scanners. The scanning resolution is an important factor since it affects the amount of detail on the image and, as a consequence, the whole analysis process. A low-resolution image containing a small number of dots per centimetre, expressed as dots per inch, makes spot detection less precise. A high-resolution image, however, will contain a large number of dots per inch and produce a large image file. This can slow down image analysis. An image with 250 dots per inch or so is a commonly used compromise. Scanners produce images in TIFF or other common formats, with a depth of 12 or 16 bits. This provides a range of 4,096 or 65,536 grey levels, respectively. Once the gels have been scanned, their images can then be opened by the 2-DE software for computational analysis. The 2-DE analysis software has a variety of ways of displaying and viewing gel images, assisting the user in image navigation and manipulation. As an example, Fig. 6.1 gives some different viewpoints of the same gel image.

The initial step in the analysis of 2-DE gel images is spot detection. This is done by defining the spot boundary, segmenting the spot contours from the background and quantifying each spot by assigning values such as intensity, area and volume. Methods can simulate the physicochemical properties of protein spots, by nature of a 2-D Gaussian function or a diffusion model, or

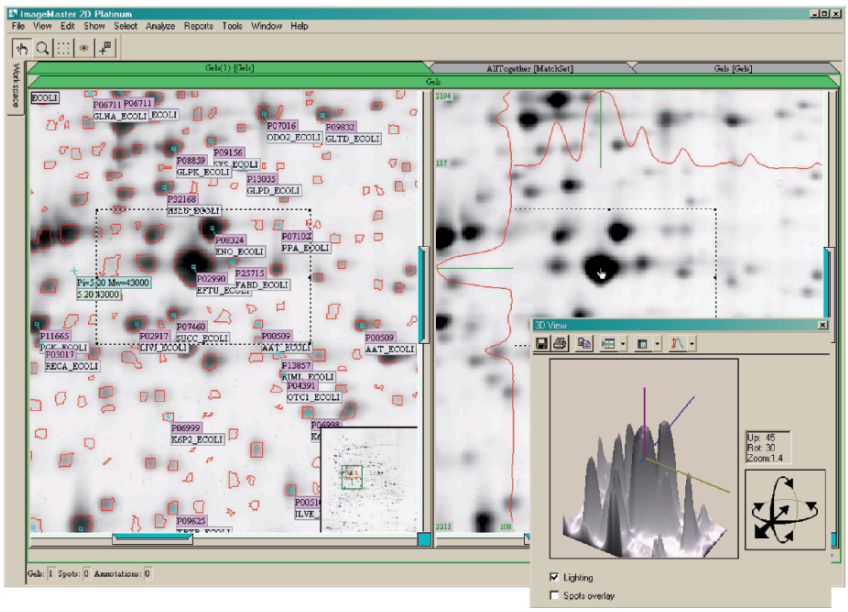


Fig. 6.1 *Escherichia coli* reference map from the SWISS-2DPAGE database (*left*) compared with a different strain of the same bacteria (*right*). The 3-D window (*bottom right*) highlights one region from the reference map

by using image-processing algorithms such as a watershed transformation (Pleissner et al. 1999). These algorithms are used to solve a number of difficult issues which are intrinsic to the experimental 2-DE method. These include overlapped or close neighbouring spots, vertical or horizontal streaks, faint or oversaturated spots and very complex regions with a high number of spots. A comprehensive review of the existing algorithms for spot detection is given by Dowsey et al. (2003). A good spot detection result is a prerequisite for a good quantitation and, consequently, good 2-D gel image analysis.

After protein spot detection, gel images are often matched to each other. The basic gel-matching process compares two gel images to find spots which represent the same protein in both of the images. Currently, two main algorithmic principles are used to match gel images: the matching of spots that have been previously detected (Appel et al. 1997b) or the matching of whole gel images based on their intensity distributions but without previous detection of spots (Smilansky 2001).

The former is more robust when matching images with large distortions or very different expression patterns. The latter, because of its rapidity and ease of use, is more appropriate when gel images are very similar. In any case, preassigning corresponding spot or pixel pairs (also called landmarks or anchors) may improve the efficiency of a matching algorithm. Irrespective of the matching method used by the software, it is essential to carefully evaluate the match results. It is hard to evaluate matching results and thus to know the accuracy of matching methods, and hence manual correction of matching results may be necessary.

After the matching of gel images, most researchers seek to find proteins that are expressed at different levels in the different gels. This is usually carried out with statistical tests. In the case where the classification of each sample is known, image-analysis packages apply common statistical tests such as Student's *t* test, Wilcoxon and Kolmogorov. For more sophisticated analyses, as in the case of blind studies where the gels are not separated into classes, image-analysis packages offer other approaches, such as clustering methods and factorial or principal component analysis. In fact, a blind analysis allows gels to be clustered into groups, for the most significantly expressed proteins to be discovered and for unforeseen differences in samples to be found. It can give new insights on possible subpopulation arrangements and their protein expression.

6.2.2 Applications to Different Proteomics Approaches

Experimental studies have different objectives, and require different approaches for extracting relevant information from gels with image-analysis software. The approaches differ in the number of 2-DE gel images to be analysed and the kind of comparisons that are performed.

6.2.2.1 Single-Gel Analysis

Research groups interested in discovering all or as many as possible proteins expressed in a 2-D gel use image-analysis software to visualise and annotate the gel image. Typically, the sample has not been previously studied with the same 2-DE method and does not have a corresponding reference map available in a 2-DE database (see Chap. 7). Once the proteins of the gel have been identified, most often with MS, the software enables annotation of the gel image with the corresponding protein names and other related data. The resulting protein map can be integrated into a public 2-DE database as a reference map and hence will be available to other research groups. Figure 6.1 is an example of the 2-D reference map of *Escherichia coli* (Tonella et al. 1998) available in the SWISS-2DPAGE database (Hoogland et al. 2004). This figure also illustrates the comparison of the reference map of *E. coli* with another 2-D gel obtained from another strain of *E. coli*. In this particular case, the two gels are from comparable samples and have been subjected to the same experimental procedure.

6.2.2.2 Groups of Gels

The most common objective in the analysis of groups of 2-DE gels is to find variations in protein-expression profiles between biological samples from different conditions. These conditions might include disease versus control, different time points, drug-treated versus control, or different individuals in a population. The variations in protein expression are recognised visually as spots that appear or disappear (indicating that the protein is unique to a gel state), as differences in spot intensity (indicating upregulated or downregulated proteins), or as shifts in a spot's position (indicating possible post-translational modifications, alternate splicing or polymorphisms). Gel image analysis software is required to locate these variations precisely, and to focus MS efforts on protein identification and characterisation. Figure 6.2 is an example of a differential analysis between two groups of gels. Each group has three replicate gels. Control gels belong to one group (top row), while gels from drug-treated samples are in a second group (bottom row). Notably, this figure shows that two neighbouring spots present in the control sample are absent from the drug-treated sample. Either these proteins are not expressed in the drug-treated samples or they have undergone post-translational modifications, causing them to migrate to another position on the 2-D gel.

6.2.2.3 Two-Dimensional Difference Gel Electrophoresis

The 2-D difference gel electrophoresis (DIGE) technique, described in Chap. 4, has changed the landscape for the computational analysis of 2-DE gels. In this

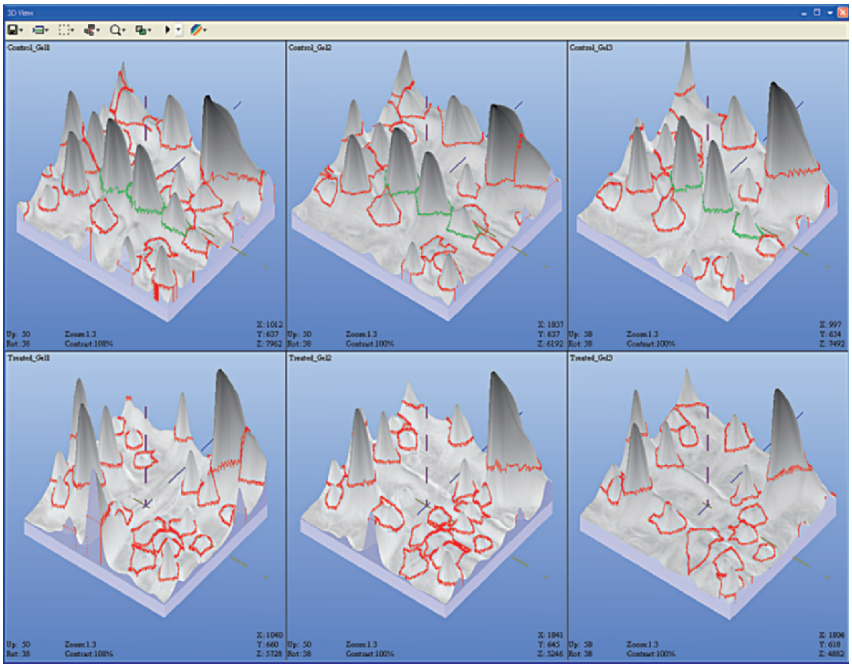


Fig. 6.2 Image analysis used to compare a group of gels. The *top row* shows a 3-D view of a region on three control gels, and the *bottom row* shows the corresponding regions on three treated gels. A pair of spots present in the control gels (*centre*) is not seen in the treated gel

method, up to three samples are each labelled with one spectrally resolvable fluorescent dye (Cy2, Cy3 or Cy5), then mixed together and co-migrated in a single gel. The 2-D gel is then scanned with lasers of different wavelengths, allowing the detection of each of the different samples in the same gel. The resulting images can be compared using image-analysis software. DIGE helps reduce variations due to gel-specific experimental factors, as up to three samples are on the same gel. Identical proteins will be localised to the same position on the gels, making the spot-detection procedure easier and avoiding possible errors in this step and in the matching step. Furthermore, quantitation is more accurate since an internal common standard can be included during migration. The internal standard is created, for example, by pooling aliquots of all biological samples in the experiment, labelling with one of the dyes and running together with the individual samples. It is an average image of all samples, with all proteins of the experiment represented.

The comparison of groups of DIGE gel images is, in practice, very similar to the comparison of groups of non-DIGE gels; however, specific algorithms are required for the spot detection and matching procedures. Most of the software packages listed in Table 6.1 analyse this kind of fluorescent gel

image. Like for non-DIGE gels, they use statistics to localise differentially expressed proteins and have other statistical tools, such as clustering, for exploratory data analysis. Applications of such software range from the discovery of biomarkers of pancreatic cancer (Yu et al. 2005) to the quantitative analysis of bacterial proteins (Gade et al. 2005).

6.3 Liquid Chromatography–Mass Spectrometry

The intensive use of MS in proteomics has opened a new domain in proteome imaging. The representation of LC-MS datasets as a 2-D plot can highlight nuances of data not necessarily seen when displayed as 1-D chromatographs or spectra. In a LC-MS image, the axes represent the retention time derived from LC and the mass-to-charge (m/z) values derived from MS, while the grey levels of the image describe the MS signal intensities.

The idea of displaying LC-MS data as an image is quite recent (Berger et al. 2002; Palmblad et al. 2002), and so is the software applied to it. The known tools are divided into two categories: those commercialised by life-science companies (such as Decyder MS from GE Healthcare) and those developed by academic institutes, such as Pep3D (Li et al. 2004) and Mzmine (Katajamaa and Oresic 2005). MSight (Palagi et al. 2005), used here to illustrate the applications of LC-MS images, falls between these two categories. It is a package underpinned by Melanie, the commercial software described in Sect. 6.2, and it is developed by the Proteome Informatics Group at the Swiss Institute of Bioinformatics, an academic institute. The current MSight version (1.0) is freely accessible through the ExPASy server (MSight 2007) and it allows the visualisation of MS runs. At the time of writing this chapter, MSight release 2.0 alpha, which performs matching and comparison of MS runs, was freely available on the ExPASy server; however, it may be commercialised at any time.

6.3.1 First Steps in Liquid Chromatography–Mass Spectrometry Image Analysis

The reliability and utility of comparative protein-expression profiling with LC-MS images relies on accurate measurements of the relative abundance of the proteins or peptides present in the sample. To achieve accurate measurements, the software has to address issues such as automated peak detection, peak alignment and matching, and finally peptide quantitation. These are similar issues to those faced in the use of 2-D gel image analysis software.

The first step when dealing with LC-MS data is to open the data files. MSight, as well as the other image-analysis applications, accepts data generated from the majority of mass spectrometers from Bruker, Waters or ABI-SCIEX. All the image-analysis applications support the mzXML format (Pedrioli et al. 2004),

and other export and conversion filters are also available. The Human Proteome Organisation's Proteomics Standards Initiative has been working on a generalised standard representation of MS data, named mzData (Orchard et al. 2005) and most of the LC-MS software supports this format as well.

If the data are noisy, filtering can be used in the LC-MS data processing. This step removes the peaks with weakest intensities, which correspond to background chemical noise, thus reducing the complexity of spectra and facilitating peak detection. The peak-detection step itself then looks for the monoisotopic peaks by deisotoping, and determines ion charge states. Finally, it clusters the isotopic peaks of the same corresponding mass value into one single peak signal. This procedure keeps only the peaks of interest from the enormous quantity of data.

Ideally, identical proteins or peptides analysed in the same LC-MS platform should have the same retention time, molecular weight and signal intensity; however, owing to experimental variations, this is not usually the case. While m/z values depend on mass accuracy and the resolution of the mass spectro-meter, the retention times depend entirely on the chromatographic method used. Peaks from the same peptide or protein can match closely in m/z values, but the retention times between the runs can vary significantly. A peak-alignment step corrects these variations and finds corresponding peaks across different LC-MS runs. Once the runs have been aligned, peak intensities can be quantified and can be compared and statistically analysed in order to find proteins and peptides that show significant differences in their expression levels.

6.3.2 Applications to Different Proteomics Approaches

One of the main uses of LC-MS imaging software is data visualisation and navigation. Different views and zooms are possible. For example, on the left of Fig. 6.3 four LC-MS images are shown, each with a different zoom factor. Also shown is a 1-D view of a number of superimposed spectra (bottom right), a 3-D view of a selected region of a LC-MS run (above the spectra) and a workspace window used to manage different images and projects. It is convenient to have a global view of the sum of all spectra, using the redundancy of LC and MS sampling rates, and to be able to jump easily to selected isotopic peaks. Displaying the same data in different modes can help to distinguish isotopic peaks of the same peptide or even different overlapping peptides, something which would not be possible when looking at a single spectrum at a time.

6.3.2.1 Monitoring Experiments and Post-translational Modifications

LC-MS image software is particularly useful for monitoring problems with samples and experiments, including sample contamination with undesired proteins or artefacts on the chromatography column. The software and

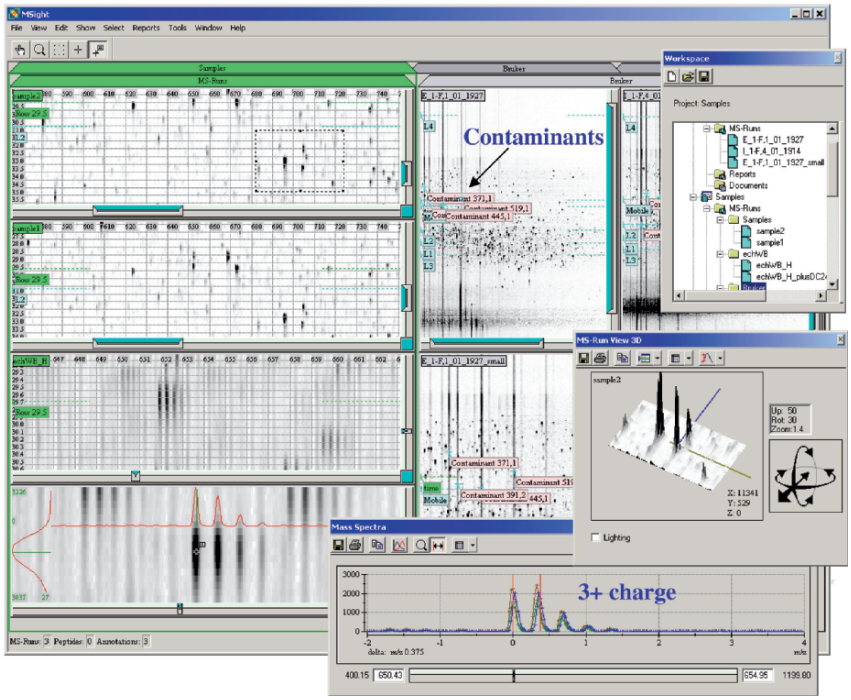


Fig. 6.3 Global view of liquid chromatography (LC)–mass spectrometry (MS) analyses with MSight

images are also useful to monitor the presence of mass calibrators deliberately included in the sample so as to quantitatively or qualitatively measure other signals. In Fig. 6.3, examples of column contaminants are seen as vertical stripes in the images. Of great interest for this type of proteome imaging is the capacity to find post-translational modifications of peptides. Visual analysis of single spectra and 2-D images helps to find and interpret small molecular variations such as the oxidation or glycosylation of peptides. Figure 6.4 illustrates the visual interpretation of oxidised peptides, where the data were generated from a 2-DE gel spot digested with trypsin. LC-MS-MS image analysis complements the study by providing an interpretable global view. The precursor ions selected for MS-MS analysis, the names of the proteins identified, the ion sequences and their charges may be linked to the LC-MS images with annotations such as the one shown in Fig. 6.5.

6.3.2.2 Sample Populations

The ultimate objective of LC-MS imaging software is to compare two or more analyses, find differences between them and to quantify these variations. This

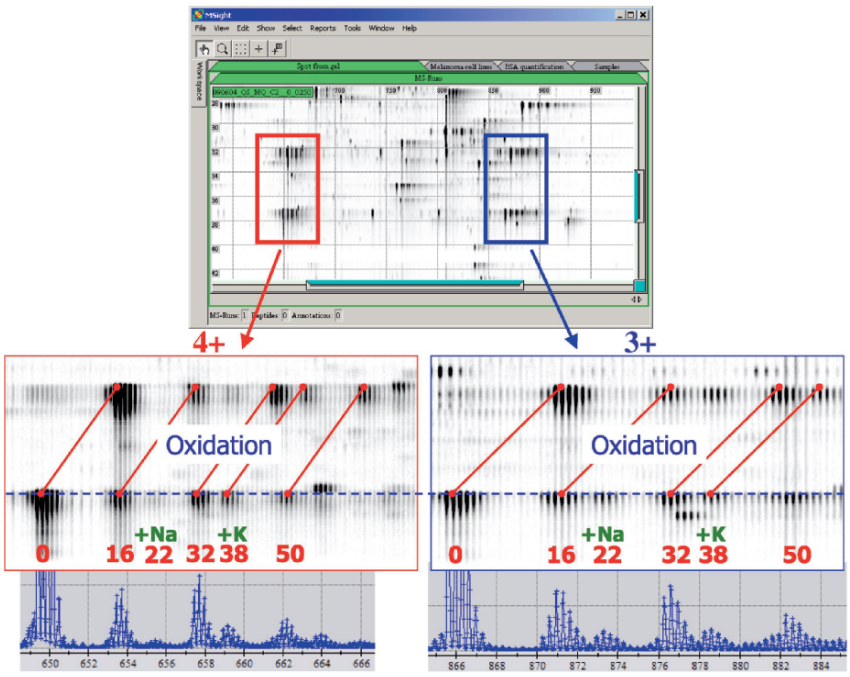


Fig. 6.4 Analysis of a post-translational modification. Oxidation plus adducts of sodium and potassium are shown

is similar in concept to the comparison of multiple 2-D gel images. Once peaks have been detected and matched across runs, LC-MS images can be analysed automatically. In this respect, LC-MS image analysis represents a promising alternative to isotope-based quantitative proteomics, as proteins that are present at different amounts in the compared samples can be seen without using any isotopic labelling method. Figure 6.5 illustrates this case when sample compositions are known and specific proteins are tracked between LC-MS runs. Samples were obtained from a 32–45-kDa fraction of lysate from a culture of a B-cell line (W. Bienvenut and M. Quadroni, personal communication). They were digested with trypsin and separated by reversed-phase capillary LC coupled to a SCIEX/Applied Biosystems QSTAR quadrupole time-of-flight (TOF) mass spectrometer equipped with an electrospray ionisation source. In the LC-MS image (Fig. 6.5, top row), the total amount of protein analysed was about 1 pmol. In the following rows, 26, 83 and 520 fmol of trypsin-digested bovine serum albumin (BSA) were, respectively, added to sample. The size of the isotope spots of one of the BSA peptides (identified by LC-MS-MS as being LGEYGFQNALIVR) gradually increases as the protein concentration increases.

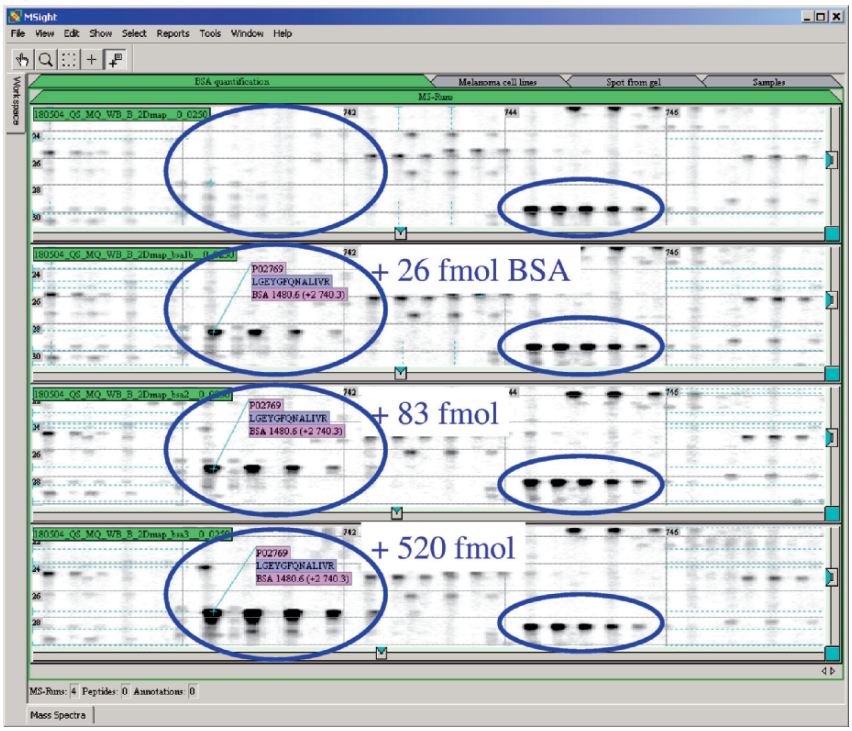


Fig. 6.5 Quantitation with LC-MS imaging. A peptide from bovine serum albumin is absent in the first run (*top row*, *left*). The size of the isotope spots increase with changes in peptide concentration. On the *right*, *circled peptides* do not change from run to run, indicating the equal presence of another peptide

Zhang et al. (2005) provided a further example of this. Serum protein N-linked glycosylated peptides analysed via LC-MS imaging were used to discriminate untreated normal mice from genetically identical mice with carcinogen-induced skin cancer. In this study, the Pep3D software was used to visualise the images, and identifications and companion software were used to make an unsupervised hierarchical clustering of peptides to distinguish cancer samples from normal samples.

6.4 The Molecular Scanner

The development of the software and tools presented so far was driven by data, specifically that of 2-DE gel images and spectra from LC-MS. Novel analytical approaches are also available that combine electrophoresis with direct matrix-assisted laser desorption/ionisation (MALDI) TOF MS analysis. These

also produce thousands of spectra at a time and are highly dependent on proteome imaging. To overcome both automation limitations and the long processes required to combine gels and MS techniques, and to help interpret results, Hochstrasser and co-workers (Bienvenut et al. 1999) created a technology called the molecular scanner. This molecular scanner allows the visualisation and classification of biological samples at the molecular level, while maximising the benefit of parallel protein transfer, protein profile imaging and MS protein identification. In this technique, proteins are separated either by 1-D sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis (PAGE) (separation by molecular weight), 1-D isoelectric focusing (separation by isoelectric point) or 2-D PAGE. They are then digested in parallel during a transblotting process, called one-step digestion transfer, through a membrane containing trypsin immobilised on a poly(vinylidene difluoride) capture membrane (Bienvenut et al. 1999). A thin layer of MALDI ionising matrix solution is then deposited on the whole capture membrane and, after drying, a film of gold is deposited onto the membrane by anodic vaporisation to reduce charge buildup on the non-conductive membrane during MS (Scherl et al. 2005). The capture membrane is then completely scanned with a MALDI TOF-TOF mass spectrometer. Spectra are automatically acquired from each position on the membrane (one spectrum per *x*/*y* coordinate) and all resulting peptide mass and fragment data are used for protein identification to analyse the complete sample. Automatic identification is performed by protein-identification software tools such as Aldente (Tuloup et al. 2003) and ChemApplex (Parker 2002) for peptide mass fingerprinting, and Phenyx (2007) for peptide fragment fingerprinting. Imaging software is finally used to reconstruct the image of the original gel by using spectral-intensity and protein-identification results. At the end of the whole process, a fully annotated protein map image is created.

Images from the molecular scanner provide a simple way to navigate through the large volumes of data generated by the scanning process. Images of different samples can also be compared in a semiautomated differential proteome analysis. Three tools have been created to help analyse these images, each for specific experimental purposes and functions:

1. *The molecular scanner imaging tool.* Software has been developed at the Swiss Institute of Bioinformatics to identify and visualise proteins after their separation by 2-DE and analysis with the molecular scanner. The software achieves this by querying external databases to identify proteins and then uses spectral and identification data to reconstruct annotated gel-like images. To date, it has been applied to the analysis of proteins from *E. coli* (Müller et al. 2002a) and proteins from human plasma (Müller et al. 2002b). The challenge of reconstructing the gel-like images was in clustering peptide masses according to the similarity of the spatial distributions of their signal intensities and the position of any neighbouring proteins. The software produces different views of the data, as illustrated

in Fig. 6.6. Thanks to this visualisation tool, it is possible to confirm observations such as the localisation of identical peptides masses in several different spots and the existence of single spots that contain more than one protein. Besides this, many low-abundance proteins can be identified and many of the false positives associated with peptide mass fingerprinting identification can be discarded.

2. *BioMap*. This tool was developed by Stoeckli et al. (2002) to generate MS image maps for any mass-to-charge ratio and to view expression profiles for scans of tissue sections. It allows the visualisation and distribution of one selected peptide on a whole region with its intensity information, the mass spectra and annotations. Figure 6.7 illustrates this with sections of mouse brain tissue, showing the distribution of a large number of peptides according to their localisation (Rohner et al. 2005). As two peptides with the same mass but different sequence may be found in the same place, the identification of proteins by MS-MS is required for unambiguous protein identification. The technique of locating compounds directly in tissue sections as well as the use of BioMap will be discussed in Sect. 6.5.
3. *MSight*. The MSight software, described in Sect. 6.3, can also be used to generate *m*/*z* data images from the molecular scanner process. However, instead of using the LC retention time as a coordinate, it uses molecular-weight data from 1-D SDS-PAGE gels or protein pI estimates from a 1-D isoelectric focusing gel. The major advantage of MSight is that it provides

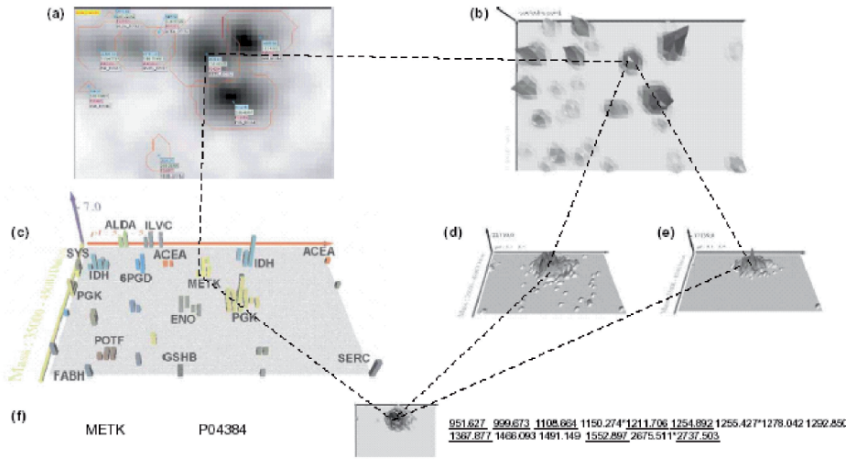


Fig. 6.6 Protein visualisation and identification from an *E. coli* 2-D gel following molecular scanner analysis. The area represented corresponds to 1,536 MS spectra measured on a surface of about 1.2 cm^2 from a mini-2-D gel. *a* The total MS intensity image, *b* the smoothed spot centre images, *c* protein positioning, *d*, *e* the positions and MS intensities of the *m*/*z* values 951.627 and 1,108.664 Da, respectively, on the whole surface, and *f* one example of an identified protein METK *S*-adenosylmethionine synthetase. (Reprinted from Binz et al. 2004 with permission from Elsevier)

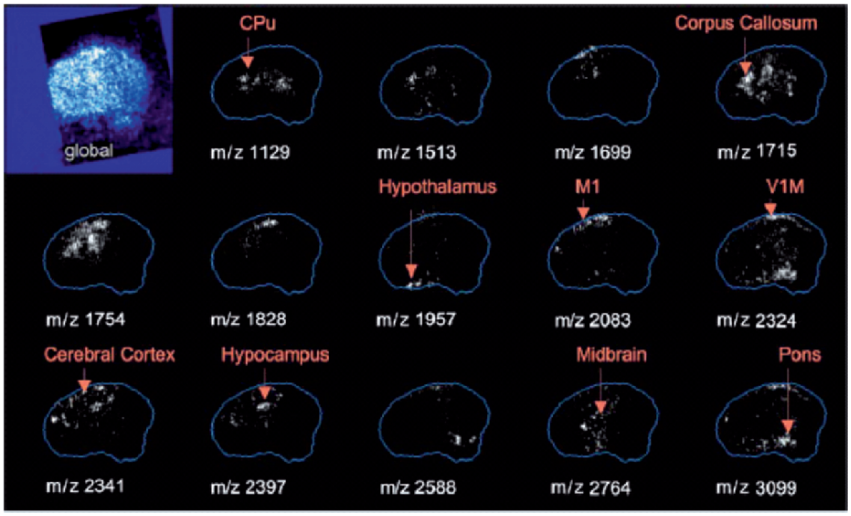


Fig. 6.7 A BioMap® image of brain tissue sections following the molecular scanner process. Visualisation of specific peptides in different brain loci. *CPu* caudate putamen (striatum), *M1* primary motor cortex, *V1M* primary visual cortex, monocular region. (Reprinted from Rohner et al. 2005 with permission from Elsevier)

an easy way to monitor technical issues such as the quality of a separation, the homogeneity of matrix deposition and any problems with mass calibration. These are verified by the presence of an internal standard deposited along with the sample and/or over the entire membrane. For example, with the help of MSight, matrix deposition could be controlled to avoid peptide diffusion (C.G Zimmerman, unpublished results). Figure 6.8 shows a composite image of MS data obtained from membrane proteins of *Staphylococcus aureus* strain N315 separated by 1-D SDS-PAGE and analysed by the molecular scanner. It shows the 1-D SDS-PAGE gel, the MSight image, the BioMap image and the corresponding protein identifications made with Aldente.

The tools described above were developed to manage large amounts of peptide mass fingerprinting data. However, the complexity of the protein complement of biological samples requires information from MS-MS for unambiguous protein identification. Ideally, this should be included in these tools soon. Further developments of such tools might include quantitative information generated either by MS or by MS-MS analysis, as well as the automatic annotation of protein properties such as post-translational modifications.

Regarding the molecular scanner technique per se, its major advantages are the avoidance of sequential excision of hundreds of bands or spots from 1-DE or 2-DE gels and a dramatic decrease in the time required to establish a

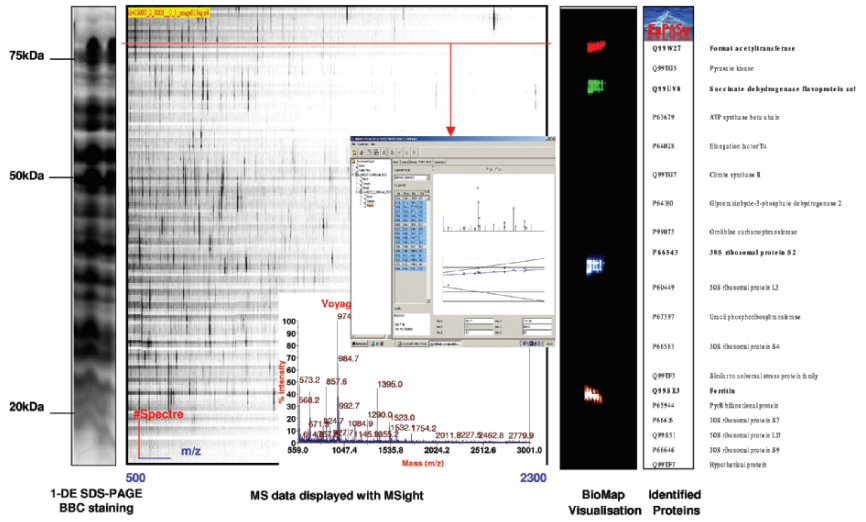


Fig. 6.8 Visualisation of membrane protein profiling of *Staphylococcus aureus* N315 strain after the molecular scanner process. On the *left*, the 1-D sodium dodecyl sulfate polyacrylamide gel electrophoresis (*SDS-PAGE*) gel, in the *middle*, the MSight image, a spectrum and the Aldente identification window, and on the *right*, the BioMap image and identification of the proteins

complete protein map of one biological sample (Nadler et al. 2004). On the other hand, one of its major drawbacks is that sensitivity is currently at the picomolar level owing to unresolved problems with the capture membrane. Protein profiling by the molecular scanner is clearly dependent both on wet laboratory technologies used for protein separation and on image-visualisation tools. The molecular scanner method generates relevant information complementary to that obtained from other imaging methods and can be applied to a range of biological samples. Other techniques have been developed to create protein profiling along with the identification and localisation of proteins, such as the imaging MS (IMS) approach, described next.

6.5 Imaging Mass Spectrometry

Medical imaging techniques, such as computer tomography and magnetic resonance imaging, play an important role in the study of anatomical, physiological and functional information; however, these images do not display protein distributions. Histochemical methods provide 2-D maps of protein distribution by way of labels, affinity tags or expression markers, but they are limited to the simultaneous analysis of only a few known proteins at a time. An alternative method has been developed by Caprioli and Stoeckli (Stoeckli

et al. 2001; Caprioli et al. 1997) for MALDI IMS on tissue sections to image protein distribution with the help of high-resolution MS.

6.5.1 Imaging Mass Spectrometry – Technical Aspects

In a typical IMS procedure, fresh frozen tissue sections are cut with a cryostat and then mounted either on indium tin oxide coated conductive glass slides (Galicia et al. 2002) or on metallic surfaces such as MALDI plates (Stoeckli et al. 2001). Glass slides are used to perform an optical histology evaluation and, in parallel, protein analysis. The section is then coated with a thin layer of MALDI ionising matrix solution and the target plate is inserted into a MALDI MS analyser. The MALDI process on tissue sections demands a relatively high laser pulsed energy compared with conventional MS analysis, where the proteins are already extracted. Among the available MS techniques, three are currently applied to biological experiments to improve tissue imaging (Heeren 2005). They differ mainly in the scanning resolution:

- Laser microprobe mass analyser *(LAMMA) or microprobe imaging*. The microprobe mode of MALDI (also called LAMMA) uses a microfocused UV laser beam to fire at a single, defined spot on the sample (Hillenkamp and Karas 1990; Spengler and Hubert 2002). The resulting mass spectrum is stored along the spatial coordinates of the spot. A new region is then illuminated and another mass spectrum recorded. This process is repeated until the entire sample area has been examined and mass spectra associated with each specific location have been obtained. As a result, the resolution is determined by the spot size. In a well-optimised MALDI system, this can be down to 25 µm (Todd et al. 2001). Microprobe MALDI is the IMS reference method; however, other more specialised methods exist, such as secondary ion MS (SIMS) and microscope imaging, which both improve spatial resolution.
- *SIMS*. This uses an energetic atomic ion beam to generate secondary ions from surfaces that are subsequently analysed by MS. The association of MALDI and SIMS aims to combine high mass fragmentation and high spatial resolution (Todd et al. 2001). While LAMMA commonly generates ions with a maximum of 100 kDa, and typical pixel sizes greater than 25 µm, SIMS routinely generates images with less than 1-µm spatial resolution but has low sensitivity for high masses. A number of limitations still exist, however. For instance, the final resolution is governed by the size of the matrix crystals, which is a consequence of the crucial matrix deposition step. If crystals are large, the spatial resolution is low and it provides high intensity signals (McDonnell et al. 2003). Conversely, when crystals are small, the spatial resolution is high and the signals can be weak. The deposition of a thin layer of gold seems to enhance molecular ion yields, making a good compromise between resolution and crystal size. Maarten Altelaar et al.

(2005) and Stoeckli et al. (2005) obtained some preliminary results with gold deposition on brain tissue sections to follow drug distribution.
- *Microscope imaging*. The speed at which images of a few square centimetres are generated is still low with both LAMMA and SIMS imaging. The images can take up to 24 h to be acquired since the laser beam is fired at a single, defined spot on the sample at a time. The microscope imaging approach reduces this problem by generating a whole picture at a time (much like a photograph) for each *m*/*z* ratio. In this way a mass-to-charge ratio (*m*/*z*) separated series of molecular images is generated, showing spatial detail from within the laser spot. Thus, the spatial resolution is only limited by the quality of the ion optics (laser beam) and the spatial resolution of the detector, and not by the size of the laser beam. The microscope mode offers a resolution of up to 2 μm (Luxembourg et al. 2004).

6.5.2 Imaging Mass Spectrometry – Applications

IMS is most commonly used to evaluate the distribution of proteins in a section of tissue. One major inconvenience is the necessity to know the target protein which is to be visualised, since this analysis is based on intact proteins or small molecules (such as drugs and metabolites). Thus, the research for a biomarker by IMS needs the preliminary identification and validation of the target by other conventional methods. A major advantage is to make the visualisation of mass analytes possible by keeping spatial information without labelling the protein. Among the recent and exciting applications of IMS we can find:

- *Drug tracking*. Known drug compounds can be directly located in tissue sections through conventional autoradiography and correlated with the distribution of a radiolabelled target. IMS has the advantage that it may be used to investigate the cellular distribution of well-established and prospective drugs without labelling of a protein (Rubakhin et al. 2005). The advantage of IMS is to be able to track, in a single experiment, the distribution of mass values of the drug and its metabolites, and it can thus be involved in many steps of drug discovery.
- *Displaying morphological information*. Linking protein-distribution images and histological images is important to ensure the medical accuracy of IMS information. Chaurand et al. (2004) investigated a method to stain histological tissue sections and perform direct analysis by IMS. This staining procedure helped to track protein changes in morphological structures and simultaneously avoided the constraint of aligning consecutive stained sections with an original analysed sample. In the study of Chaurand et al., over 400 different proteins were monitored on the epididymis of a sexually mature mouse and several tissue-specific proteins were observed in the caput and in the corpus tissues. In another large-scale imaging analysis, IMS was used to scan a whole mouse body section where not only known

parent compounds were discriminated but also their metabolites (Rohner et al. 2005).

- *Image classification*. Large numbers of IMS images, after generation, can be analysed and classified. For example, images from a large population study on 20 brain tumour biopsies have been classified by cluster analysis (Schwartz et al. 2004). This shows the potential ability of IMS to differentiate morphological differences among groups. This study demonstrated the protein-profiling similarity within two white matter brain regions as well as the pattern distinction between white matter and cortex regions. It also demonstrated the ability of IMS to discriminate between different histological grades of glioma (non-tumour, grade II and grade III tumours). Moreover, in this study, a tumour of the neural crest region was found to be different from all other tumours and from normal brain tissue, showing the robustness of IMS to identify tumours of unknown origin or to distinguish between two tumours. In another study, 80 lung tumour samples were classified (Yanagisawa et al. 2003).

IMS is in an exponential phase of development. It is already showing great promise, and chances are high that it will be used in the future in clinical studies and in biology. With high-resolution mass spectrometers, the analysis of single cells obtained by laser capture microdissection to localise proteins/peptides in different cell compartments will soon be feasible, showing the real applicability of IMS to molecular imaging.

6.6 Conclusion

Proteomic technologies have great potential for making significant contributions to the fields of biology and medicine. Whereas most current medical imaging technologies, such as magnetic resonance imaging and computed tomography, rely on identifying anatomical relationships between different organs and tissues, the future will most likely rely on molecular or cellular imaging. Proteome imaging might have the greatest impact on the identification of new drug targets and new prospective biomarkers for medical diagnostics and prognostics. 2-DE gels, LC-MS, the molecular scanner and IMS will be key technologies for these molecular assessments, and the capacity to visualise their data through images will certainly consolidate their role. Their images provide complementary information to that obtained from conventional medical imaging by avoiding the use of biochemical targets such as antibodies or labelling. So far, the software and tools created to manipulate proteome images have confirmed the capacity of the intrinsic techniques to identify proteins, as well as to closely study and compare proteomes. As experimental techniques evolve and stabilise, the better these tools are adapted, the better the representation of proteomic data as multidimensional images will be.

Acknowledgements. The authors would like to acknowledge Willy Bienvenut (The Beatson Institute for Cancer Research, Glasgow, UK) and Manfredo Quadroni (Protein Analysis Facility, Lausanne University, Switzerland) for providing the LC-MS data used to generate Figs. 6.3–6.5. We would also like to acknowledge the Biomedical Proteomics Research Group for the gels shown in Fig. 6.1.

References

Anderson NL, Taylor J, Scandora AE, Coulter BP, Anderson NG (1981) The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. Clin Chem 27:1807–1820

Appel RD, Hochstrasser DF, Funk M, Vargas JR, Pellegrini C, Muller AF, Scherrer JR (1991) The MELANIE project: from a biopsy to automatic protein map interpretation by computer. Electrophoresis 12:722–735

Appel RD, Palagi PM, Walther D, Vargas JR, Sanchez JC, Ravier F, Pasquali C, Hochstrasser DF (1997a) Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images: I. Features and user interface. Electrophoresis 18:2724–2734

Appel RD, Vargas JR, Palagi PM, Walther D, Hochstrasser DF (1997b) Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms. Electrophoresis 18:2735–2748

Berger SJ, Lee SW, Anderson GA, Pasa-Tolic L, Tolic N, Shen Y, Zhao R, Smith RD (2002) High-throughput global peptide proteomic analysis by combining stable isotope amino acid labeling and data-dependent multiplexed-MS/MS. Anal Chem 74:4994–5000

Bienvenut WV, Sanchez JC, Karmime A, Rouge V, Rose K, Binz PA, Hochstrasser DF (1999) Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot. Anal Chem 71:4800–4807

Binz PA, Muller M, Hoogland C, Zimmermann C, Pasquarello C, Corthals G, Sanchez JC, Hochstrasser DF, Appel RD (2004) The molecular scanner: concept and developments. Curr Opin Biotechnol 15:17–23

Caprioli RM, Farmer TB, Gile J (1997) Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. Anal Chem 69:4751–4760

Chaurand P, Schwartz SA, Caprioli RM (2004) Profiling and imaging proteins in tissue sections by MS. Anal Chem 76:87A–93A

Dowsey AW, Dunn MJ, Yang GZ (2003) The role of bioinformatics in two-dimensional gel electrophoresis. Proteomics 3:1567–1596

Gade D, Gobom J, Rabus R (2005) Proteomic analysis of carbohydrate catabolism and regulation in the marine bacterium *Rhodopirellula baltica*. Proteomics 5:3672–3683

Galicia MC, Vertes A, Callahan JH (2002) Atmospheric pressure matrix-assisted laser desorption/ionization in transmission geometry. Anal Chem 74:1891–1895

Garrels JI (1989) The QUEST system for quantitative analysis of two-dimensional gels. J Biol Chem 264:5269–5282

Heeren RM (2005) Proteome imaging: a closer look at life's organization. Proteomics 5:4316–4326

Hillenkamp F, Karas M (1990) Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. Methods Enzymol 193:280–295

Hoogland C, Mostaguir K, Sanchez JC, Hochstrasser DF, Appel RD (2004) SWISS-2DPAGE, ten years later. Proteomics 4:2352–2356

Katajamaa M, Oresic M (2005) Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics 6:179–191. doi:10.1186/1471-2105-6-179

Lemkin PF, Lester EP (1989) Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modeling. Electrophoresis 10:122–140

Lemkin PF, Lipkin LE (1981) GELLAB: a computer system for two-dimensional gel electrophoresis analysis. III. Multiple two-dimensional gel analysis. Comput Biomed Res 14:407–446

Li XJ, Pedrioli PG, Eng J, Martin D, Yi EC, Lee H, Aebersold R (2004) A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry. Anal Chem 76:3856–3860

Luxembourg SL, Mize TH, McDonnell LA, Heeren RM. (2004) High-spatial resolution mass spectrometric imaging of peptide and protein distributions on a surface. Anal Chem 76:5339–5344

Maarten Altelaar AF, Ponsioen B, Jalink K, Heeren RM, Piersma SR (2005) Imaging mass spectrometry study of the spatial behaviour of membrane molecules in single neuroblastomas cells. In: Proceedings of the 53rd ASMS conference on mass spectrometry, San Antonio, TX, 5–9 June 2005

McDonnell LA, Mize TH, Luxembourg SL, Koster S, Eijkel GB, Verpoorte E, de Rooij NF, Heeren RM (2003) Using matrix peaks to map topography: increased mass resolution and enhanced sensitivity in chemical imaging. Anal Chem 75:4373–4381

Müller M, Gras R, Appel RD, Bienvenut WV, Hochstrasser DF (2002a) Visualization and analysis of molecular scanner peptide mass spectra. J Am Soc Mass Spectrom 13:221–231

Müller M, Gras R, Binz PA, Hochstrasser DF, Appel RD (2002b) Molecular scanner experiment with human plasma: improving protein identification by using intensity distributions of matching peptide masses. Proteomics 2:1413–1425

MSight (2007) Swiss Institute of Bioinfomatics, Geneva. http://www.expasy.org/MSight/. Cited 1 March 2007

Nadler TK, Wagenfeld BG, Huang Y, Lotti RJ, Parker KC, Vella GJ (2004) Electronic Western blot of matrix-assisted laser desorption/ionization mass spectrometric-identified polypeptides from parallel processed gel-separated proteins. Anal Biochem 332:337–348

Olson AD, Miller MJ (1988) Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. Anal Biochem 169:49–70

Orchard S, Hermjakob H, Binz PA, Hoogland C, Taylor CF, Zhu W, Julian RK Jr, Apweiler R (2005) Further steps towards data standardisation: the Proteomic Standards Initiative HUPO 3(rd) annual congress, Beijing 25–27(th) October, 2004. Proteomics 5:337–339

Palagi PM, Walther D, Quadroni M, Catherinet S, Burgess J, Zimmermann-Ivol CG, Sanchez JC, Binz PA, Hochstrasser DF, Appel RD (2005) MSight: an image analysis software for liquid chromatography-mass spectrometry. Proteomics 5:2381–2384

Palmblad M, Ramstrom M, Markides KE, Hakansson P, Bergquist J (2002) Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. Anal Chem 74:5826–5830

Parker KC (2002) Scoring methods in MALDI peptide mass fingerprinting: ChemScore, and the ChemApplex program. J Am Soc Mass Spectrom 13:22–39

Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R (2004) A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 22:1459–1466

Phenyx (2007) Geneva Bioinformatics, Geneva. http://www.genebio.com/products/ protein_id.html. Cited 1 March 2007

Pleissner KP, Hoffmann F, Kriegel K, Wenk C, Wegner S, Sahlstrom A, Oswald H, Alt H, Fleck E (1999) New algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gel databases. Electrophoresis 20:755–765

Rohner TC, Staab D, Stoeckli M (2005) MALDI mass spectrometric imaging of biological tissue sections. Mech Ageing Dev 126:177–185

Rubakhin SS, Jurchen JC, Monroe EB, Sweedler JV (2005) Imaging mass spectrometry: fundamentals and applications to drug discovery. Drug Discov Today 10:823–837

Scherl A, Zimmermann-Ivol CG, Di DJ, Vaezzadeh AR, Binz PA, Amez-Droz M, Cochard R, Sanchez JC, Gluckmann M, Hochstrasser DF (2005) Gold coating of non-conductive

membranes before matrix-assisted laser desorption/ionization tandem mass spectrometric analysis prevents charging effect. Rapid Commun Mass Spectrom 19:605–610

Schwartz SA, Weil RJ, Johnson MD, Toms SA, Caprioli RM (2004) Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression. Clin Cancer Res 10:981–987

Skolnick MM (1982) An approach to completely automatic comparison of two-dimensional electrophoresis gels. Clin Chem 28:979–986

Smilansky Z (2001) Automatic registration for images of two-dimensional protein gels. Electrophoresis 22:1616–1626

Spengler B, Hubert M (2002) Scanning microprobe matrix-assisted laser desorption ionization (SMALDI) mass spectrometry: instrumentation for sub-micrometer resolved LDI and MALDI surface analysis. J Am Soc Mass Spectrom 13:735–748

Stoeckli M, Chaurand P, Hallahan DE, Caprioli RM (2001) Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. Nat Med 7:493–496

Stoeckli M, Staab D, Staufenbiel M, Wiederhold KH, Signor L (2002) Molecular imaging of amyloid beta peptides in mouse brain sections using mass spectrometry. Anal Biochem 311:33–39

Stoeckli M, Knochenmuss R, McCombie G, Staab D, Rohner TC (2005) MALDI MSI of compounds and metabolites in whole-body tissue sections. In: Proceedings of the 53rd ASMS conference on mass spectrometry, San Antonio, TX, 5–9 June 2005

Todd PJ, Schaaff TG, Chaurand P, Caprioli RM (2001) Organic ion imaging of biological tissue with secondary ion mass spectrometry and matrix-assisted laser desorption/ionization. J Mass Spectrom 36:355–369

Tonella L, Walsh BJ, Sanchez JC, Ou K, Wilkins MR, Tyler M, Frutiger S, Gooley AA, Pescaru I, Appel RD, Yan JX, Bairoch A, Hoogland C, Morch FS, Hughes GJ, Williams KL, Hochstrasser DF (1998) '98 *Escherichia coli* SWISS-2DPAGE database update. Electrophoresis 19:1960–1971

Tuloup M, Hernandez C, Coro I, Hoogland C, Binz PA, Appel RD (2003) Aldente and BioGraph : an improved peptide mass fingerprinting protein identification environment. In: Ducret A, Eberle AN, Faupel MD et al (eds) Proceedings of the Swiss Proteomics Society 2003 congress: understanding biological systems. FontisMedia, Lausanne, pp 174–176

Vo KP, Miller MJ, Geiduschek EP, Nielsen C, Olson A, Xuong NH (1981) Computer analysis of two-dimensional gels. Anal Biochem 112:258–271

Yanagisawa K, Shyr Y, Xu BJ, Massion PP, Larsen PH, White BC, Roberts JR, Edgerton M, Gonzalez A, Nadaf S, Moore JH, Caprioli RM, Carbone DP (2003) Proteomic patterns of tumour subsets in non-small-cell lung cancer. Lancet 362:433–439

Yu KH, Rustgi AK, Blair IA (2005) Characterization of proteins in human pancreatic cancer serum using differential gel electrophoresis and tandem mass spectrometry. J Proteome Res 4:1742–1751

Zhang H, Yi EC, Li XJ, Mallick P, Kelly-Spratt KS, Masselon CD, Camp DG 2nd, Smith RD, Kemp CJ, Aebersold R (2005) High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry. Mol Cell Proteomics 4:144–155

7 Data Integration in Proteomics

Frédérique Lisacek, Christine Hoogland, Lydie Bougueleret, and Amos Bairoch

Abstract

The present chapter attempts to cover the recent initiatives directed towards representing, displaying and processing protein-related data suitable for interpreting proteomics data. Examples of integrated proteomics data are described and commented on. Challenging issues of integration are briefly discussed as well as directions for ongoing and future work. In particular, we consider the overlap between data integration, integrative biology and systems biology.

7.1 Introduction

The broad workflow of proteomics usually encompasses protein extraction, separation and analysis by mass spectrometry. Resulting mass spectra are subsequently processed, and mass data are then matched with sequence data from databases to identify the proteins present. Thus, in a first approximation, the raw output from a proteomic study usually consists of a list of proteins along with accompanying expression profiles relevant to the experimental conditions under study. The characterisation of the selected proteins is the next stage and is often a very difficult task. It involves not only determining the precise form of each protein (e.g. splice variant, phenotype, post-translational processing), but also the investigation of their possible interactions. The latter studies raise a number of issues that are common to many bioinformatics applications. In fact, these issues are common ones given that sequences constitute the core data type for all of these applications.

Various independent sources are available for the manual or automated extraction of relevant information related to a sequence or a collection of sequences.[1] However, the reliability of all extracted information needs to be ascertained. The different sources need to be cross-checked, and further digging into mass data is possibly involved. Only then can contextual constraints be reliably identified that characterise protein structure, function,

[1]In the specific context of proteomics, the focus is of course on amino acid sequences.

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

modifications and interactions. These characteristics constitute what is often called 'sequence annotation'. Well-annotated proteins are clearly valuable for interpreting experimental results.

Automated procedures for gathering multisource and heterogeneous information are commonly acknowledged as data-integration initiatives. By definition, integration entails producing a synthetic picture, which involves two distinct tasks: centralising information in one given location and the blending of this information, given an underlying principle. A mountain of technical difficulties relative to format-compatibility problems have channelled most bioinformatics efforts towards tackling the first task. This has led to information 'piling up' in one location. But the accumulation of properties of biological entities is as informative as listing ingredients for cooking, where the recipe is missing! Our incomplete knowledge of biology cannot help us resolve how these descriptive features should be grouped or when the groups should be processed. And a recipe is precisely about quantification and chronology.

The cooking analogy also emphasises the necessary 'blending-in' step of integration, which is too often overlooked. In fact, the analysis of texture is the hidden key to the blending process. In much the same way that some ingredients mix or do not mix smoothly, multisource data can complement or not complement each other. The texture or the granularity of information will determine the way by which information should be merged to become synthetic. Defining and understanding varying grains of information allows the definition of information levels. Technically speaking, a synthetic picture is produced each time zooming out from one level to the next is made possible. In other words, a zooming point corresponds to a change of texture. Hence, clear definitions of levels and zooming operations are essential requirements of the blending-in step of integration (Lisacek et al. 2004).

The integration of gene-to-protein data follows such an outline. All steps are visible, for instance, in Ensembl (Hubbard et al. 2005). The known transitions from a high-level chromosome to a low-level protein sequence set an appropriate scheme for structuring gathered pieces relative to genomic sequences. DNA sequences are consistently mapped with multiple transcripts and related to spliced introns and translatable exons. Gene loci along a chromosome are set as references for defining zooming operations. In this case, integration reflects meaningful principles of molecular biology (the central dogma) that includes an identified chronology of events.

At this stage, the underlying principles of the blending step for integration need to be specified for proteomics. Up until now, mainly two motivations have guided integration procedures. First, centralisation and homogeneity were required for connecting and harmonising experimental workflows. This is a technical challenge that has been successfully taken up in proteomics by some authors as shown further in this chapter (Hoogland et al. 2004; Desiere et al. 2005). Second, centralisation and homogeneity were driven by biological considerations. This is a key issue in biology and bioinformatics that is poorly addressed.

The move towards integration was initiated years ago by some of the most established and recognised bioinformatics resources (e.g. Swiss-Prot, KEGG, FlyBase). They centralised information from multiple sources and have circumvented the blending step by complementing data with manual annotation. The resultant knowledge is comparable to that of an encyclopaedia.[2] Unfortunately, our lack of understanding of laws governing protein folding and interactivity and our incomplete knowledge of pathways and cellular activities have hindered the design of appropriate and comprehensible models for the blending of protein-related information. So far, gene ontology (GO) (Harris et al. 2004) remains the most popular initiative for organising biologically interpretable protein data at molecular and cellular levels. But in GO, knowledge is currently unevenly represented owing to the relative novelty of this effort and also our limited understanding of biology. Ontology design has recently become an active field in bioinformatics (Schulze-Kremer 2002), and this trend further justifies the need for the determination of principles that set a more stable basis for data integration.

More recent initiatives of data integration have implied that quantification of information is the next hurdle. Indeed, a meaningful and useful description of a complex entity or a process is a set of attributes with a matched weighting. These weights express the relative importance of each component of the entity or of each event in a process. Determining descriptors or attributes is often relatively straightforward. However, our lack of biological knowledge will hinder our capacity to make an accurate system of weighting. A number of bioinformatics databases provide graphical views of interconnected objects but no associated weighting to help support data interpretation. As a result, these views do not always improve our understanding. Paradoxically, data visualisation is increasingly used to help define more accurate weighting schemes. Varying views of data are proposed to human awareness for a finer selection of information.

A recent and striking shift in data collection and representation is seen in the move from bioinformatics databases to atlases (EMAGE, Baldock et al. 2003; Protein Atlas, Ulhen et al. 2005; Novartis Atlas, Su et al. 2004; PeptideAtlas, Desiere et al. 2005). For example, the PeptideAtlas and the Protein Atlas have recently been introduced as resources for browsing and querying protein data. This generation of atlases, resources conventionally seen as books of maps, illustrations and drawings, may be a reflection of our inability to blend information in appropriate ways. In this manner, data exploration may have become an indispensable preliminary step towards integration. Interestingly, the navigation between different views in an atlas might help us refine the definitions required for the overall task of biological data integration.

[2]An encyclopaedia is defined as a comprehensive compilation of a body of knowledge that gives information on many aspects of the subject treated. This is definitely the case of highly curated databases. Incidentally, Swiss-Prot is designated as a knowledgebase, not an encyclopaedia.

At this stage, the tasks of integration can be summarised as (1) the selection of sources from which information is extracted and pooled in one location, (2) the quantification and interconnection of gathered information that includes the assessment of the relative importance of pieces of information, (3) the visualisation of interconnected information and (4) the gradual definition and implementation of one or more underlying principles for structuring information (i.e. ontologies and models). This last task involves questions related to the chronology and dynamics of events that are only starting to be addressed in data integration. More generally, a series of unsolved questions is attached to each of these tasks.

This chapter is structured as follows. Examples of integrated proteomics data are described and commented on. Then, the more challenging issues of integration are briefly discussed as well as directions for ongoing and future work. In particular, we consider the overlap between data integration, integrative biology and systems biology.

7.2 Integration As Gathering and Cross-Linking Information

7.2.1 Selection of Sources and Quantification

An enormous variety of bioinformatics data resources are available to the community. To help naïve users, *Nucleic Acids Research* publishes a yearly database issue that summarises those which are Internet-accessible. In this, it can be seen that the number of databases is in constant growth. Amazingly, the Molecular Biology Database Collection is sorted under a number of headings that change from year to year. In other words, there is no standardised and persistent categorisation of resources. As a result, the selection of resources remains a subjective and geography-dependent task.

Reliability and exhaustiveness are two common criteria that are used to assess the quality of a source. Reliability often comes through usage and steadily increases with time. Exhaustiveness is, in contrast, a changing criterion as detailed in the next section. A third and important criterion is time-resistance or the frequency of update. DNA and protein sequence data can very quickly become outdated. The rapid evolution of technologies in molecular biology, the rapid growth of many high-throughput methods and the ever-increasing flow of incoming data require regular and rigorous consistency checks. New data have to be internally checked and cross-checked as well. Databases updated frequently are more trustworthy that those that are released once only.

7.2.1.1 Trends in Databases

A fact-based assessment can be undertaken on any particular set of proteins and their associated bioinformatics resources (databases and tools). Uneven

data production over many years has had the result that a great deal is known about some proteins but hardly anything is known about others. For instance, databases of protein structures contain more proteins that form crystals easily, rather than transmembrane proteins that do not easily crystallise.

For many years data resources have had rapidly increasing numbers of entries, reflecting an incessant production of data. Efforts were often guided by criteria such as maximum coverage of a topic or of a species, therefore tending towards exhaustiveness. As a result, information in a lot of large databases is often redundant. Alternatively, it is averaged and specificity is lost.

The rather recent release of genome sequence data has modified this view. An exhaustive set of proteins for a species, perhaps excluding splice variants, is now accepted as being a defined and manageable number of entities. The number of entries of a genome or proteome database is thus not expected to grow, but information related to each entry should grow dramatically. Accordingly, we might expect the properties of objects to become exhaustive. Database expansion has shifted from breadth to depth and from objects to properties to generate a clearer picture of biology.

As a consequence of the former definition of exhaustiveness, biological data were often collected because they were available as opposed to being deliberately selected. Resulting trends can be identified in databases. For instance, most protein family databases are biased towards enzyme-related domains, given the traditional tight knit between structural domains and enzymatic activity in protein studies. Once a bias is made explicit, its impact on the quality of annotation can be assessed and relevant questions can be set. As a corollary, available information is weighed to counteract the effect of that bias.

7.2.1.2 Data Evolution

In a review published in 1997, one of us (Bairoch 1997) listed trends that were foreseen as gaining importance in the future. As forecasted, the exponential increase in the volume of information has given rise to new types of databases and major efforts in data integration.

Fast processing is particularly needed given the current intensive effort for sequencing complete genomes. However, speed and quantity are often achieved to the detriment of quality. This issue is well introduced in Gattiker et al. (2003), where high standards for producing reliable protein annotations are described. Among others, a relevant strategy involves gathering sequences into consistent families while carefully defining similarity criteria. Grouping criteria do not necessarily reflect a global similarity of amino acid sequences. Some proteins can be functionally equivalent though very diverse at the sequence level.

A new generation of curated and comprehensive data resources has indeed emerged as a possible solution to the critical issue of information overflow (Matys et al. 2003; Cooper et al. 2003; Schomburg et al. 2004). Those resources

include non-redundant and exhaustive data as well as appropriate analysis tools to explore, visualise and analyse the many aspects of data. They are usually developed at a high cost of human expertise.

7.2.2 Biology Inspired Cross-Linking

7.2.2.1 The UniProt Universal Protein Knowledgebase

Since its creation in 1986, Swiss-Prot has been consistently considered as the 'gold standard' for protein sequence and associated data. The encyclopaedic nature of the database and its high quality of sequence annotation have contributed to this reputation. In 1996, shortly after the first release of early genome sequences, an automated translation of the EMBL nucleotide sequence database (TrEMBL) was introduced to keep up with the rapid accumulation of this sequence data. TrEMBL was intended as a supplement to Swiss-Prot: TrEMBL entries are generated fully automatically while strictly following the Swiss-Prot format (see further in this section).

In parallel, the Protein Information Resource (PIR) was also created in the mid-1980s and the corresponding Protein Sequence Database (PSD) was created to help keep up with the massive production of nucleic acid sequences (Wu et al. 2003).

In 2003, Swiss-Prot, TrEMBL and PIR-PSD were merged into the UniProt Knowledgebase (UniProtKB). UniProtKB has become the reference resource of annotated protein sequences, complemented with functional information (Bairoch et al. 2005). All suitable PIR-PSD sequences missing from Swiss-Prot and TrEMBL were incorporated into UniProtKB and bidirectional cross-references were created to allow the easy tracking of PIR-PSD entries.[3]

UniProtKB comprises two parts: (1) a section of fully, manually annotated records resulting from literature information extraction and curator-evaluated computational analysis (UniProtKB/Swiss-Prot) and (2) a section with computationally analysed records potentially awaiting full manual annotation (UniProtKB/TrEMBL). UniProtKB provides a broad range of links between sequence data and all possible protein properties, including 3-D structure, biochemical activity and functional and structural classifications. The structure of UniProtKB entries is essentially the same as that of former Swiss-Prot entries (Bairoch 1997). Data integration in UniProtKB includes consistent cross-references between sequence data and a range of protein properties. UniProtKB/TrEMBL results from applying automatic annotation methods. It includes the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ nucleotide sequence databases.

[3]In 2006, the transfer into UniProt of references and experimentally verified data present in PIR but missing from Swiss-Prot and TrEMBL is ongoing.

Filtering rules are implemented to guarantee the consistency of automatic annotation (Apweiler 2001).

In UniProtKB, biologists carry out manual annotation based on literature survey and sequence analysis. This manual annotation effort is divided into projects that are species-dependent. The focus is mostly dictated by the incoming flow of complete genome data and includes among others (1) the Human Proteome Initiative (HPI) project, for the annotation of proteins in human and other mammal species; (2) the High-Quality Automated and Manual Annotation of Microbial Proteomes (HAMAP) project for the annotation of proteins in fully sequenced prokaryote species; and (3) the Plant Proteome Annotation Program (PPAP) for the annotation of proteins in fully sequenced plant species.

The main goal of the HPI project is to annotate all known human sequences according to the quality standards of UniProtKB/Swiss-Prot (O'Donovan et al. 2001). The information for each known protein includes the description of its function, its domain structure, subcellular localisation, post-translational modifications, variants and similarities to other proteins. The HPI contains several subcomponents: (1) annotation of all known human proteins; (2) annotation of mammalian orthologues of human proteins; (3) annotation of all known human polymorphisms at the protein sequence level; (4) annotation of all known post-translational modifications in human proteins; and (5) tight links to structural information.

The main goal of the HAMAP project is to automate the annotation of a significant percentage of proteins originating from bacterial and archaeal genome-sequencing projects, with no decrease in quality. It is also used to annotate proteins encoded by complete plant and algal plastid genomes (e.g. chloroplasts, cyanelles, apicoplasts, non-photosynthetic plastids). The automatic annotation relies on a rule-based system (Gattiker et al. 2003). The resulting annotation is only kept if it matches the quality of manual annotation. Many checks are performed in order to prevent the propagation of erroneous annotation and to identify problematic cases; these are channelled to manual curation.

HAMAP families store information on protein orthologues. Expert curators manually generate orthologous microbial protein families. They are used for the high-quality automatic annotation of microbial proteomes. Each family is composed of the following data: (1) annotation that is propagated to member entries (e.g. protein name, keywords); (2) computed features (e.g. export signals, transmembrane regions) that may be applied to entries by using appropriate prediction programs; (3) alignments of a representative set of entries; and (4) profiles that are automatically generated from the alignments which serve for the identification of new members in complete microbial proteomes.

The PPAP project is devoted to the annotation of plant-specific proteins and protein families (Schneider et al. 2005). Proteins of *Arabidopsis thaliana* have been targeted first. *Arabidopsis* proteins in UniProtKB/Swiss-Prot are

correlated with genomic data, and therefore, have an ordered locus name or Arabidopsis Genome Initiative (AGI) number based on the sequential ordering of the genes on the chromosomes. Proteins missing this information are either in regions of the genome not yet sequenced or are poorly defined from the results of gene prediction. Annotation of *Arabidopsis* proteins relies extensively on the presence of such ordered locus names. AGI numbers allow a clear distinction between two almost identical homologous genes and simple conflicts due to sequencing errors. Moreover, a wealth of *Arabidopsis* full-length complementary DNA sequences contribute to refining gene models and subsequently setting cleaner definitions of protein sequences. Since 2005 the other PPAP annotation target species has been *Orysa sativa* (rice) (International Rice Genome Sequencing Project 2005). Automatic annotation of the rice genome predicts the presence of more than 55,000 genes. However, many putative genes with no homology to *Arabidopsis* counterparts may be erroneous predictions, or sequences that are never translated into functional proteins in vivo.

These targeted initiatives emphasise the current efforts invested into protein annotation in completely sequenced organisms. The corresponding sets of proteins can be extracted by querying the database with 'complete proteome' as a UniProtKB/Swiss-Prot keyword.

7.2.2.2 Human Protein Atlas

The Human Protein Atlas (HPA) recently designed and constructed by Uhlen et al. (2005) studies the expression of normal and diseased tissues with antibody-based proteomics. It aims to systematically generate high-quality antibodies to all non-redundant human proteins and use these to localise all proteins in human tissues. It is also studying the effects of post-translational modifications and protein isoforms on possible functional variations. The non-redundant set of human proteins is defined as one product from each gene locus. In other words, antibodies are targeted to epitopes that are shared among the various protein forms. They are produced using a high-throughput method that involves the cloning and the expression of protein epitope signature tags (Nilsson et al. 2005).

HPA contains histological images from sections of human tissue samples. Each antibody in the database is used for the immunohistochemical staining of normal and cancerous tissues. A brown-black staining reveals the localisation of a protein of interest. The tissue section is also histochemically stained to visualise microscopic features. Tissue microarrays are used for immunohistochemical staining of a large number and variety of normal and cancerous tissues. An extended set of controlled vocabularies for tissues is included.

These efforts have resulted in an antibody-based Protein Atlas that includes expression and localisation profiles in normal human tissues and

dozens of different cancers. This information can be accessed and browsed in an online database, starting from a gene or a tissue name, from a chromosomal location or an antibody identifier. Images are magnified at convenience. Release 2.0 of HPA (30 October 2006) contains over 1,500 antibodies representing over 1,300 different proteins and approximately 1.2 million immunohistological images. All Protein Atlas entries are linked to their cognate UniProtKB entries where links are provided back to HPA. As such, it provides the first link between sequence data and systematic protein localisation and expression. As suggested in Sect. 7.1, data integration is intended in this case as a consistent cross-reference between sequence data and protein-expression information.

7.2.3 Integrating Elements of the Proteomics Workflow

7.2.3.1 High-Throughput Data: Standards and Repositories

High-throughput analytical methods have produced an ever-growing flow of new data. This situation has spawned a number of initiatives to address the issue of data storage and data standardisation. The transcriptomics community has led by example, introducing a data format termed MIAME – the minimum information about a microarray experiment. Following the same principle, a MIAPE format – minimum information about a proteomics experiment – is currently being defined. These formats are set "to enable the unambiguous interpretation of the results of an experiment and to allow its reproduction" (Microarray Gene Expression Data Society 2005).

The MIAPE project is carried out by a working group in charge of defining standardised formats for proteomics experiments (General Proteomics Standards 2002). This working group is one of three, all launched through the Proteomics Standards Initiative (PSI) of the Human Proteome Organisation (HUPO). The kick-off meeting took place in Washington in April 2002. The PSI instigators advocated setting the definitions of community standards for data representation in proteomics. Such standards are expected to facilitate data comparison, exchange and verification. They have been regularly discussed in the literature since then (Taylor et al. 2003; Pedrioli et al. 2004). Besides an involvement in setting a standardised general proteomics format, PSI supports other working groups in key areas of proteomics. These include 2-D gel electrophoresis, mass spectrometry and protein–protein interaction data (Proteomics Standards Initiative 2002).

The HUPO PSI is also closely related to the development of the Proteomics Identifications database (PRIDE). PRIDE is put together to provide the proteomics community with a public repository for protein and peptide identifications together with the evidence supporting these identifications (Martens et al. 2005). PRIDE is a centralised, standards-compliant, public data repository for proteomics data. Each entry in the database

contains the proteins identified, a list of peptides used to make identifications, the tissue, the experiment conducted, the conditions of the experiment, any post-translational modifications to those peptides and links to any publication describing the experiment. PRIDE has a dual objective. It is designed to provide (1) a common data-exchange format and repository to support proteomics literature publications and (2) a reference set of tissue-based identifications for use by the community.

Other repository initiatives are more advanced than PRIDE and appear as mature resources in Sect. 7.2.4. Lastly, a new initiative called the Computational Portal and Analysis System (CPAS) was very recently launched. It is presented as a set of Web-based bioinformatics and collaboration tools to help scientists store, analyse and share data from high-throughput experiments. It includes proteomics data and other experimental data (CPAS 2006).

7.2.3.2 SWISS-2DPAGE

SWISS-2DPAGE is an annotated database that assembles data on proteins from a variety of human and mouse biological samples as well as from *Arabidopsis thaliana*, *Dictyostelium discoideum*, *Escherichia coli*, *Saccharomyces cerevisiae* and *Staphylococcus aureus*. In all cases, proteins have been identified on 2-D polyacrylamide gel electrophoresis (PAGE) reference maps. SWISS-2DPAGE provides links between sequence data and protein expression. Most recorded proteins have been identified by one or more methods, including mass spectrometry, microsequencing, immunoblotting, gel comparison and amino acid composition.

The SWISS-2DPAGE database was the first 2-D electrophoresis (2-DE) federated database available on the Internet (Appel et al. 1993). Since then, it has been continuously accessible and expanded (Hoogland et al. 2004), and contains close to 40 maps. Various types of information (such as genome data, organism-specific data, protein families or domains, polymorphisms, mutations, structure, metabolic pathway) are brought together by cross-linking to other resources such as UniProtKB, PubMed and other federated 2-DE databases (HSC-2DPAGE, PHCI-2DPAGE, Siena-2DPAGE).

Each protein entry includes mapping procedures, physiological and pathological information, experimental data (isoelectric point, molecular weight, amino acid composition) and bibliographical references. Where proteins have been identified by mass spectrometry, peptide mass data are provided. In addition, SWISS-2DPAGE provides numerous 2-D PAGE images. These show the location of proteins identified on the 2-D gel, as well as a theoretical region computed from protein sequences which indicate where unidentified proteins might be found.

The data in SWISS-2DPAGE were initially stored as flat files. Recently, however, this was changed to a relational format, to assist in managing data consistency and maintenance (Mostaguir et al. 2003). The current schema is

flexible enough to facilitate integration of new types of data and is able to be adapted to specific needs. Various developments have been made to obtain consistent information and facilitate data extension from external sources. For instance, protein function or taxonomy is automatically brought up to date from information available in UniProtKB.

A number of query tools are unique to SWISS-2D PAGE. These are:

1. *Get protein list for a reference map*: This retrieves a table of all the protein entries identified on a given reference map, with all relevant 2-DE information (spot serial number, isoelectric point, molecular weight, mapping procedure, references).
2. *Get region on 2D gels for sequence*: This computes the estimated location for a user-entered sequence. The estimation is obtained according to the computed isoelectric point and molecular weight of the sequence.

In SWISS-2DPAGE, data integration is intended to allow a consistent incorporation of elements of the proteomics workflow with protein data. Information is gathered in a single location to help guide further investigations. For instance, a reference map of a 2-D gel can be consulted for selecting spots to be screened. Unidentified spots versus proteins that are already identified and mapped can be targeted more directly.

7.2.3.3 PeptideAtlas and the Global Proteome Machine

The PeptideAtlas (Desiere et al. 2005) was designed to store proteomics data generated by high-throughput methods. It is actually presented as an expandable resource for integration of data from diverse proteomics experiments. This initiative is driven by the assumption that protein-expression data can contribute to the annotation of eukaryotic genomes.

Data production follows the classic workflow of liquid chromatography–tandem mass spectrometry : protein extraction and digestion, peptide mass spectrometry analysis and sequence matching and final mapping on the corresponding genome. Standardisation of protocols of peptide data generation and peptide mass data description is obviously needed for data submission to PeptideAtlas. Mass spectrometry data must comply with the mzXML standard (Pedrioli et al. 2004) and with mzData (PSI). As previously discussed, other initiatives are addressing the issue of data standardisation in proteomics (e.g. MIAPE).

Tandem mass spectrometry spectra are stored in the PeptideAtlas database. Each statistically validated assignment of a peptide to a mass spectrum is recorded. A range of confidence thresholds is optionally made available for selecting peptides likely to match proteins. False-positive rates are correspondingly estimated. A database scheme supports different builds of PeptideAtlas, several versions of Ensembl, a range of eukaryotic organisms

and several reference protein sequence sets. PeptideAtlas currently includes data produced at the Institute for Systems Biology, which hosts the resource. As it is intended as a data repository, a large set of published data are also available for download.

The Global Proteome Machine (GPM) (Craig et al. 2004) is another similar approach to storing tandem mass spectrometry data. The database is used on its own, to provide answers to specific queries, as well as to serve as an index to experimental information stored in XML documents. The underlying schema serves as both an extension and a simplification of the MIAPE idea, for the purpose of validating observed protein coverage and peptide fragmentation data.

7.2.3.4 Other Noteworthy Efforts

All initiatives undertaken in proteomics to achieve data integration cannot be exhaustively reviewed in this chapter. Several summaries of links are available for further exploration. For instance, a very large selection of 2D-PAGE database servers is provided at WORLD-2DPAGE List (1995). Table 7.1 also gives a handful of references to recognised resources.

7.2.4 Integration As a Federated Effort

7.2.4.1 Proteomics Servers

ExPASy (ExPASy Proteomics Server 1993) was created in 1993 (Appel et al. 1994) and is still developed and maintained at the Swiss Institute of Bioinformatics

Table 7.1 Some integrated databases that are useful resources for proteome research. This table is not an exhaustive list

Name	URL	Reference
General purpose		
Integr8	http://www.ebi.ac.uk/integr8	Kersey et al. (2005)
Proteome browser	http://genome.ucsc.edu	Hsu et al. (2005)
Plant proteomics		
PROTICdb	http://cms.moulon.inra.fr/proticdb/protic/home	Ferry-Dumazet et al. (2005)
Microbiology		
Proteome Web	http://proteomeweb.anl.gov	Babnigg and Giometti (2003)
Proteome Database System for Microbial Research	http://www.mpiib-berlin.mpg.de/2D-PAGE/	Pleissner et al. (2004)

(Gasteiger et al. 2003). ExPASy is dedicated to the federation of databases and tools that are relevant to proteomic studies. The server hosts the following databases that are essentially developed and maintained at the same physical location:

- UniProtKB, described in Sect. 7.2.2
- SWISS-2DPAGE, described in Sect. 7.2.3
- The PROSITE database of protein domains and families
- The ENZYME repository of information relative to the nomenclature of enzymes
- The SWISS-MODEL repository, a database of automatically generated 3-D protein structural models

All databases available on ExPASy are *explicitly* cross-referenced to other molecular biology databases or resources available on the Internet. Approximately 30 *implicit* links to other additional resources are created on demand when certain views of UniProtKB are generated. This concept is targeted at data collections that do not have their own system of unique identifiers, but can be referenced via identifiers such as accession numbers or gene names. GeneCards (1997) is an example of a database implicitly linked to UniProtKB – it only shares an identifier with UniProtKB (the HUGO-approved gene name). Implicit links are a specific feature of ExPASy and are not available on other Web servers or in the UniProtKB data files downloadable by file transfer protocol. They significantly enhance database interoperability and strengthen the role of UniProtKB as a central hub for the interconnection of molecular biology resources.

ExPASy has an extensive collection of software tools (ExPASy Proteomics Tools 1994). Some of them are targeted towards access and display of the databases listed above. Others can be used for data analysis, such as the prediction of protein sequence features or the processing of proteomics data originating from 2-D PAGE and mass spectrometry experiments. ExPASy tools that are specifically designed for mass spectrometry data analysis perform computations and predictions while using annotations documented in the UniProtKB feature tables (e.g. the input form of peptide mass fingerprinting engine Aldente includes the optional selection of a range of protein features). This assists the detection of possible splice variants, post-translational modifications or protein processing.

The European Bioinformatics Institute (EBI) adevelops and maintains the PRIDE proteomics database (see Sect. 7.2.3) and federates a very wide spectrum of protein-related databases (European Institute of Bioinformatics Protein Databases 2006). However, the software tools available at EBI are not specifically targeted at mass spectrometry data analysis. Conversely, the Institute for Systems Biology offers a wide panel of software tools for proteomic studies (Seattle Proteome Center 2004) but does not develop nor does it maintain a wide variety of protein-related databases, with the

exception of PeptideAtlas. ExPASy remains the server that best covers the diversity of requirements for proteomics researchers.

7.2.4.2 Semantic Web Approach

The federation of a number of Internet resources can involve the exchange of data between these resources. However, crosstalk is hampered by source heterogeneity in terms of format, content and structure. The World Wide Web Consortium (W3C) has been developing interoperable technologies since 2001, under the specific name of the 'Semantic Web' (World Wide Web Consortium 1994). The declaration of the main principles of the Semantic Web (Koivunen and Miller 2001) demonstrates the potential relevance of this approach to problems of biological data integration. Six principles have set the basis for discussion on the topic:

1. *Principle 1: Everything can be identified by uniform resource identifiers (URI)*[4]*)*. "People, places, and things in the physical world can be referred to in the Semantic Web by using a variety of identifiers. Anyone who has control over a part of Web namespace can create a URI and say that it identifies something in the physical world." This statement obviously applies only to the situation restricted to biological resources.
2. *Principle 2: Resources and links can have types*. The current Web consists of resources and links. Resources are Web documents that rarely contain explicit explanations of their possible use and the nature of links to other Web documents. This type of explanation would be considered as metadata. Metadata could be described by relationships such as 'depends on', 'is a version of', 'has subject' or 'authors'. This information is definitely missing in biological resources.
3. *Principle 3: Partial information is tolerated*. The Semantic Web tools need to tolerate data decay and reuse of addresses. More generally this statement is central in the discussion on biological data formalisation and standardisation. The flexibility of the Resource Description Framework (RDF) versus the commonly chosen XML technology is indisputable, as detailed in Wang et al. (2005). The possibility of accommodating incomplete or even erroneous knowledge is indispensable in biology.
4. *Principle 4: There is no need for absolute truth*. "Not everything found on the Web is true and the Semantic Web does not change that in any way". This statement clearly holds in biology. The question of source reliability was already mentioned in Sect. 7.2.1.
5. *Principle 5: Evolution is supported*. "It is common that similar concepts are often defined by different groups of people in different places or even by

[4]In relation to uniform resource locators (URL).

the same group at different times. It would often be beneficial to combine the data available on the Web that uses these concepts." This statement is very true for the biological sciences. "The Semantic Web provides communities with tools that can be used to resolve ambiguities and clarify inconsistencies. Also new information can be added without insisting that the old has to be modified."

6. *Principle 6: Minimalist design*. "The Semantic Web makes the simple things simple, and the complex things possible. The aim of the W3C activity is to standardize no more than is necessary." This statement is food for thought in biology.

The principles listed above were proposed to help improve the management of heterogeneous sources. This assessment seems particularly appropriate for applications in life sciences. So far, only one reference explicitly details the implementation of the Semantic Web (Cheung et al. 2005). However, querying the resource on the corresponding Web site (Integrative YeastHub Project 2006) is still quite challenging.

7.3 Integration As Blending of Information

The remaining sections mainly focus on the semantic aspects of integration, that is, the diverse means of organising biological knowledge for corroborating, structuring and possibly merging pieces of information. Note that an important goal of formalising knowledge is predicting new knowledge. Prediction has long been, and is still envisaged, through the mining of data. Indeed, data mining is traditionally associated with knowledge discovery (Wikipedia 2001). Prediction is now also widely approached through simulation of processes. The following will shed some light on these issues.

7.3.1 Textual Information

Roughly speaking, biological data are either stored in databases (as seen in previous sections) or in published literature. It is in fact commonplace to state that biological knowledge is mainly textual knowledge. Annotations are written as text. Moreover, the best justification of an annotation is a published article. Biological data integration depends very often on mining and extracting information in texts. In turn, this information undermines or strengthens facts stated or derived in or from databases. As a result, the field of text mining for biology is currently expanding. For example, a range of methods are currently being used to populate protein–protein interaction databases (e.g. IntAct 2004) and to complement protein annotations. In scientific papers, the syntax of the English language, in particular verbs (e.g. 'inhibit', 'bind',

'associate', 'activate'), can be exploited to identify expressions describing molecular interactions of a subject protein and complementary proteins. This approach is known as natural language processing. Extracted information is structured but not necessarily quantified. Alternatively, expressions describing molecular interactions can be detected irrespective of the grammatical role of words, simply by counting the presence of terms of a predefined thesaurus. Extracted information is quantified but not structured. This approach is basically statistical. These two main strategies for parsing text have been designed, developed and improved over decades. Applications span a broad range of problems of information processing such as text summarisation or translation of language. However, questions that are central in biology still challenge most natural language processing methods. Biological knowledge is characterised by highly variable terminologies and a subsequent lack of consensus for describing entities and processes. This situation generates ambiguity and rough approximations, as exemplified in the first textbook dedicated to text mining in biology (Ananiadou and McNaught 2006). A number of open problems remain. High-quality annotation has, however, become more and more dependent upon good text-mining tools. Even though text mining is a side issue of the present chapter, it plays a central role in ontology building, which is the focus of the next section.

7.3.2 Ontologies

Protein data, as hinted in Sect. 7.2.1, are described in a collection of databases. A number of these databases are actually databases of classifications. These include protein family databases defined upon sequence similarity (e.g. InterPro 2001), structural databases defined upon structure similarity (e.g. SCOP 1994) and enzyme databases defined upon the enzymatic classification (e.g. ENZYME 1994). In each case a similarity measure is defined, and thresholds are then used as criteria for grouping proteins together. Each similarity measure is independent of one another. Except for UniProtKB cross-links, GO terms are the only biologically meaningful links between independent classifications of proteins. In other words, GO is used for structuring interpretable protein data and, therefore, for integrating these data. However, 'structure' is yet to be defined.

GO terms are organised in three independent categories: biological process, molecular function and cellular component. As pointed out in Bodenreider et al. (2005), the main flaw of GO is the hierarchical structure that prevents linking concepts from one level to the other. Assume a molecular function was matched with an EC number (e.g. EC 3.1.1.3). Assume this EC number is mapped on a pathway (glycerolipid metabolism in our example). This information cannot be found directly in GO. It is necessary to use another resource to correlate the information. For example, a UniProtKB entry may be cross-linked to the following two GO numbers:

1. GO:0004806; Molecular function: triacylglycerol lipase activity
2. GO:0006641; Biological process: triacylglycerol metabolism

A user can infer the link between the enzymatic activity and the appropriate metabolic process from this co-occurrence of terms. A machine could infer it as well in simple cases such as this example. However, ambiguous situations may arise if a specific enzymatic activity is part of several distinct metabolic pathways. Ironically, the ambiguity is not necessarily shared by an expert biochemist who would rely on contextual information for distinguishing the relevant pathway. This information is unfortunately not stored in GO, even though it may be available from textbooks. The limiting use of 'is a' or 'part of' for describing relations between biological entities is discussed in Schulze-Kremer (2002). Besides, ambiguity is often poorly managed by automatic procedures since conflicts are frequently solved on a case-by-case basis. Schulze-Kremer (2002) points out important distinctions between ontologies and controlled vocabularies and through a list of shortcomings considers GO as representing the latter category.

A variety of ontologies have been developed. A number of authors have insisted on the necessity of a concerted effort for the building of ontologies. The Open Biomedical Ontologies (2004) project reflects this trend as it centralises information and makes reference to all ontologies available and in use in the life sciences. Several types of ontologies are defined, and cover various descriptions of entities at stake. Experimental protocols, molecular or chemical processes and anatomy are some of the most targeted topics. Relevant terms of each topic are gradually developed into trees, which express the relations between those terms. Then, the correspondence between ontologies needs to be established. For instance, Soldatova and King (2005) would rather consider ontology of experiments as an intermediate between upper ontologies and specific scientific domain ontologies. Lan et al. (2003) suggest a species-dependent approach for defining and using structural and functional ontologies. At present, these issues are actively debated.

7.3.3 Examples of Visualisation Tools Merging Several Sources

Ontologies, as mentioned earlier, are mainly populated with information extracted from the literature, although databases are also relevant sources of data. In fact, experimental and digital data are now constantly mixed in biological studies. This situation is well illustrated in the proteomics-based analysis of the interferon response in human liver cells (Yan et al. 2004). The global protein expression in human liver carcinoma cells was compared depending on the presence or the absence of IFN-α treatment. Close to 1,400 proteins were identified by mass spectrometry, and 60% of these proteins could be mapped to a portion of the GO carrying annotated human genes. The GO terms associated with IFN-induced and

IFN-repressed proteins related the former to the immune response or signalling pathways, whereas the latter appeared to be mainly involved in development and RNA transport. This broad distinction was the first trend determined in the identified proteins. Further integration was achieved by examining a protein–protein interaction database. Then, data were visualised using Cytoscape (2001), a popular open-source visualisation package. It was implemented for visualising molecular interaction networks and integrating these interactions with other data. Needless to say, important conclusions drawn in this study were established through human expertise. Graph views of interconnected objects with limited quantification of edges generate general, not specific, insights. Such data visualisation is thought to help define more accurate information, which in turn can refine views of data. In fact, data-mining techniques are only powerful once a problem is well understood. Most bioinformatics tools produce rough outputs that guide the interpretation of biological results but do not necessarily produce biological insight directly. The present trend is towards exploratory tools.

The previous example shows that the synthetic aspect of integration is yet to be designed. This aspect depends on the definition of hierarchical levels. The basis of integration for protein data should obviously express transitions from sequence to structure, from structure to interacting partners and from interacting partners to pathways. These transitions are likely to involve diverse properties of proteins and hence various representations. Furthermore, protein properties need to be put back into context and contexts have to be explored. As stated in Sect. 7.1, a synthetic picture is produced each time the 'zooming out' from one level to the next is made possible, and zooming operations are essential for exploratory purposes (Lisacek et al. 2004). This point is further illustrated in Aloy and Russell (2005), where protein data are shown as part of a network of interactions down to a 3-D structure. Likewise in Nikitin et al. (2004), bacterial proteins are shown as combinations of family signatures in the context of a pathway and up to the context of a proteome. Visualisation tools are clearly a key issue of integration.

7.3.4 From Data Integration to Systems Biology

Integrative biology can be considered as a forerunner to systems biology. Integrative biology was put forward as a new discipline some 15 years ago as a natural extension of the progress made in developmental biology (University of California, Berkeley 2005). Mathematical models of morphogenesis and pattern formation that are much older approaches were also part of this integrative effort. In fact, the new theme emerged from those disciplines of biology integrating the time dimension.

Studying and determining underlying the biological principles for the integration of data is precisely the concern of integrative biology. Organisms

and their environment are considered on a series of hierarchical levels from gene products to whole organisms. Integrative physiology even promotes considering single molecules prior to proteins (Noble and Boyd 1993). In fact, boundaries between integrative biology/physiology and systems biology are now blurred by the common goal of these disciplines (Liu 2005). In all cases, structuring knowledge into levels (ordered by magnitude) and determining rules that would allow navigating between levels are required. Biological organisms are considered as distributed systems. They are represented as networks (Sect. 7.3.3). In this context, regulation appears as a central process for rationalising observed outputs of the system. Regulation is de facto a key theme of systems biology studies but the relationships between the topology, the dynamics and the underlying logics of these networks are still very difficult to establish (Schlitt and Brazma 2005).

Systems biology initiated many integrative projects, especially in proteomics (Institute for Systems Biology 2006). Some of the most interesting and challenging issues of systems biology are discussed in Zhu et al. (2003). They are summarised as (1) the lack of appropriate data, (2) the 'time invariance' of experimental data and (3) weak definitions of levels. The first issue may appear paradoxical. We are flooded by massive data production, yet we need new observational viewpoints to completely understand biology. This point meets the second issue. Pathways and maps often describe static information flow between molecular entities, whereas biological processes are temporal and spatial in nature. The third issue targets multiple description methods used in models that confuse our understanding.

The question comes down to formalising a language that would potentially explain rather than just describe entities and their relations. This issue is humorously debated in Lazebnik (2002). However, the need for a language to help formalise biological knowledge is the serious and important conclusion of the discussion. The issue is widely appreciated (Campbell 1982; Konopka 1997; Lisacek 2004) and still requires further debate. The incompleteness of knowledge is exacerbated by the underappreciated influence of context. Interestingly, the recognition of context-sensitive languages remains an NP-complete problem.[5]

A current answer to formalisation is the development of a range of simulation schemes. The goal of simulation is to quantify the behaviour of a system. The yield of a reaction, the steps of a molecular pathway, up to the full network of interacting entities that characterise a cellular activity are modelled and the behaviour of the resulting systems is then tested in response to defined perturbations. For instance, Moleculizer simulates a biochemical network that quantifies the contribution of protein complexes with

[5]In computer science, a problem is said to be NP-complete when no algorithm can reach the solution in polynomial (i.e. reachable) time.

a stochastic model (Lok and Brent 2005). Several other projects aim at designing virtual cells (E-Cell International Research Project 2003;. National Resource for Cell Analysis and Modeling 2005). Furthermore, recent initiatives are under way to merge efforts invested in ontology building, data representation and modelling. In particular, dedicated languages such as systems biology markup language (SBML) (Hucka et al. 2003) are under development in an attempt to support data representation and modelling. SBML is XML-based and is used or converted to in dozens of available software packages (Systems Biology Markup Language 2003). As in the case of mass spectrometry data exchange described in Sect. 7.2.3, comparable initiatives are being launched for the exchange of biochemical pathway data. This type of endeavour is likely to address key problems associated with defining principles of integration for synthesising information.

7.4 Concluding Remarks

The overall goal of proteome research is to generate a dynamic description of the molecular players within one or several biological processes. This can be viewed as rationalising experimental as well as digital data. Classically, methods for the integration of data fall into two opposed categories, namely data-driven versus model-driven. A data-driven approach attempts to rationalise data with little, if no, predetermined hypothesis. A model-driven approach relies on a preset model that guides the rationalisation of data. Most successful attempts to integrate and visualise data rely on mixing both approaches. Proteomics databases are a wealth of information that can complement incomplete models of biological processes.

References

Aloy P, Russell RB (2005) Structure-based systems biology: a zoom lens for the cell. FEBS Lett 579:1854–1858

Ananiadou S, McNaught J (2006) Introduction to text mining for biology. In: Ananiadou S, McNaught J (eds) Text mining for biology and biomedicine. Artech House, London, pp 1–12

Appel RD, Sanchez JC, Bairoch A, Golaz O, Miu M, Vargas JR, Hochstrasser DF (1993) SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. Electrophoresis 14:1232–1238

Appel RD, Bairoch A, Hochstrasser DF (1994) A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. Trends Biochem Sci 19:258–260

Apweiler R (2001) Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. Brief Bioinform 2:9–18

Babnigg, G, Giometti CS (2003) ProteomeWeb: a web-based interface for the display and interrogation of proteomes. Proteomics 3:584–600

Bairoch A (1997) Proteome databases. In: Wilkins MR, Williams KL, Appel RD, Hochstrasser DF (eds) Proteome research: new frontiers in functional genomics. Springer, Berlin, pp 93–132

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33:D154–159

Baldock RA, Bard JB, Burger A, Burton N, Christiansen J, Feng G, Hill B, Houghton D, Kaufman M, Rao J, Sharpe J, Ross A, Stevenson P, Venkataraman S, Waterhouse A, Yang Y, Davidson DR (2003) EMAP and EMAGE: a framework for understanding spatially organized data. Neuroinformatics 1:309–325

Bodenreider O, Aubry M, Burgun A (2005) Non-lexical approaches to identifying associative relations in the gene ontology. Proc Pac Symp Biocomput 91–102

Campbell, J (1982) Grammatical man: information, entropy, language, and life. Simon and Schuster, New York, pp 1–319

Cheung KH, Yip KY, Smith A, Deknikker R, Masiar A, Gerstein M (2005) YeastHub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics 21(Suppl 1):i85–i96

Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH (2003) GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. Nucleic Acids Res 31:511–513.

CPAS (2006) https://cpas.fhcrc.org/Project/home/begin.view. Cited 18 Jan 2007

Craig R, Cortens JP, Beavis RC (2004) Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 3:1234–1242

Cytoscape (2001) http://www.cytoscape.org/. Cited 18 Jan 2007

Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, Fausto N, Hafen E, Hood L, Katze MG, Kennedy KA, Kregenow F, Lee H, Lin B, Martin D, Ranish JA, Rawlings DJ, Samelson LE, ShiioY, Watts JD, Wollscheid B, Wright ME, Yan W, Yang L, Yi EC, Zhang H, Aebersold R (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol 6:R9

E-Cell International Research Project (2003) http://www.e-cell.org. Cited 18 Jan 2007

ENZYME (1994) Swiss Institute of Bioinformatics, Geneva. http://www.expasy.org/enzyme/. Cited 18 Jan 2007

European Institute of Bioinformatics Protein Databases (2006) European Institute of Bioinformatics, Hinxton. http://www.ebi.ac.uk/Databases/protein.html. Cited 18 Jan 2007

ExPASy Proteomics Server (1993) Swiss Institute of Bioinformatics, Geneva. http://www.expasy.org. Cited 18 Jan 2007

ExPASy Proteomics tools (1994) Swiss Institute of Bioinformatics, Geneva. http://www.expasy.org/tools/ Cited 18 Jan 2007

Ferry-Dumazet H, Houel G, Montalent P, Moreau L, Langella O, Negroni L, Vincent D, Lalanne C, de Daruvar A, Plomion C, Zivy M, Joets J (2005) PROTICdb: a web-based application to store, track, query, and compare plant proteome data. Proteomics 5:2069–2081

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31:3784–3788

Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJ, Lachaize C, Veuthey AL, Gasteiger E, Bairoch A (2003) Automated annotation of microbial proteomes in SWISS-PROT. Comput Biol Chem 27:49–58

GeneCards (1997) Weizmann Institute, Rehovot. http://www.genecards.org. Cited 18 Jan 2007

Harris MA, Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–261

General Proteomics Standards (2002) http://psidev.sourceforge.net/gps/index.html. Cited 18 Jan 2007

Hoogland C, Mostaguir K, Sanchez JC, Hochstrasser DF, Appel RD (2004) SWISS-2DPAGE, ten years later. Proteomics 4:2352–2356

Hsu F, Pringle TH, Kuhn RM, Karolchik D, Diekhans M, Haussler D, Kent WJ (2005) The UCSC Proteome Browser. Nucleic Acids Res 33:D454–458

Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D,Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E (2005) Ensembl 2005. Nucleic Acids Res 33:D447–453

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531

IntAct (2004) European Institute of Bioinformatics, Hinxton. http://www.ebi.ac.uk/intact/. Cited 18 Jan 2007

Institute for Systems Biology (2006) http://www.systemsbiology.org/. Cited 18 Jan 2007

Integrative YeastHub Project (2006) Yale Center for Medical Informatics, New Haven. http://yeasthub.gersteinlab.org. Cited 18 Jan 2007

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

InterPro (2001) European Institute of Bioinformatics, Hinxton. http://www.ebi.ac.uk/interpro/. Cited 18 Jan 2007

Kersey P, Bower L, Morris L, Horne A, Petryszak R, Kanz C, Kanapin A, Das U, Michoud K, Phan I, Gattiker A, Kulikova T, Faruque N, Duggan K, Mclaren P, Reimholz B, Duret L, Penel S, Reuter I, Apweiler R (2005) Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res 33:D297–302

Koivunen M-R, Miller E (2001) W3C Semantic Web Activity. Proceedings of the Semantic Web kick-off seminar, Helsinki, Finland, November 2001. http://www.cs.helsinki.fi/u/eahyvone/stes/semanticweb/kick-off/proceedings.html

Konopka AK (1997) Theoretical molecular biology. In Meyers RA (ed) Encyclopedia of molecular biology and molecular medicine, vol 6. VCH, Weinheim, pp 37–53

Lan N, Montelione GT, Gerstein M (2003) Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. Curr Opin Chem Biol 7:44–54

Lazebnik Y (2002) Can a biologist fix a radio? Or, what I learned while studying apoptosis. Cancer Cell 2:179–182

Lisacek F (2004) Methods of computational genomics: an overview. In: Konopka AK, Crabbe JC (eds) The compact handbook of computational biology. Dekker, New York, pp 279–342

Lisacek F, Chichester C, Gonnet P, Jaillet O, Kappus S, Nikitin F, Roland P, Rossier G, Truong L, Appel RD (2004) Shaping biological knowledge: applications in proteomics. Comp Funct Genomics 5:190–195

Liu ET (2005) Systems biology, integrative biology, predictive biology. Cell 121:505–506

Lok L, Brent R (2005) Automatic generation of cellular reaction networks with Moleculizer 1.0. Nat Biotechnol 23:131–136

Martens L, Hermjakob H, Jones P, Taylor C, Gevaert K, Vandekerckhove J, Apweiler R (2005) PRIDE: The PRoteomics IDEntifications database. Proteomics 5:3537–3545

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31:374–378

Microarray Gene Expression Data Society (2005) http://www.mged.org/Workgroups/MIAME/miame.html. Cited 18 Jan 2007

Mostaguir K, Hoogland C, Binz PA, Appel RD (2003) The Make 2D-DB II package: conversion of federated two-dimensional gel electrophoresis databases into a relational format and interconnection of distributed databases. Proteomics 3:1441–1444

National Resource for Cell Analysis and Modeling (2005) http://www.nrcam.uchc.edu/. Cited 18 Jan 2007

Nilsson P, Paavilainen L, Larsson K, Odling J, Sundberg M, Andersson AC, Kampf C, Persson A, Al-Khalili Szigyarto C, Ottosson J, Bjorling E, Hober S, Wernerus H, Wester K, Ponten F, Uhlen M (2005) Towards a human proteome atlas: high-throughput generation of mono-specific antibodies for tissue profiling. Proteomics 5:4327–4337

Nikitin F, Rance B, Itoh M, Kanehisa M, Lisacek F (2004) Using protein motif combinations to update KEGG pathway maps and orthologue tables. Genome Inform 15:266–275

Noble D, Boyd CAR (1993) The challenge of integrative physiology. In: Boyd CAR, Noble D (wds) The logic of life. Oxford University Press, Oxford, pp 1–13

O'Donovan C, Apweiler R, Bairoch A (2001) The human proteomics initiative (HPI). Trends Biotechnol 19:178–181

Open Biomedical Ontologies (2004) http://obo.sourceforge.net/browse.html. Cited 18 Jan 2007

Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R (2004) A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 22:1459–1466

Pleissner KP, Schmelzer P, Wehrl W, Jungblut PR (2004) Presentation of differentially regulated proteins within a web-accessible proteome database system of microorganisms. Proteomics 4:2987–2990

Proteomics Standards Initiative (2002) http://psidev.sourceforge.net/. Cited 18 Jan 2007

Seattle Proteome Center (2004) Proteomics tools. http://tools.proteomecenter.org/ software.php. Cited 18 Jan 2007

Schlitt T, Brazma A (2005) Modelling gene networks at different organisational levels. FEBS Lett 579:1859–1866

Schneider M, Bairoch A, Wu CH, Apweiler R (2005) Plant protein annotation in the UniProt knowledgebase. Plant Physiol 138:59–66

Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 32:D431–433

Schulze-Kremer S (2002) Ontologies for molecular biology and bioinformatics. In Silico Biol 2:179–193

SCOP (1994) Medical Research Council, Cambridge. http://scop.mrc-lmb.cam.ac.uk/scop/. Cited 18 Jan 2007

Soldatova LN, King RD (2005) Are the current ontologies in biology good ontologies? Nat Biotechnol 23:1095–1098

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067

Systems Biology Markup Language (2003) http://sbml.org/index.psp. Cited 18 Jan 2007

Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I, Mohammed S, Deery MJ, Howard JA, Dunkley T, Aebersold R, Kell DB, Lilley KS, Roepstorff P, Yates JR 3rd, Brass A, Brown AJ, Cash P, Gaskell SJ, Hubbard SJ, Oliver SG (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. Nat Biotechnol 21:247–254

Uhlen M, Bjorling E, Agaton C, Szigyarto CA, Amini B, Andersen E, Andersson AC, Angelidou P, Asplund A, Asplund C, Berglund L, Bergstrom K, Brumer H, Cerjan D, Ekstrom M, Elobeid A, Eriksson C, Fagerberg L, Falk R, Fall J, Forsberg M, Bjorklund MG, Gumbel K, Halimi A, Hallin I, Hamsten C, Hansson M, Hedhammar M, Hercules G, Kampf C, Larsson K, Lindskog M, Lodewyckx W, Lund J, Lundeberg J, Magnusson K, Malm E, Nilsson P, Odling J, Oksvold P, Olsson I, Oster E, Ottosson J, Paavilainen L, Persson A, Rimini R, Rockberg J, Runeson M, Sivertsson A, Skollermo A, Steen J, Stenvall M, Sterky F, Stromberg S, Sundberg M, Tegel H, Tourle S, Wahlund E, Walden A, Wan J, Wernerus H, Westberg J, Wester K, Wrethagen U, Xu LL, Hober S, Ponten F (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol Cell Proteomics 4:1920–1932

University of California, Berkeley (2005) Department of Integrative Biology. http://ib.berkeley.edu/aboutib/index.php. Cited 18 Jan 2007

Wang X, Gorlitsky R, Almeida JS (2005) From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nat Biotechnol 23:1099–1103

Wikipedia (2001) Data mining. http://en.wikipedia.org/wiki/Data_mining. Cited 18 Jan 2007

WORLD-2DPAGE List (1995) Swiss Institute of Bioinformatics, Geneva. http://www.expasy.org/ch2d/2d-index.html. Cited 18 Jan 2007

World Wide Web Consortium (1994) http://www.w3.org. Cited 18 Jan 2007

Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC (2003) The Protein Information Resource. Nucleic Acids Res 31:345–347

Yan W, Lee H, Yi EC, Reiss D, Shannon P, Kwieciszewski BK, Coito C, Li XJ, Keller A, Eng J, Galitski T, Goodlett DR, Aebersold R, Katze MG (2004) System-based proteomic analysis of the interferon response in human liver cells. Genome Biol 5:R54

Zhu H, Huang S, Dhar P (2003) The next step in systems biology: simulating the temporospatial dynamics of molecular network. BioEssays 26:68–72

8 Protein–Protein Interactions

Anne-Claude Gavin

Abstract

Protein–protein interactions are central to postgenome biology. Complex cellular functions are almost always the result of the coordinated action of several proteins, acting in molecular assemblies or pathways to achieve a particular task. The wiring schemes provided by protein interaction networks are believed to contribute to our understanding of the convoluted relationships between genomes and phenotypes. In human, impairment of pathway flow or deregulated connections can lead to abnormalities. The majority of targets of current therapeutics cluster in a limited number of pathways. One of the goals of this chapter is to exemplify and illustrate the importance of protein–protein interactions for human biology and pathology: how protein interaction networks contribute to the understanding of phenotype and more particularly human genetic disorders, helping the gene-finding process and bringing a molecular framework to the genetic heterogeneity and pleiotropy often observed in human syndromes. An important section of this chapter is also dedicated to a review of the proteomic technologies developed and adapted to chart protein–protein interactions on a whole cell or organism scale. Their respective advantages, shortcomings and limitations are presented. Finally, recent successes in the development of small molecules that target and interfere with protein–protein interaction sites open new avenues towards the design of drugs that specifically abrogate or modulate disease-relevant interactions. This constitutes the last topic of the chapter.

8.1 Introduction

Protein–protein interactions are of central importance to postgenome biology. Although proteins are the main effectors of the genomic message, they purvey only part of the message by themselves. However, proteins rarely act alone. Complex cellular functions are almost always the result of the co-ordinated action of several proteins acting in molecular assemblies or pathways to achieve a particular task (Alberts 1998). The wiring schemes provided by protein interaction networks are believed to contribute to our understanding

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

of the relationships between genomes and phenotypes (Jeong et al. 2001; Rubin 2001; Brunner and van Driel 2004). From an evolutionary perspective, the apparent complexity observed at the level of an organism does not necessarily relate to genome size. For example, the size of the human genome does not differ substantially from that of a much simpler organism, the worm *Caenorhabditis elegans*. Besides events such as alternative splicing and post-translational modifications (see Chap. 5), it is becoming apparent that the contextual combination of the gene products generates molecular diversity and may contribute to phenotypic complexity. In human, impairment of pathway flow or deregulated connections can lead to abnormalities. The majority of targets of current therapeutics cluster in a limited number of these cellular pathways (Brown and Superti-Furga 2003).

The goal of this chapter is to exemplify and illustrate the importance of protein–protein interactions for human biology and pathology. We will explore how protein interaction networks contribute to the understanding of phenotype and human genetic disorders, how they can help the gene-finding process and how they bring a molecular framework to the genetic heterogeneity and pleiotropy often observed in human syndromes. Traditionally, protein–protein interaction studies have focused on a few, selected gene products in a particular physiological context. Recently, more holistic strategies have been executed with the aim of understanding global protein–protein interaction networks of a cell or organism. An important section of this chapter is dedicated to a review of the proteomic technologies developed for the study of protein–protein interactions on a proteome-wide scale. Their respective advantages, shortcomings and limitations are discussed. From a simplistic point of view, the analysis of protein complexes or interactions should contribute to the reconstruction of disease-relevant or pharmacologically relevant pathways at molecular levels and to the identification of novel drug targets. It should also help to understand the mechanism of action and side effects of therapeutic compounds. Recent successes in the development of small molecules that target and interfere with protein–protein interactions open new avenues towards the design of drugs that specifically abrogate or modulate disease-relevant interactions. A discussion of this will constitute the last topic of the chapter.

8.2 Protein–Protein Interactions in Human Diseases: Altered Protein Connectivity Leads to Disorder

Over 1,500 human disease genes have been identified and more than 2,000 monogenic syndromes have been catalogued in the Online Mendelian Inheritance in Man database (OMIM Online Mendelian Inheritance in Man 2005). These provide a rich source of functional and phenotypic data.

Interestingly, syndromes are rarely the consequence of a complete gene knockout. Even if nonsense or missense mutations occasionally lead to the expression of very unstable proteins, a partially or totally non-functional product may still accumulate. Mutations that affect protein–protein interactions are not uncommon. Table 8.1, though far from comprehensive, provides some well-studied examples. Mutations in receptors that affect interactions with the cognate peptide ligands are evident. For example, mutations in the fibroblast growth factor receptor 2 (FGFR2) selectively increase its affinity for FGF2 (Anderson et al. 1998) and lead to Apert's syndrome. This is characterised by skull malformation, syndactyly and mental deficiency. Other instances concern mutations of proteins that function as scaffolds or enzyme modulators but possess no intrinsic catalytic activity. Familial

Table 8.1 Altered protein–protein interaction in human diseases

Disease and syndrome	Mutated gene product	Interacting partner	References
Apert syndrome	Fibroblast growth factor receptor 2: FGFR2	FGF	Anderson et al. (1998)
Familial melanoma	Tumour-suppressor gene: p16 (INK4)	Cyclin-dependent kinases (CDK4, CDK6)	Yarbrough et al. (1999); Cammett et al. (2003)
CADASIL	NOTCH3	NOTCH3, Fringe	Zweifel et al. (2003); Arboleda-Velasquez et al. (2005)
Bare lymphocyte syndrome	RFXANK	RFX complex	Wiszniewski et al. (2003)
Branchio-oto-renal/branchio-otic syndromes	SIX1	EYA1	Ruf et al. (2004)
Adrenoleukodystrophy	ATP-binding cassette transporter: ABCD1	ABCD1	Zhou (2004)
Holt–Oram syndrome	Tbx5	NKX2.5	Fan et al. (2003)
ICF syndrome	Methytransferase gene: DNMT3B	Unknown	Chen et al. (2004)
Giant axonal neuropathy	Gigaxonin	MAP1B-LC	Ding et al. (2002)
Hereditary nonpolyposis colorectal cancer	Mismatch repair gene: MLH1	PMS2	Guerrette et al. (1999)

melanoma is caused by mutations in the tumour suppressor, p16(INK4), that preclude its association with the cyclin-dependent kinases (CDK4, CDK6) (Yarbrough et al. 1999). Enzymes can also be involved in protein–protein interactions. For instance, DNMT3B is a DNA methyltransferase implicated in the ICF (immunodeficiency, centromeric instability, facial anomalies) syndrome, a rare autosomal recessive disorder. Missense mutations characterised in ICF patients map within the catalytic site and also affect an N-terminal PWWP domain, involved in protein–protein interactions (Shirohzu et al. 2002). Further mutations have been characterised that prevent the assembly of functional multiprotein complexes. A good example is an RFXANK gene mutant that fails to assemble the regulatory factor X complex (an obligate transcription factor required for the expression of MHC class II genes), leading to the bare lymphocyte syndrome (Wiszniewski et al. 2003). Finally, a variety of abnormal or erroneous interactions between brain proteins can result in the formation of toxic aggregates of proteinaceous fibrils. These 'fatal attractions' (Trojanowski and Lee 2000) are associated with a variety of neurodegenerative disorders, such as sporadic and familial Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis and prion encephalopathies.

These selected examples by no means represent a comprehensive inventory. Rather they illustrate that the spatial and temporal orchestration of the many enzyme activities, required for the proper functioning of the cell, through extensive and highly regulated protein–protein interaction networks bears remarkable functional relevance. Mutational lesions or environmental factors impairing the pathway flow or deregulating connections lead to abnormalities, as surely as interferences with the catalytically active sites.

8.3 Charting Protein–Protein Interactions

Proteins and their interactions have traditionally been studied one by one. These detailed, small-scale studies have elucidated the function of many proteins. Furthermore, a wide variety of modular binding domains with specificity for distinct sequence motifs have been mapped, contributing to our understanding of the puzzle of protein–protein interaction (Pawson et al. 2002). Whilst useful, decades of one-at-a-time approaches have contributed functional information for only 5–10% of all predicted genes (Vidal 2005). More recently, the elucidation of the full repertoire of genes in several organisms and significant breakthroughs in the fields of functional genomics and proteomics (see Chaps. 2–7) have paved the road for holistic protein–protein interaction analyses. Two main types of approaches have been adopted (Table 8.2). First is the yeast two-hybrid system that allows the mapping of binary or pairwise associations. Second is affinity-purification methods

Table 8.2 Overview of large-scale protein–protein interaction studies. Note that some studies have analysed interactions between all proteins, whilst others have analysed protein interactions in specific pathways

Method	Organism (proteins in proteome)	Interactions	References
Yeast two-hybrid	*Helicobacter pylori* (1,590)	~1,520	Rain et al. (2001)
	S. accharomyces cerevisiae (6,500)	~4,500	Ito et al. (2001)
	S. cerevisiae (6,500)	~1,000	Uetz et al. (2000)
	Caenorhabditis elegans (~20,000)	~5,000	Li et al. (2004)
	Drosophila melanogaster (~14,000)	~20,400	Giot et al. (2003)
	D. melanogaster (~14,000)	~2,300	Formstecher et al. (2005)
Affinity chromatography			
Tandem affinity purification /mass spectrometry	*S. cerevisiae* (6,500) *Escherichia coli* (4,300) *H. sapiens* (~30,000)	~4,100 ~5,250 ~220	Gavin et al. (2002) Butland et al. (2005) Bouwmeester et al. (2004)
Immunoaffinity purification/mass spectrometry	*S. cerevisiae* (6,500)	~3,620	Ho et al. (2002)
Luminescence-based mammalian interactome mapping	*Homo sapiens* (~30,000)	~950	Barrios-Rodiles et al. (2005)
Others			
Protein microarrays	*S. cerevisiae* (6,500)	~40	Zhu et al. (2001)

coupled to mass-spectrometric protein identification, designed for the characterisation of protein complexes. Alternative strategies are also emerging, like protein chips and luminescence-based mammalian interactome mapping (LUMIER). Fluorescence-based interaction assays such as fluorescence resonance energy transfer (FRET), hold great promise as they allow the monitoring of protein–protein interaction inside a living cell. These technologies will not be discussed as they have not yet been applied on a large-scale; however, they have been recently reviewed (Presley 2005; Wallrabe and Periasamy 2005). Finally, it is worth noting that a number of databases have been developed that integrate protein–protein interaction data from sources such as large-scale physical or genetic interactions and literature mining. These databases, outlined in Table 8.3, provide a very rich source of information.

Table 8.3 Protein–protein interaction databases available on the Internet and their URLs. See also The Jena Centre for Bioinformatics (2005)

Database	Internet URL	Species	Interaction
The GRID	http://biodata.mshri.on.ca/grid/servlet/Index	*C. elegans*, *D. melanogaster*, *S. cerevisiae*	G, B, C, P
SGD	http://www.yeastgenome.org/	*S. cerevisiae*	G, B, C, P
CYGD	http://mips.gsf.de/genre/proj/yeast/	*S. cerevisiae*	G, B, C
BIND	http://bind.ca/	*Arabidopsis thaliana*, *Bos taurus*, *C. elegans*, *D. melanogaster*, *Gallus gallus*, *H. pylori*, *HIV1*, *H. sapiens*, *Rattus norvegicus*, *S. cerevisiae*, *Mus musculus*	B, C, P
HPRD	http://www.hprd.org/	*H. sapiens*	B, C, P
DIP	http://dip.doe-mbi.ucla.edu/	*C. elegans*, *D. melanogaster*, *E. coli*, *H. pylori*, *H. sapiens*, *M. musculus*, *R. norvegicus*, *S. cerevisiae*	B, C
MINT	http://160.80.34.4/mint/index.php	*B. taurus*, *C. elegans*, *D. melanogaster*, *E. coli*, *H. pylori*, *H. sapiens*, *M. musculus*, *R. norvegicus*, *S. cerevisiae*	B
IntAct	http://www.ebi.ac.uk/intact/index.jsp	*C. elegans*, *D. melanogaster*, *E. coli*, *H. sapiens*, *M. musculus*, *S. cerevisiae*	B
Biocarta	http://www.biocarta.com/	*H. sapiens*, *M. musculus*	P
MPPI	http://mips.gsf.de/proj/ppi/	Mammals	B
PathCalling	http://curatools.curagen.com/cgi-bin/com.curagen.portal.servlet.PortalYeastList	*S. cerevisiae*	G, B

G genetic, *D* domain, *B* binary, *C* complexes, *P* pathways.

8.3.1 Characterisation of All Coding Sequences in an Organism

All strategies for the large-scale study of protein–protein interactions rely on the introduction of a tag, a short peptide or protein domain, in frame with the proteins of interest. This allows for affinity-based purifications (affinity purification and mass-spectrometric approaches, protein chips) and/or for the detection and the monitoring of protein–protein interactions (yeast two-hybrid, fluorescence-based approaches). In all cases, the proteome-wide analysis of protein–protein interactions in an organism requires the prior identification of all coding sequences or open reading frames (ORFs), their cloning and/or manipulation. For a relatively simple organism such as *Saccharomyces cerevisiae*, which has very little alternative splicing, the strategy is relatively straightforward and involves genome-scale PCR amplification directly from genomic DNA (Hudson et al. 1997). For higher organisms, which have very complex intron–exon structures, the source material for cloning is a comprehensive complementary DNA (cDNA) library. Success thus depends on the presence of a full-length cDNA for the gene of interest. As a consequence, low-abundance or very long messenger RNAs (mRNAs) are usually subject to higher failure rates. Importantly, only a single spliced mRNA is usually chosen as a representative of each ORF. Accordingly, genome-wide cloning in higher organisms entails the loss of some of the natural diversity of proteins isoforms.

Several efforts are currently under way aiming at the production of non-redundant collections of ORFs in various organisms. These includes the *Caenorhabditis elegans* ORFeome project (Lamesch et al. 2004), the Berkeley *Drosophila melanogaster* Genome Project (DGP) (Stapleton et al. 2002) as well as several efforts to generate human and mouse gene collections such as the Mammalian Gene Collection (MGC) (Gerhard et al. 2004), the Unigene set (Bussow et al. 2000) the Full-Length Expression (FLEXGene) repository (Brizuela et al. 2001, Brizuela 2002) and the Integrated Molecular Analysis of Genomes and their Expression (The IMAGE Consortium 1999) cDNA collection.

8.3.2 Monitoring Binary Interactions: the Yeast Two-Hybrid System

The yeast two-hybrid system is a genetic, ex vivo assay that allows the discovery of binary protein–protein interactions. Its principle relies on the modular nature of many transcription factors that contain both a site-specific DNA binding domain (DBD) and a transcriptional activation domain (AD) that recruits the transcriptional machinery to the promoters (Fig. 8.1). The interaction between a 'bait' fusion protein (hybrid protein X-DBD) and a 'prey' fusion protein (hybrid protein Y-AD) reconstitutes a functional transcription factor which turns on the expression of a reporter gene or selectable marker (Fields and Song 1989).

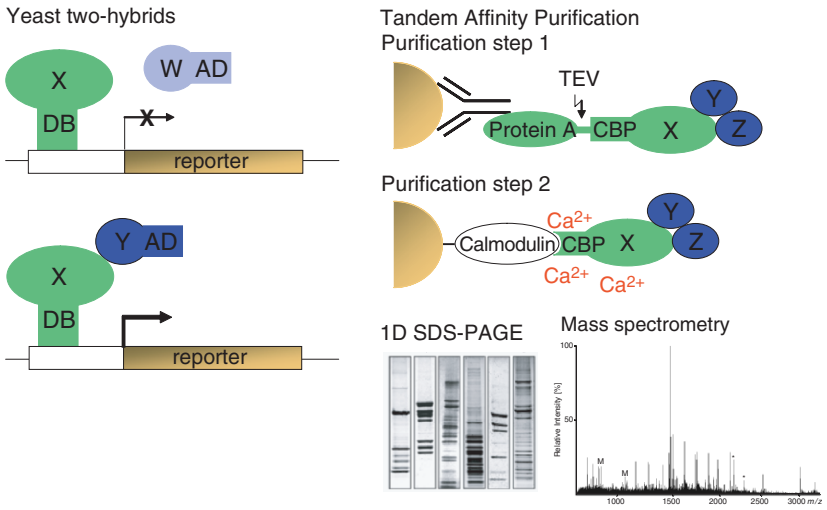


Fig. 8.1 *Left*: The yeast two-hybrid system relies on the modular nature of many transcription factors that contain both a site-specific DNA binding domain (*DB*) and a transcriptional activation domain (*AD*). The interaction between a 'bait' fusion (protein X-DB hybrid protein) and a complementary 'prey' fusion (protein Y-AD hybrid protein) reconstitutes a functional transcription factor which turns on the expression of the reporter gene. Note that the 'prey' fusion W-AD was not complementary, and the reporter gene was not expressed. *Right:* The tandem affinity purification (TAP) protocol sequentially utilises two epitope tags, protein A and calmodulin binding peptide (*CBP*). The TAP-fusion protein *X* is expressed in cells and a protein complex can assemble under physiological conditions with the endogenous components (*Y*, *Z*). The protein complex is purified by a two-step affinity purification: firstly, via immunoglobulin resin that binds to the protein A tag and secondly via calmodulin resin that binds the CBP tag in the presence of calcium. The protein complexes can finally be analysed by one dimensional sodium dodecyl sulfate polyacrylamide gel electrophoresis (*1D SDS-PAGE*) or by mass spectrometry. *TEV Tobacco etch virus*

The yeast two-hybrid system is very economical, scalable and its use rapidly evolved to genome-wide strategies. Several experimental settings have been successfully exploited for high-throughput approaches. The most elegant, but also the most labour-intensive, is the systematic one-by-one testing of all possible protein combinations. For instance, in yeast, 192 bait protein fusions have been screened against all 6,000 yeast prey protein fusions (Uetz et al. 2000). Other approaches have screened each bait protein against a library of pooled cloned prey proteins (Giot et al. 2003; Li et al. 2004; Rain et al. 2001; Uetz et al. 2000). Finally, a many-to-many strategy has been used which involves the screening of pools of bait proteins against pools of prey proteins (Ito et al. 2001). Compared with the one-by-one approach, the screening of pools of bait or prey requires the sequencing of the plasmids to identify the interacting protein partner. It is however, by far the most economical, allows the highest throughput, but usually yields fewer interactions.

The key advantage of the two-hybrid system, besides scalability and flexibility, is its capacity to detect transient interactions and those that are very weak. This includes interactions with dissociation constants down to 10^{-7} M, which correspond to the weakest interactions that occur within the cell. The system, however, has drawbacks due to its ex vivo nature. Expressed fusion proteins are forced to the nucleus, which may not be where they are usually localised. Membrane proteins, for example, are usually not compatible with such a nuclear-based assay. The bait or prey proteins may not undergo appropriate post-translational modification. Similarly, interactions that involve co-operative, allosteric events or chaperone-assisted assembly may not occur correctly in the nucleus. The yeast two-hybrid system measures the physical propensity or potential of different proteins to 'fit' or interact. Because of its ex vivo nature, however, this does not necessarily reflect physiological events; in vivo under physiological condition the proteins may be expressed at different times, in different tissues or organelles. Finally, transcription factors, as well as other proteins (about 5–10% of gene products), can autoactivate the transcription of reporter genes. Such proteins are thus unsuitable to the approach.

Several modified versions of the yeast two-hybrid system have been developed that address some of the abovementioned drawbacks. They involve the reconstitution of modular proteins other than transcription factors that enable the analyses of proteins not amenable to the 'classical' two-hybrid assay (essentially membranes and transcription factors). they include the SOS (Aronheim et al. 1997) or the Ras recruitment systems (Broder et al. 1998), the G-protein-based screening assay (Ehrhard et al. 2000) and the split-ubiquitin system (Stagljar et al. 1998). Although some of these assays are apparently robust (Iyer et al. 2005), none of them have yet been used in high throughput proteome-wide screens.

8.3.3 Analysis of Protein Complexes by Affinity Purification and Mass Spectrometry

A critical regulatory level in the interpretation of genetic information lies within the dynamic assembly and disassembly of protein complexes, or molecular machines. This must occur at an appropriate time and space to serve the necessary cellular tasks. Assembly, for example, often requires energy-driven conformational changes, specific post-translational modifications, chaperone-assistance and particular order of addition of individual components (Goh et al. 2004). Because of this, it is not easy to extrapolate the composition of protein complexes on the basis of the sum of the individual binary interactions (Aloy and Russell 2002; Bader and Hogue 2002; Edwards et al. 2002; Jansen et al. 2002; Kemmeren et al. 2002; von Mering et al. 2002). Characterisation of protein complexes traditionally requires their purification from their living environment, the cell.

The emergence of sensitive and high-throughput mass spectrometry methods, allowing the detection and fragmentation of peptides in the lower femtomolar range (see Chaps. 3, 5), has fuelled the development of methods employing the biochemical purification of whole cellular assemblies. Pioneering work used specific antibodies combined with mass spectrometry to identify protein complexes in various cells and tissues. One example is the protein assembled around the neurotransmitter receptor from the mouse synapse (Husi et al. 2000; Becamel et al. 2002). Importantly, antibodies allow the retrieval of endogenously expressed, native proteins from cells or even tissues, in very close to physiological conditions. These approaches however, have been traditionally limited by the availability of specific antibodies or other affinity-capturing agents.

Generic approaches for the purification of protein complexes are available that use affinity techniques to capture 'bait' proteins fused to an epitope tag. An antibody directed against the tag, instead of the bait protein, is typically used. A variety of epitope tags have been developed in the past, such as Myc, HA, Flag, KT3 and GST. Corresponding antibodies or affinity resins (e.g. anti-Myc, anti-HA, anti-Flag, anti-KT3, glutathione) are commercially available. Many different ORFs can be fused to the same epitope tag in parallel, expressed in the appropriate cell type and retrieved using the same affinity resins. The protein components of the purified protein complexes are identified either by direct tandem mass spectrometry or by peptide mass fingerprinting after one-dimensional gel electrophoresis (see Chap. 3).

Affinity capture approaches have been recently adapted to proteome-wide analysis of protein complexes. Ho et al. (2002) used a one-step immunoaffinity purification strategy based on the Flag epitope tag and anti-Flag system. They characterised proteins assembled around 725 yeast bait proteins involved in the DNA-damage response and in cell signalling (kinases, phosphatases and regulatory subunits). The analysis identified interactions between 1,578 different proteins.

In the purification of protein complexes, techniques are required that bring high discriminatory power against any non-specific protein background whilst purifying the true components of a protein complex. Unfortunately, traditional high-stringency washing steps that involve high concentrations of salt or detergents are not always compatible with the preservation of integrity of a protein complex. To address these issues, the tandem affinity purification (TAP) method was developed. It sequentially utilises two epitope tags, *Staphylococcus* protein A and calmodulin-binding peptide (CBP), instead of just one (Rigaut et al. 1999; Puig et al. 2001). The TAP-fusion protein, which acts as the bait, is expressed in the cells (Fig. 8.1). A protein complex can assemble under physiological conditions with its endogenous components. The tagged protein, along with associated partners, is retrieved by two steps of affinity purification. First, an immunoglobulin resin is used to bind to the protein A tag. The protein complex is specifically eluted from the immunoglobulin resin by protease cleavage, using *Tobacco*

etch virus (TEV) protease. Importantly, the TEV protease cuts very specifically at a seven amino acid sequence that has been introduced between the protein A and the CBP tags. The TEV cleavage sequence is only found in a few human, mouse or yeast proteins, ensuring that proteins in the retrieved complexes are not digested. In a second affinity step, the complex is immobilised to calmodulin-coated beads via the CBP tag. This step removes the TEV and further contaminants. As the CBP–calmodulin interaction is calcium-dependent, a second specific elution step is achieved through the removal of calcium with a chelating agent (ethylene glycol tetraacetic acid).

The suitability of TAP/mass spectrometry for proteome-wide analyses was first demonstrated in yeast, where 231 protein complexes were characterised from the analysis of 1,739 yeast genes (Gavin et al. 2002). Of them, 134 protein complexes were new. The approach was subsequently applied to a number of high-throughput analyses in a variety of cell types, including the bacterium *Escherichia coli* (Butland et al. 2005) and human (Bouwmeester et al. 2004).

The key advantage of the affinity purification/mass-spectrometric approach is that complexes are produced in vivo, under conditions which are very close to their normal physiology. The actual molecular assemblies form in the appropriate cellular compartment. As the approach is not limited to one cell type, it is possible to monitor changes in complex composition in different cell lines, during development or following various cell treatments. Importantly, only one protein in a complex is cloned and tagged. All other components of the complex are native and will reflect the natural diversity of protein isoforms (see Chap. 5). Among the affinity-purification approaches, TAP has proven generally applicable to proteome-wide screens. The protein complexes are kept under native conditions throughout the purification procedure. As it uses two steps of purification, it is efficient at reducing non-specific protein binding, and overly stringent wash conditions can be avoided. Accordingly, TAP-purified complexes have been successfully used for other applications such as electron microscopy studies (Aloy et al. 2002, 2004).

The main drawbacks of TAP are that affinity purification/mass-spectrometric methods are not generally designed to monitor very labile or transient interactions (generally K_d is in the mid-nanomolar region or below; unpublished data). In addition, the fusion with an epitope tag may sometimes interfere with the biological function of the tagged protein through modification of protein folding, its recruitment within a protein complex or its subcellular localisation. These risks can be significantly reduced by creating and analysing both N-terminal and C-terminal protein fusions in parallel. Finally, in systems or organisms where homologous recombination is not efficient, such as in mammalian cells, the TAP tag cannot be directly fused to the targeted proteins at its chromosomal locus. In such cases, the TAP fusion is usually overexpressed from a plasmid. This can lead to aberrant localisation, protein aggregation or toxicity. Generally, tight controls over the expression levels and the bait-protein localisations should be included in any systematic screen (Bouwmeester et al. 2004).

8.3.4 Luminescence-Based Mammalian Interactome Mapping

LUMIER is the most recent adaptation of the affinity-purification approaches discussed in previous sections. It is based on the co-expression of a *Renilla* luciferase (RL) fusion protein together with individual Flag-tagged candidate partners. The interactions are determined by performing an RL enzymatic assay on immunoprecipitates generated with an anti-Flag tag antibody. A proof of principle analysis involved ten core members of the TGF-β signalling pathway fused to RL and 518 mouse cDNAs fused to Flag (Barrios-Rodiles et al. 2005). The method was readily adapted and optimised in a 96-well format. An interesting aspect of this strategy is the possibility to derive semi-quantitative information on the intensity of the interaction.

8.3.5 Protein Microarrays

In recent years, it has become routine to use DNA microarrays to probe the expression of thousands of genes in parallel. Similarly, protein microarrays have been developed to provide a means for the rapid and parallel screening of large numbers of proteins for biochemical activities, protein–protein interactions, protein–lipid, protein–nucleic acid and protein–small chemical interactions. The first high-density proteome microarray consisted of 5,800 GST-His tag fusion proteins for yeast (Zhu et al. 2001). Such protein microarrays are currently commercially available. Details of the procedure have been extensively reviewed elsewhere (Schweitzer et al. 2003; Poetz et al. 2005b).

In principle, protein microarrays are amenable to proteome-wide screens for protein–protein interactions; however, they have only been used to date for the identification of new calmodulin- and phospholipid-binding proteins, and to monitor domain–domain and antigen–antibody interactions (Zhu et al. 2001; Espejo et al. 2002; Hiller et al. 2002; Poetz et al. 2005a).

8.3.6 Data Quality

As many interactions in large-scale studies are not rigorously quality controlled, large-scale approaches may suffer from the pitfalls of spurious interactions (false positives) and undetected genuine associations (false negatives). Recently, several analyses have attempted to evaluate the quality of the various protein–protein interaction datasets either by integrating additional functional annotation or through comparison with reference protein interaction sets (Table 8.2). It was thus estimated that the false-positive rate for two-hybrid data was 50–90% and was about 30% for mass-spectrometric approaches (Mrowka et al. 2001; Gavin et al. 2002; Sprinzak et al. 2003). A more optimistic outlook was obtained in recent analyses, where 65% of the two-hybrid interactions were positively confirmed by protein co-precipitation experiments (Li et al. 2004).

Another important observation is the remarkably low overlap observed between data from the various large-scale analyses of protein–protein interactions. For example, comparison of the two largest two-hybrid studies performed in yeast to date revealed that only 15–20% of the high-confidence, core interactions were identified in both datasets (Ito et al. 2001; Bader and Hogue 2002; von Mering et al. 2002). Intrinsic differences in the experimental setup, for example different stringency criteria or screen design (one-by-one screens versus pools of baits/preys) are expected to contribute to the heterogeneity of the results. Only 7% overlap between yeast two-hybrid and TAP/mass spectrometry datasets (Gavin et al. 2002) is probably not so surprising. As already outlined, the two approaches monitor different properties of proteins, namely their ability to interact (yeast two-hybrid) and their ability to stably integrate in protein complexes (mass spectrometry based techniques). Both methods are highly complementary and should be applied in combination to produce comprehensive protein–protein interaction maps of a given cell. Finally, the fact that none of the numerous large-scale analyses of yeast protein–protein interactions have redefined more than 50% of the known interactions (von Mering et al. 2002) indicates that interaction screens have to date not been run to saturation. Even for the relatively simple and extensively studied organism *S. cerevisiae*, many more interactions remain to be discovered.

8.4 Biological and Biomedical Applications

8.4.1 Charting of Diseases and Pharmacologically Relevant Pathways

The majority of protein targets for current therapeutics cluster in a limited number of cellular pathways (Brown and Superti-Furga 2003). The selective elucidation of the pathways central to human diseases is expected to provide a molecular framework for the interpretation of genetic links and help discover new targets with higher chemical tractability.

It has long been recognised that mutations in different genomic loci might lead to similar or related human syndromes. It is generally assumed that this so-called genetic heterogeneity reflects disruptions in proteins that participate in a common interaction network, such as ligand–receptor interactions, the different subunits of a multiprotein complex or proteins that function at different steps of a signalling or biochemical pathway (Brunner and van Driel 2004). A striking example is provided by the nine genes associated with the Fanconi anaemia (FA) syndrome. Subsequent to their identification, the nine genes were shown to cooperate in a common network, the FA/BRCA2 pathway involved in DNA repair. Notably, seven of the FA proteins form a multiprotein complex (Mace et al. 2005).

Generally, our appreciation of the pathways central to human diseases remains scant. As our knowledge becomes more comprehensive, one might expect a remarkable synergy to be seen between the interactome and the phenome. A thorough understanding of the protein–protein interaction networks formed around a mutated gene product is expected to focus efforts in identifying new genes involved in similar syndromes. Reciprocally, the clustering of genes leading to common or similar phenotypes may guide the reconstruction and interpretation of interaction networks.

Several recent efforts have generated comprehensive views of the proteins that are functionally involved in pathways associated with major human abnormalities. For instance, Tewari et al. (2004) combined the yeast two-hybrid approaches with systematic multiple-perturbation methods to systematically study the TGF-β signalling pathway in *C. elegans*. TGF-β signalling is involved in a variety of diseases, such as familial primary pulmonary hypertension, hereditary chondrodysplasia and tumour-predisposition syndromes (Massague et al. 2000). The analysis discovered eight new TGF-β pathway modulators. In human, the systematic mapping, by TAP/mass spectrometry, of the protein interaction network around 32 components of the proinflammatory TNF-α/NF-κ B signalling cascade led to the identification of 221 molecular associations. The analysis of the network, and directed functional perturbation studies using RNA interference, highlighted ten new functional modulators that provided significant insight into the logic of the pathway as well as new candidate targets for pharmacological intervention (Bouwmeester et al. 2004).

8.4.2 Lessons Learned from Global Interaction Analyses in Yeast

It has long been known that proteins can have more than one cellular function. At genetic levels, such phenomena have been proposed to account for gene pleiotropy. This occurs when different mutations in a single gene cause multiple phenotypes or diseases with seemingly unrelated symptoms (Hodgkin 1998). A good illustration is provided by mutations in the *XPD* (ERCC2) gene that cause either xeroderma pigmentosum (XP), a DNA-repair disorder, or trichothiodystrophy (TTD), a disease characterised by mental retardation, unusual faces, ichthyonic skin and a reduced stature. XPD is a subunit of the TFIIH basal transcription factor complex implicated in both transcriptional regulation and DNA repair. The two functions of the TFIIH complex involve the selective recruitment of specific factors (Egly 2001). The most parsimonious explanation and current hypothesis is that different mutations in *XPD* affect either the DNA-repair function of TFIIH or its transcriptional role, resulting in XP or TTD, respectively (Lehmann 2001). Table 8.4 illustrates how protein multifunctionality can be explained by the tendency of proteins to associate with different binding partners in various contexts (organelles, cell types or tissues). They can therefore have diverse specificities, biochemical activities and functions.

Table 8.4 Example of multifunctional proteins

Protein	Complexes	Biological function	References
Yotiao (scaffold for PKA and PP1)	NMDA receptor	Excitatory synaptic transmission, brain development, learning and memory	Marx et al. (2002)
	K^+ channel KCNQ1/KCNE1	β adrenergic receptor-dependent regulation of cardiac action potential	
TRRAP	TRRAP/TIP60 histone acetylase	Promoter-specific histone acetylation	Park et al. (2001)
	PCAF complex (Spt-Ada-Gcn5-acetyl-transferase)	Promoter-specific histone acetylation	
Hsp70 (chaperones)	Hsp70/Hsp90 multichaperone machinery	Protein folding	Johnson et al. (2002)
	Class I histone deacetylases	Nucleosome remodeling	
Hsp90 (chaperones)	Hsp70/Hsp90 multichaperone machinery	Protein folding	Forsythe et al. (2001)
	p23(co-chaperone)/ telomerase	Telomere maintenance	
Lipoamide dehydrogenase	Branched-chain ketoacid dehydrogenase	Amino acid catabolism	Odievre et al. (2005)
	Pyruvate dehydrogenase	Glycolysis	
	α-Ketoglutarate dehydrogenase	Tricarboxylic acid cycle	
Phosphofructokinase	Phosphofructokinase	Glycolysis and gluconeogenesis	Vertessy et al. (1997)
	Tubulin	Organisation of microtubules	
Glucokinase	Glucokinase-regulatory protein complex	Blood glucose homeostasis	Danial et al. (2003)
	PKA/PP1/Wiskott–Aldrich family member WAVE-1/ glucokinase	Orchestrates glycolysis and apoptosis	

A recent genome-wide analysis of protein complexes in *S. cerevisiae* raised the view that protein multifunctionality may be a more general phenomenon than initially thought (Gavin et al. 2002; Gavin and Superti-Furga 2003). Different protein complexes appear to use the same protein to execute different biological functions. Strikingly, about 37% of proteins were found in more than one protein complex (Krause et al. 2004). Proteins, just like the domains they are made of, seem to be used in a combinatorial manner in a variety of 'molecular machines'. Protein modularity or multifunctionality has been proposed to support parsimonious increases in organismal complexity even with a relatively constant number of genes. The understanding of protein modularity in higher eukaryotes may not only contribute a molecular framework for the explanation of genetic traits, like genetic pleiotropy, but is also expected to contribute to the selection of more specific and safer drug targets.

8.4.3 An Emerging Application: the Development of Small-Molecule Protein–Protein Interaction Inhibitors

The current generation of drugs is designed to target the active sites of enzymes or the ligand-binding pockets of membrane-bound receptors. From a purely conceptual point of view, the pharmacological interference with protein–protein interactions may provide an alternative and probably more subtle means for the modulation of a discrete protein function. A key feature of protein interactions is their variety, and it is this that makes them attractive as new drug targets. However, unlike the interactions of enzymes with substrates, protein–protein interactions usually do not occur tightly or in deep ligand-binding pockets. Instead, they involve large, flat surfaces. For many years, in contrast to known exceptions involving natural products or peptides and peptidomimetics (Sillerud and Larson 2005; Table 8.5), the prevailing view was that protein–protein interactions would continue to be difficult to target, especially with conventional small-molecule chemistry.

In recent years, an increasing number of publications have reported the successful modulation of protein–protein interactions using traditional small molecules (Table 8.5). These have either directly targeted the protein–protein interface or acted through the binding of an allosteric site (Pagliaro et al. 2004). These successes mainly resulted from the development of sensitive high-throughput screening assays for the detection of protein interactions, such as luminescence resonance energy transfer (Bergendahl et al. 2003), variations of the two-hybrid system (Zhao et al. 2004) or fluorescence polarisation assays. For example, Degterev et al. (2001) devised an ingenious high-throughput fluorescence polarisation-based screening strategy that monitored the displacement of fluorescently labelled Bak-BH3 peptide from Bcl-X_L protein in solution. They identified two classes of compounds (BH3Is) that disrupted the Bak-BH3/Bcl-X_L interaction with affinities in the low micromolar range (Table 8.5).

Table 8.5 Drugs and small molecules that interfere with protein–protein interactions

Targets	Small molecule/drug	Stage	References
Natural products and natural-product-like			
FKBP12/mTOR	Rapamycin	Approved	Huang et al. (2003)
Calcineurin/FK506-binding protein	FK506	Approved	Jin and Harrison (2002)
Calcineurin/cyclophilin	Cyclosporin	Approved	Jin and Harrison (2002)
€Tubulin/Tubulin	Vinblastine	Approved	Wilson et al. (1999)
SH3-mediated interactions	UCS15A	Discovery	Oneyama et al. (2002)
Tcf4/β-catenin	PKF118-310 (IC_{50} 0.8 μM), PKF115-584 (IC_{50} 3.2 μM)	Discovery	Lepourcelet et al. (2004)
Peptide and peptidomimetics			
Myc/Max	IIA4B20 (IC_{50} 75 μM), IIA6B17 (IC_{50} 50 μM)	Discovery	Berg et al. (2002)
Stat3/Stat3	ISS 610 (IC_{50} 50 μM)	Discovery	Turkson et al. (2004)
Synthetic small molecules			
MDM2/p53	Nutlin-3 (IC_{50} 0.09 μM)	Discovery	Vassilev et al. (2004)
	Norbornane derivatives		Buolamwini et al. (2005)
	Synthetic chalcones		Buolamwini et al. (2005)
	Pyrazolidinedione sulfonamide		Buolamwini et al. (2005)
	1,4-Benzodiazepine-2,5-diones		Buolamwini et al. (2005)
LFA1/sICAM1	BIRT377 (K_d 25.8 nM)	Discovery	Kelly et al. (1999)
Bak-BH3/Bcl2	HA14 (IC_{50} 9 μM), BH3I-1 (K_d 2.4–12.5 μM); BH3I-2 (K_d 4.1–6.4 μM)	Discovery	Wang et al. (2000); Degterev et al. (2001)
MMP2/integrin α(V)β(3)	TSRI265	Discovery	Silletti et al. (2001)
SH3-mediated interactions	2-Aminoquinolines (K_d 20 μM)	Discovery	Inglis et al. (2004)
CRM1/Rev-NES	PKF050-638	Discovery	Daelemans et al. (2002)
ESX/Sur-2	Wrenchnolol	Discovery	Shimogawa et al. (2004)
CREB/CBP	KG-501	Discovery	Best et al. (2004)
NGF/p75	PD90780	Discovery	Colquhoun et al. (2004)
I L-2/IL-2Rα	Sulfamide- and urea-based small-molecule antagonists (IC_{50} 0.6 μM)	Discovery	Waal et al. (2005)

Advances in the structural biology of protein–protein interactions have also facilitated the emergence of new strategies for the design and the development of small-molecule inhibitors of protein interactions. These include computer-based screenings, which led to the identification of the HA14 compound that blocks the association between Bak-BH3 and Bcl2 (Wang et al. 2000); the surface activity relationships by NMR – the focused targeting of interaction 'hotspots' that are believed to make particularly important contribution to the energy of protein–protein interaction (Bogan and Thorn 1998); the fragment-based drug-discovery approaches, where small organic molecule fragments with low affinities for the interaction interfaces are linked together so that their affinities are combined (Arkin and Wells 2004); and the use of bifunctional small molecules that sterically block interactions through targeted covalent modifications (Way 2000).

Although most of these technologies are very new and the compounds identified so far are of relatively low affinity, the recent advances suggest that the modulation of protein–protein interactions may become a viable alternative for the design of small-molecule inhibitors of protein targets.

8.5 Future Directions

Biology relies on the concerted molecular interaction of proteins operationally organised in complexes, pathways and networks. With the development of large-scale strategies for the analysis of protein–protein interactions, including yeast two-hybrid system and TAP/mass spectrometry, protein–protein interactions have taken a centre stage in the postgenomic era. Comprehensive cartographies of several pathways that map behind major human abnormalities have been established. Proteome-scale, protein–protein interaction screens have also been performed in model organisms and provide a molecular framework for the interpretation of simple genetic data, such as gene essentiality (Jeong et al. 2001). This has shown that proteins that are lethal when deleted have a greater tendency to be central in interaction networks. Notwithstanding this, there has been limited use of protein–protein interaction maps to predict the behaviour of whole and complex systems.

Currently, protein–protein interaction maps provide us with a relatively static wiring scheme of all physical associations. As they do not qualify the types of relationships between the various components, they still lack the information that describes the essential logic of the network. In the future, the systematic integration with functional data such as gene knockout and systematic perturbations should fill this gap and provide us with more meaningful maps of functional significance.

Protein-interaction networks are dynamic, and this dynamism is crucial for the function of the cell. However, large-scale analyses of protein

interactions have failed to address the dynamic nature of the molecular circuits to date. Many of the broadly used approaches for the monitoring of protein–protein interactions, although powerful, study the expression of proteins under non-physiological conditions. This is the case for ex vitro systems, where the regulation and the fine-tuning of the molecular interactions at cellular and physiological levels are usually lost. The monitoring of the dynamic recruitment of specific network components, in a spatial and temporal sense, requires their isolation from a normal physiological environment in the cell. In this respect, the TAP/mass-spectrometric approach has proven a very powerful method and shows promise for the systematic monitoring of loss or gain of pathway components following cellular perturbation. Ultimately, the further development of fluorescence-based interaction assays, such as FRET, for larger-scale quantification and monitoring of protein interaction within a living cell is also likely to provide striking insights into the biology.

Acknowledgement. I thank Albert Heck for critical reading of the manuscript.

References

Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92:291–294

Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 99:5896–5901

Aloy P, Ciccarelli FD, Leutwein C, Gavin AC, Superti-Furga G, Bork P, Bottcher B, Russell RB (2002) A complex prediction: three-dimensional model of the yeast exosome. EMBO Rep 3:628–635

Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L, Russell RB (2004) Structure-based assembly of protein complexes in yeast. Science 303:2026–2029

Anderson J, Burns HD, Enriquez-Harris P, Wilkie AO, Heath JK (1998) Apert syndrome mutations in fibroblast growth factor receptor 2 exhibit increased affinity for FGF ligand. Hum Mol Genet 7:1475–1483

Arboleda-Velasquez JF, Rampal R, Fung E, Darland DC, Liu M, Martinez MC, Donahue CP, Navarro-Gonzalez MF, Libby P, D'Amore PA, Aikawa M, Haltiwanger RS, Kosik KS (2005) CADASIL mutations impair Notch3 glycosylation by Fringe. Hum Mol Genet 14:1631–1639

Arkin MR, Wells JA (2004) Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. Nat Rev Drug Discov 3:301–317

Aronheim A, Zandi E, Hennemann H, Elledge SJ, Karin M (1997) Isolation of an AP-1 repressor by a novel method for detecting protein-protein interactions. Mol Cell Biol 17:3094–3102

Bader GD, Hogue CW (2002) Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol 20:991–997

Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, Donovan RS, Shinjo F, Liu Y, Dembowy J, Taylor IW, Luga V, Przulj N, Robinson M, Suzuki H, Hayashizaki Y, Jurisica I, Wrana JL (2005) High-throughput mapping of a dynamic signaling network in mammalian cells. Science 307:1621–1625

Becamel C, Alonso G, Galeotti N, Demey E, Jouin P, Ullmer C, Dumuis A, Bockaert J, Marin P (2002) Synaptic multiprotein complexes associated with 5-HT(2C) receptors: a proteomic approach. EMBO J 21:2332–2342

Berg T, Cohen SB, Desharnais J, Sonderegger C, Maslyar DJ, Goldberg J, Boger DL, Vogt PK (2002) Small-molecule antagonists of Myc/Max dimerization inhibit Myc-induced transformation of chicken embryo fibroblasts. Proc Natl Acad Sci USA 99:3830–3835

Bergendahl V, Heyduk T, Burgess RR (2003) Luminescence resonance energy transfer-based high-throughput screening assay for inhibitors of essential protein-protein interactions in bacterial RNA polymerase. Appl Environ Microbiol 69:1492–1498

Best JL, Amezcua CA, Mayr B, Flechner L, Murawsky CM, Emerson B, Zor T, Gardner KH, Montminy M (2004) Identification of small-molecule antagonists that inhibit an activator: coactivator interaction. Proc Natl Acad Sci USA 101:17622–17627

Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 280:1–9

Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C et al. (2004) A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. Nat Cell Biol 6:97–105

Brizuela L, Braun P, LaBaer J (2001) FLEXGene repository: from sequenced genomes to gene repositories for high-throughput functional biology and proteomics. Mol Biochem Parasitol 118:155–165

Brizuela L, Richardson A, Marsischky G, Labaer J (2002) The FLEXGene repository: exploiting the fruits of the genome projects by creating a needed resource to face the challenges of the post-genomic era. Arch Med Res 33:318–324

Broder YC, Katz S, Aronheim A (1998) The ras recruitment system, a novel approach to the study of protein-protein interactions. Curr Biol 8:1121–1124

Brown D, Superti-Furga G (2003) Rediscovering the sweet spot in drug discovery. Drug Discov Today 8:1067–1077

Brunner HG, van Driel MA (2004) From syndrome families to functional genomics. Nat Rev Genet 5:545–551

Buolamwini JK, Addo J, Kamath S, Patil S, Mason D, Ores M (2005) Small molecule antagonists of the MDM2 oncoprotein as anticancer agents. Curr Cancer Drug Targets 5:57–68

Bussow K, Nordhoff E, Lubbert C, Lehrach H, Walter G (2000) A human cDNA library for high-throughput protein expression screening. Genomics 65:1–8

Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. Nature 433:531–537

Cammett TJ, Luo L, Peng ZY (2003) Design and characterization of a hyperstable p16INK4a that restores Cdk4 binding activity when combined with oncogenic mutations. J Mol Biol 327:285–297

Chen T, Tsujimoto N, Li E (2004) The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin. Mol Cell Biol 24:9048–9058

Colquhoun A, Lawrance GM, Shamovsky IL, Riopelle RJ, Ross GM (2004) Differential activity of the nerve growth factor (NGF) antagonist PD90780 [7-(benzolylamino)-4,9-dihydro-4-methyl-9-oxo-pyrazolo[5,1-b]quinazoline-2-carboxylic acid] suggests altered NGF-p75NTR interactions in the presence of TrkA. J Pharmacol Exp Ther 310:505–511

Daelemans D, Afonina E, Nilsson J, Werner G, Kjems J, De Clercq E, Pavlakis GN, Vandamme AM (2002) A synthetic HIV-1 Rev inhibitor interfering with the CRM1-mediated nuclear export. Proc Natl Acad Sci USA 99:14440–14445

Danial NN, Gramm CF, Scorrano L, Zhang CY, Krauss S, Ranger AM, Datta SR, Greenberg ME, Licklider LJ, Lowell BB, Gygi SP, Korsmeyer SJ (2003) BAD and glucokinase reside in a mitochondrial complex that integrates glycolysis and apoptosis. Nature 424:952–956

Degterev A, Lugovskoy A, Cardone M, Mulley B, Wagner G, Mitchison T, Yuan J (2001) Identification of small-molecule inhibitors of interaction between the BH3 domain and Bcl-xL. Nat Cell Biol 3:173–182

Ding J, Liu JJ, Kowal AS, Nardine T, Bhattacharya P, Lee A, Yang Y (2002) Microtubule-associated protein 1B: a neuronal binding partner for gigaxonin. J Cell Biol 158:427–433

Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 18:529–536

Egly JM (2001) The 14th Datta lecture. TFIIH: from transcription to clinic. FEBS Lett 498: 124–128

Ehrhard KN, Jacoby JJ, Fu XY, Jahn R, Dohlman HG (2000) Use of G-protein fusions to monitor integral membrane protein-protein interactions in yeast. Nat Biotechnol 18:1075–1079

Espejo A, Cote J, Bednarek A, Richard S, Bedford MT (2002) A protein-domain microarray identifies novel protein-protein interactions. Biochem J 367:697–702

Fan C, Liu M, Wang Q (2003) Functional analysis of TBX5 missense mutations associated with Holt-Oram syndrome. J Biol Chem 278:8780–8785

Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. Nature 340:245–246

Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, Trehin A, Reverdy C et al. (2005) Protein interaction mapping: a *Drosophila* case study. Genome Res 15:376–384

Forsythe HL, Jarvis JL, Turner JW, Elmore LW, Holt SE (2001) Stable association of hsp90 and p23, but not hsp70, with active human telomerase. J Biol Chem 276:15571–15574

Gavin AC, Superti-Furga G (2003) Protein complexes and proteome organization from yeast to man. Curr Opin Chem Biol 7:21–27

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147

Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G Klein SL et al. (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res 14:2121–2127

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL et al. (2003) A protein interaction map of *Drosophila melanogaster*. Science 302:1727–1736

Goh CS, Milburn D, Gerstein M (2004) Conformational changes associated with protein-protein interactions. Curr Opin Struct Biol 14:104–109

Guerrette S, Acharya S, Fishel R (1999) The interaction of the human MutL homologues in hereditary nonpolyposis colon cancer. J Biol Chem 274:6336–6341

Hiller R, Laffer S, Harwanegg C, Huber M, Schmidt WM, Twardosz A, Barletta B et al. (2002) Microarrayed allergen molecules: diagnostic gatekeepers for allergy treatment. FASEB J 16:414–416

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415:180–183

Hodgkin J (1998) Seven types of pleiotropy. Int J Dev Biol 42:501–505

Huang S, Bjornsti MA, Houghton PJ (2003) Rapamycins: mechanism of action and cellular resistance. Cancer Biol Ther 2:222–232

Hudson JR Jr, Dawson EP, Rushing KL, Jackson CH, Lockshon D, Conover D, Lanciault C, Harris JR, Simmons SJ, Rothstein R, Fields S (1997) The complete set of predicted genes from *Saccharomyces cerevisiae* in a readily usable form. Genome Res 7:1169–1173

Husi H, Ward MA, Choudhary JS, Blackstock WP, Grant SG (2000) Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. Nat Neurosci 3:661–669

Inglis SR, Stojkoski C, Branson KM, Cawthray JF, Fritz D, Wiadrowski E, Pyke SM, Booker GW (2004) Identification and specificity studies of small-molecule ligands for SH3 protein domains. J Med Chem 47:5405–5417

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574

Iyer K, Burkle L, Auerbach D, Thaminy S, Dinkel M, Engels K, Stagljar I (2005) Utilizing the split-ubiquitin membrane yeast two-hybrid system to identify protein-protein interactions of integral membrane proteins. Sci STKE 275:13

Jansen R, Greenbaum D, Gerstein M (2002) Relating whole-genome expression data with protein-protein interactions. Genome Res 12:37–46

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42

Jin L, Harrison SC (2002) Crystal structure of human calcineurin complexed with cyclosporin A and human cyclophilin. Proc Natl Acad Sci USA 99:13522–13526

Johnson CA, White DA, Lavender JS, O'Neill LP, Turner BM (2002) Human class I histone deacetylase complexes show enhanced catalytic activity in the presence of ATP and co-immunoprecipitate with the ATP-dependent chaperone protein Hsp70. J Biol Chem 277:9590–9597

Kelly TA, Jeanfavre DD, McNeil DW, Woska JR Jr, Reilly PL, Mainolfi EA, Kishimoto KM, Nabozny GH, Zinter R, Bormann BJ, Rothlein R (1999) Cutting edge: a small molecule antagonist of LFA-1-mediated cell adhesion. J Immunol 163:5173–5177

Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FC (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol Cell 9:1133–1143

Krause R, von Mering C, Bork P, Dandekar T (2004) Shared components of protein complexes-versatile building blocks or biochemical artefacts? BioEssays 26:1333–1343

Lamesch P, Milstein S, Hao T, Rosenberg J, Li N, Sequerra R, Bosak S, Doucette-Stamm L, Vandenhaute J, Hill DE, Vidal M (2004) *C. elegans* ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. Genome Res 14: 2064–2069

Lehmann AR (2001) The xeroderma pigmentosum group D (XPD) gene: one gene, two functions, three diseases. Genes Dev 15:15–23

Lepourcelet M, Chen YN, France DS, Wang H, Crews P, Petersen F, Bruseo C, Wood AW, Shivdasani RA (2004) Small-molecule antagonists of the oncogenic Tcf/beta-catenin protein complex. Cancer Cell 5:91–102

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO et al (2004) A map of the interactome network of the metazoan *C. elegans*. Science 303:540–543

Mace G, Bogliolo M, Guervilly JH, Dugas du Villard JA, Rosselli F (2005) 3R coordination by Fanconi anemia proteins. Biochimie 87:647–658

Marx SO, Kurokawa J, Reiken S, Motoike H, D'Armiento J, Marks AR, Kass RS (2002) Requirement of a macromolecular signaling complex for beta adrenergic receptor modulation of the KCNQ1-KCNE1 potassium channel. Science 295:496–499

Massague J, Blain SW, Lo RS (2000) TGFbeta signaling in growth control, cancer, and heritable disorders. Cell 103:295–309

Mrowka R, Patzak A, Herzel H (2001) Is there a bias in proteome research? Genome Res 11:1971–1973

Odievre MH, Chretien D, Munnich A, Robinson BH, Dumoulin R, Masmoudi S, Kadhom N, Rotig A, Rustin P, Bonnefont JP (2005) A novel mutation in the dihydrolipoamide dehydrogenase E3 subunit gene (DLD) resulting in an atypical form of alpha-ketoglutarate dehydrogenase deficiency. Hum Mutat 25:323–324

OMIM Online Mendelian Inheritance in Man (2005) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, and National Center for Biotechnology Information, National Library of Medicine, Bethesda. http://www.ncbi.nlm.nih.gov/omim/. Cited 29 July 2005

Oneyama C, Nakano H, Sharma SV (2002) UCS15A, a novel small molecule, SH3 domain-mediated protein-protein interaction blocking drug. Oncogene 21:2037–2050

Pagliaro L, Felding J, Audouze K, Nielsen SJ, Terry RB, Krog-Jensen C, Butcher S (2004) Emerging classes of protein-protein interaction inhibitors and new tools for their development. Curr Opin Chem Biol 8:442–449

Park J, Kunjibettu S, McMahon SB, Cole MD (2001) The ATM-related domain of TRRAP is required for histone acetyltransferase recruitment and Myc-dependent oncogenesis. Genes Dev 15:1619–1624

Pawson T, Raina M, Nash P (2002) Interaction domains: from simple binding events to complex cellular behavior. FEBS Lett 513:2–10

Poetz O, Ostendorp R, Brocks B, Schwenk JM, Stoll D, Joos TO, Templin MF (2005a) Protein microarrays for antibody profiling: specificity and affinity determination on a chip. Proteomics 5:2402–2011

Poetz O, Schwenk JM, Kramer S, Stoll D, Templin MF, Joos TO (2005b) Protein microarrays: catching the proteome. Mech Ageing Dev 126:161–170

Presley JF (2005) Imaging the secretory pathway: The past and future impact of live cell optical techniques. Biochim Biophys Acta 1744:259–272

Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods 24:218–229

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, Chemama Y, Labigne A, Legrain P (2001) The protein-protein interaction map of *Helicobacter pylori*. Nature 409:211–215

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17:1030–1032

Rubin GM (2001) The draft sequences. Comparing species. Nature 409:820–821

Ruf RG, Xu PX, Silvius D, Otto EA, Beekmann F, Muerb UT, Kumar S, Neuhaus TJ, Kemper MJ, Raymond RM, Jr., Brophy PD, Berkman J, Gattas M, Hyland V, Ruf EM, Schwartz C, Chang EH, Smith RJ, Stratakis CA, Weil D, Petit C, Hildebrandt F (2004) SIX1 mutations cause branchio-oto-renal syndrome by disruption of EYA1-SIX1-DNA complexes. Proc Natl Acad Sci USA 101:8090–8095

Schweitzer B, Predki P, Snyder M (2003) Microarrays to characterize protein interactions on a whole-proteome scale. Proteomics 3:2190–2199

Shimogawa H, Kwon Y, Mao Q, Kawazoe Y, Choi Y, Asada S, Kigoshi H, Uesugi M (2004) A wrench-shaped synthetic molecule that modulates a transcription factor-coactivator interaction. J Am Chem Soc 126:3461–3471

Shirohzu H, Kubota T, Kumazawa A, Sado T, Chijiwa T, Inagaki K, Suetake I, Tajima S, Wakui K, Miki Y, Hayashi M, Fukushima Y, Sasaki H (2002) Three novel DNMT3B mutations in Japanese patients with ICF syndrome. Am J Med Genet 112:31–37

Sillerud LO, Larson RS (2005) Design and structure of peptide and peptidomimetic antagonists of protein-protein interaction. Curr Protein Pept Sci 6:151–169

Silletti S, Kessler T, Goldberg J, Boger DL, Cheresh DA (2001) Disruption of matrix metalloproteinase 2 binding to integrin alpha vbeta 3 by an organic molecule inhibits angiogenesis and tumor growth in vivo. Proc Natl Acad Sci USA 98:119–124

Sprinzak E, Sattath S, Margalit H (2003) How reliable are experimental protein-protein interaction data? J Mol Biol 327:919–923

Stagljar I, Korostensky C, Johnsson N, te Heesen S (1998) A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. Proc Natl Acad Sci USA 95:5187–5192

Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, Wan K, Rubin GM, Celniker SE (2002) A *Drosophila* full-length cDNA resource. Genome Biol 3:Research0080

Tewari M, Hu PJ, Ahn JS, Ayivi-Guedehoussou N, Vidalain PO, Li S, Milstein S, Armstrong CM, Boxem M, Butler MD, Busiguina S, Rual JF, Ibarrola N, Chaklos ST, Bertin N, Vaglio P, Edgley ML, King KV, Albert PS, Vandenhaute J, Pandey A, Riddle DL, Ruvkun G, Vidal M (2004) Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-beta signaling network. Mol Cell 13:469–482

The IMAGE Consortium (1999) Welcome to the world's largest public collection of genes. http://image.llnl.gov. Cited 29 July 2005

The Jena Centre for Bioinformatics (2005) The Jena Centre for Bioinformatics protein-protein interaction Website. http://www.imb-jena.de/jcb/ppi/. Cited 29 July 2005

Trojanowski JQ, Lee VM (2000) "Fatal attractions" of proteins. A comprehensive hypothetical mechanism underlying Alzheimer's disease and other neurodegenerative disorders. Ann N Y Acad Sci 924 62–67

Turkson J, Kim JS, Zhang S, Yuan J, Huang M, Glenn M, Haura E, Sebti S, Hamilton AD, Jove R (2004) Novel peptidomimetic inhibitors of signal transducer and activator of transcription 3 dimerization and biological activity. Mol Cancer Ther 3:261–269

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403:623–627

Vassilev LT, Vu BT, Graves B, Carvajal D, Podlaski F, Filipovic Z, Kong N, Kammlott U, Lukacs C, Klein C, Fotouhi N, Liu EA (2004) *In vivo* activation of the p53 pathway by small-molecule antagonists of MDM2. Science 303:844–848

Vertessy BG, Kovacs J, Low P, Lehotzky A, Molnar A, Orosz F, Ovadi J (1997) Characterization of microtubule-phosphofructokinase complex: specific effects of MgATP and vinblastine. Biochemistry 36:2051–2062

Vidal M (2005) Interactome modeling. FEBS Lett. 579:1834–1838

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417:399–403

Waal ND, Yang W, Oslob JD, Arkin MR, Hyde J, Lu W, McDowell RS, Yu CH, Raimundo BC (2005) Identification of nonpeptidic small-molecule inhibitors of interleukin-2. Bioorg Med Chem Lett 15:983–987

Wallrabe H, Periasamy A (2005) Imaging protein molecules using FRET and FLIM microscopy. Curr Opin Biotechnol 16:19–27

Wang JL, Liu D, Zhang ZJ, Shan S, Han X, Srinivasula SM, Croce CM, Alnemri ES, Huang Z (2000) Structure-based discovery of an organic compound that binds Bcl-2 protein and induces apoptosis of tumor cells. Proc Natl Acad Sci USA 97:7124–7129

Way JC (2000) Covalent modification as a strategy to block protein-protein interactions with small-molecule drugs. Curr Opin Chem Biol 4:40–46

Wilson L, Panda D, Jordan MA (1999) Modulation of microtubule dynamics by drugs: a paradigm for the actions of cellular regulators. Cell Struct Funct 24:329–335

Wiszniewski W, Fondaneche MC, Louise-Plence P, Prochnicka-Chalufour A, Selz F, Picard C, Le Deist F, Eliaou JF, Fischer A, Lisowska-Grospierre B (2003) Novel mutations in the RFXANK gene: RFX complex containing *in-vitro*-generated RFXANK mutant binds the promoter without transactivating MHC II. Immunogenetics 54:747–755

Yarbrough WG, Buckmire RA, Bessho M, Liu ET (1999) Biologic and biochemical analyses of p16(INK4a) mutations from primary tumors. J Natl Cancer Inst 91:1569–1574

Zhao HF, Kiyota T, Chowdhury S, Purisima E, Banville D, Konishi Y, Shen SH (2004) A mammalian genetic system to screen for small molecules capable of disrupting protein-protein interactions. Anal Chem 76:2922–2927

Zhou HX (2004) Improving the understanding of human genetic diseases through predictions of protein structures and protein-protein interaction sites. Curr Med Chem 11:539–549

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M, Snyder M (2001) Global analysis of protein activities using proteome chips. Science 293:2101–2105

Zweifel ME, Leahy DJ, Hughson FM, Barrick D (2003) Structure and stability of the ankyrin domain of the *Drosophila* Notch receptor. Protein Sci 12 2622–2632

9 Biomedical Applications of Proteomics

Jean-Charles Sanchez, Yohann Couté, Laure Allard, Pierre Lescuyer, and Denis F. Hochstrasser

Abstract

In biomedical research, understanding and defining the origin of diseases and their effect on multiple organs is a necessity. It has to be done at several levels: genomic, transcriptomic, proteomic and metabolomic. When the origin of a disease is suspected to be monogenic, the best approach to unravel its cause is obviously genomic. When the behaviour or the aggression of a cancer and its response to therapy are related to multiple gene modifications or expression deregulation, tissue biopsies should be analysed at transcriptomic levels by reverse transcription PCR and DNA microarrays. When the disease impacts mostly on the internal environment and is due to the accumulation of 'toxic' material affecting multiple organs, proteomic and metabolomic approaches are required. The present chapter discuss how proteomics can be used in medicine for basic or applied research, in fundamental or clinical domains, to unravel disease processes or to discover biomarkers and therapeutic targets.

9.1 Introduction

The aetiology of diseases is complex. In many cases, no single factor can fully account for the phenotype. However, the potential origins of diseases can be divided into two main categories: genetic and environmental. The latter category can be further divided depending upon the causative agent: microbes and 'living' agents, or physicochemical and 'non-living' agents.

The causes of some diseases are mainly genetic, stemming from single gene deletions or mutations. These give rise to monogenic disorders such as sickle cell anaemia, cystic fibrosis or Huntington's chorea. Origins of diseases also range from environmental factors such as radiation, poisoning or other intoxication to accidents. Most diseases, however, have both a genetic and an environmental basis and multiple genes are often involved. High blood pressure and diabetes are both due to multiple gene 'defects' in conjunction with high salt and calorie intake. They are also frequently associated with an obese phenotype. Several coagulation disorders are due to a genetic predisposition,

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

the concomitant intake of contraception pills and a habit of smoking. Susceptibility or resistance to microbial infection is linked to several genetic backgrounds or phenotypic conditions such as pregnancy.

Diseases and their treatment can affect an organism's internal or external environment. Irradiation and some chemical therapies can modify DNA and lead to a greater predisposition to cancer. Cancer cells might further alter their own genomes owing to defects in their DNA-repair mechanisms. New and more aggressive cell types may thus emerge over time. Uncontrolled diabetes with high sugar level modifies our internal environment, in part, by the glycation of blood proteins. Glycated haemoglobin or haemoglobin A1c reflects the duration and magnitude of hyperglycaemia over time. Cataracts and kidney damage are typical consequences of these protein modifications.

Therefore, in biomedical research, understanding and defining the origin of diseases and their effect on multiple organs is a necessity. It has to be done at several levels, namely genomic, transcriptomic, proteomic and metabolomic. When the origin of a disease is suspected to be monogenic, the best approach to unravel its cause is obviously genomic. When the behaviour or the aggression of a cancer and its response to therapy are related to multiple gene modifications or expression deregulation, tissue biopsies should be analysed at transcriptomic levels by reverse transcription PCR and DNA microarrays. When the disease impacts mostly on the internal environment and is due to the accumulation of 'toxic' material affecting multiple organs, proteomic and metabolomic approaches are required. In juvenile diabetes with severe insulin depletion, elevated sugar levels and its consequences such as keto-acidosis and microalbuminuria are best measured at the metabolomic and proteomic levels. The accumulation of insoluble 'storage' material in the brain of Alzheimer's disease (AD) patients and its consequences on tissues, cells, biochemical pathways and protein complexes should be studied by proteomic methods.

9.2 The Application of Proteomics to Medicine

Proteomics can be used in medicine for basic or applied research, in fundamental or clinical domains, to unravel disease processes or to discover biomarkers and therapeutic targets. So far, the value of proteomics has mainly been shown in fundamental applications such as discovery tools. Proteomics has provided valuable results in highlighting protein complexes such as nucleoli, ribosomes or other organelles or identifying proteins involved in major biochemical pathways. It demonstrated its power to identify and classify, with the help of bioinformatics, proteins from pathogenic microorganisms or from several cell fractions such as membrane, nuclei or other components.

In clinical medicine, numerous applications of proteomics have been published. But, so far, no proteomic methods have entered the arena of routine

clinical chemistry laboratories. Biochemical tests should fulfil at least one of the conditions listed below, otherwise the tests will have little value in routine clinical practice. These are:

1. It should help establish the diagnosis of a disease.
2. It should provide prognostic information about a disease.
3. It should guide the therapy of a disease.
4. It should suggest preventive actions when no disease has been detected yet.
5. It should provide predictive information on potential genetic diseases where preventive action is possible.

If none of the five criteria are met, a biomedical test is not justified and should therefore not be performed. In addition, because of the tremendous cost of health care, a new test is more likely to be accepted if it does not modify the overall health expenditure, but keeps the cost of medicine constant or decreases it by replacing more costly diagnostic or interventional procedures.

One of the key elements of medical practice is not to do harm to patients. False-positive results of biomedical tests could have deleterious effects on healthy people, and false negatives may impact on patients with disease. Therefore, when one considers the development of a new test, two critical elements should be considered: the positive and negative predictive value. Positive and negative predictive values rely on the prevalence of the disease in the population studied and the false-negative and false-positive rates of the applied test. False-negative and false-positive rates are directly linked to the specificity and sensitivity of a test. The specificity and sensitivity of any test need to be excellent. In epidemiological studies, the false-negative rate has a direct effect on the value of the test, but the false-positive rate has a tremendous effect on the direct and indirect health care cost. Consequently, successful clinical proteomic applications will only be ones that develop sensitive, specific, precise and accurate tests.

Multiplying the number of tests increases the number of false-positive results. When developing and validating a large panel of tests, the size of the patient and control cohorts has to be augmented to obtain statistically significant results. It is therefore unlikely that hundreds of protein tests on a chip will ever be used in routine practice. Limited panel or even single biomarkers, however, have a great future. The following calculation simply demonstrates the existence of numerous undiscovered biomarkers. It is estimated that human cells contain more than 10,000 different proteins. Approximately 80% have a common housekeeping function, 10% are tissue-specific and 5% are cell-specific and sub-organ-specific. Even if 1% of the proteins are absolutely specific of one cell type, such as troponin I of the myocardium, 1% of 10,000 is equal to 100 potential biomarkers of damage for each cell type.

Immunoassays of any kind are the central method to measure proteins of interest in the clinic. The value of mass spectra with unidentified peaks in routine clinical laboratory practice has been demonstrated for therapeutic

drug monitoring and might be adopted more broadly in the future. It requires extreme care with the samples prior to analysis, highly reproducible workflows, comprehensive databases and powerful comparison algorithms that use sophisticated statistics.

9.3 Disease Diagnosis from Body Fluids

The use of proteomics of body fluids for the diagnosis and monitoring of disease is a difficult task. Many proteomic techniques have been applied for the profiling of body fluids, including 2-D electrophoresis (2-DE), surface-enhanced laser desorption/ionisation (SELDI), isotope-coded affinity tag (ICAT) labelling and 2-D protein fractionation. However, the real impact of proteomics on disease diagnosis and patient care to date is insignificant. During the last 10 years, a very small number of novel diagnostic tests have been launched with clear clinical utility. On the other hand, the number of published papers describing the discovery of new diagnostic markers using proteomic strategies has increased exponentially since 2000. Why is that? There are potentially five main reasons for this hiatus:

1. Most of the biomarkers are still under validation.
2. Most of the biomarkers are not specific enough (found in many different diseases) or did not prove their clinical utility.
3. The multicentric studies required for test validation are difficult and expensive.
4. The payoff for diagnostic companies is not large enough to justify their investment.
5. Incorrect proteomic discovery strategies have been used to date.

The discovery of protein patterns in plasma or serum samples has been widely investigated in the last two decades. This body fluid is easily accessible and is thought to contain a major part, if not all, of the human proteome. Unfortunately, the dynamic range of protein concentration in plasma is huge and the concentration of disease-specific proteins is relatively low. Accordingly, proteomic strategies have mostly found patterns associated with the modulation of non-specific proteins from common mechanisms, including acute-phase response to diseases and inflammation processes. The development of new technologies is thus mandatory for the application of proteomics to early diagnosis in body fluids, including plasma. Nevertheless, with the advances in new proteomic technologies such as on-the-fly biomarker identification by mass spectrometry (MS) or high-density antibody microarrays, it is likely that panels of biomarkers will emerge with potential for use in routine diagnosis.

To translate proteomic discoveries into a clinical setting, there is a need for close collaboration between basic proteomics and clinical research.

The complexity of proteomics and of human diseases requires that the design of studies be carefully planned. It should take into account at least these seven rules:

1. Before discovery projects are commenced, establish that there is a need for the test.
2. Define how the sampling (patient selection) will be undertaken.
3. Set the predictor variable and more importantly the outcome variable (gold standard).
4. Design the study to allow tests to be performed blind.
5. Calculate the sample size to obtain sufficient statistical power.
6. Be sure that there are enough patients to meet the required sample size.
7. Demonstrate the values of the test in terms of sensitivity, specificity and predictive value.

Finally, to be implemented in a clinical setting, the diagnostic markers discovered by proteomic strategies should be validated by careful confirmatory studies.

9.4 Vascular Diseases

9.4.1 Introduction

Vascular diseases are characterised by conditions that clog or weaken blood vessels. They have been widely investigated as they represent a major cause of death in industrialised countries (Minino et al. 2002; Gianazza and Sironi 2004). Vascular diseases are mainly caused by elevated levels of cholesterol and triglycerides in the blood, hypertension, hyperhomocysteinemia, diabetes and obesity, systemic inflammation and metabolic syndrome, but also a lack of physical activity, cigarette smoking, a high-fat diet, low antioxidant levels and infectious agents. Gender and age may also play a role. Vascular diseases involve many different processes. These include slow processes, such as atherosclerosis that can evolve over decades, and much more rapid events, such as infarction in the heart or the brain.

Atherosclerosis is defined as the deposition of fatty substances, cellular waste products, cholesterol, calcium and other substances in the inner lining of an artery, called plaque. Plaque can be hard and stable, or soft and unstable. Hard plaque causes artery walls to thicken and harden. Soft plaque is more likely to break apart from the walls and enter the bloodstream. This can cause a clot that can partially or totally block the flow of blood in an artery and cause heart conditions or stroke (reviewed by Lusis 2000). The main consequences of vascular diseases to the heart are angina pectoris (the heart needs more blood), congestive heart failure (the heart cannot pump enough blood to the body's other organs) and heart attack (sudden interruption or insufficiency of the blood supply to the heart).

Cerebrovascular accident, or stroke, is the third most frequent cause of death and a leading cause of disability in industrialised countries. About 10–20% of all stroke patients will suffer from an intracerebral haemorrhage, due to the rupture of a small artery within the brain parenchyma. The vast majority of patients (80–90%) will suffer from an ischemic stroke resulting from the atherothrombotic or embolic occlusion of a large, middle or small artery supplying the brain.

The pathological mechanisms leading to plaque formation and rupture still remain unknown. Diagnosis and prognosis of patients at risk and/or presenting with acute ischemia of the heart or brain still remains difficult and mainly relies on clinical examination, medical history of the patient and costly imaging techniques. This section summarises recent findings using proteomic strategies in the field of vascular diseases in general and reports specific investigations of heart attack and stroke.

9.4.2 Application of Proteomics to Vascular Diseases and Atherosclerosis

Proteomics has promise as a powerful means to explore atherosclerotic plaques and the endothelial cells composing the vessels where the deposition occurs (reviewed by Mayr et al. 2004). At present, human umbilical vascular endothelial cells are the most explored model of endothelial cells. An annotated 2-DE map of these cells is available at http://www.huvec.com (Bruneel et al. 2003).

Currently, diagnosis of atherosclerosis mainly relies on radiological imaging, but remains limited by variable efficacy, high cost and the difficulty of characterising the pathological process. The lack of circulating biomarkers for atherosclerosis led Matuszeck et al. (2003) to look for proteins that are associated with atherosclerotic plaques. They analysed perfusates of 30 patients with atherosclerosis and of seven patients without, and identified haptoglobin (Hpt) β-chain as a differentially expressed protein. This result was validated by western blot and nephelometric analysis on the plasma of the same patients, and allowed the two populations to be distinguished ($p = 0.002$). Nevertheless, as concluded by the authors, the specificity of Hpt as a circulating plasma marker of atherosclerosis remains to be demonstrated. Among the newly developed approaches for atherosclerosis plaque composition elucidation, Duran et al. (2003) investigated secreted proteins from the supernatant of cultured artery segments. For this, they studied three types of segments: normal artery segment, non-complicated plaques and complicated plaques with thrombus. Forty-two proteins were obtained for normal artery segments, 154 for segments bearing a plaque and 202 for artery segments with ruptured plaque and thrombus. These results clearly demonstrated that the number and type of secreted proteins is directly proportional to the severity and complexity of the lesion. Matrix-assisted laser desoption/ionisation (MALDI) time-of- flight (TOF) MS showed that proteins secreted by human

carotid artery segments containing a non-complicated plaque are involved in cholesterol transport (apolipoprotein B-100 and apolipoprotein A-1) or participate in the elimination of toxic radicals (superoxide dismutase and peroxiredoxin). Nevertheless, a clear correlation between plaques and identified secreted proteins remains to be achieved. Similarly, in the search for coronary atherosclerosis biomarkers, You et al. (2003) analysed proteins extracted from ten diseased and seven normal coronary arteries. The expression of ferritin light chain was 1.9-fold higher in the diseased arteries. As ferritin light chain protein mediates storage of iron in cells, this result is consistent with the 'iron hypothesis' and reinforces the idea of an association between excessive iron storage and risk of coronary artery disease (Liotta et al. 2003).

Recent studies have focused on the changes in the proteome during the progression of atherosclerotic lesions. Indeed, rupture of a plaque is the main cause of acute coronary syndrome and peripheral vascular disease. Comparing the cholesterol-laden macrophage as a cellular model of atherosclerosis stimulated with oxidised low-density lipoprotein (LDL) and with LDL, Fach et al. (2004) identified 59 upregulated proteins in association with oxidised LDL, 17 downregulated proteins and 57 that showed no change. Immunoassays of cell supernatant confirmed the overexpression of cyclophilin B and cathepsin L, downregulation of tissue inhibitor metalloproteinase 1 (TIMP1) and unchanged levels of matrix metalloproteinase 9 (MMP9). In a more recent study (Donners et al. 2005) 2-D profiles of whole-mount advanced stable lesions were compared with those of plaques containing a thrombus. They identified vinexin-β and α1-antitrypsin in the same 2-DE spot. Differential expression of vinexin-β could not be confirmed, but six isoforms of α1-antitrypsin were observed by western blot. One of the isoforms was only found in thrombus containing plaques.

9.4.3 Application of Proteomics to Cardiovascular Diseases

Proteomic analysis has been widely used in the field of cardiovascular disease, with the specific aims of (1) undertaking a survey of protein expression, (2) identifying biomarkers of specific disease states and (3) understanding the molecular mechanisms of ischemic heart events.

A study on the proteome and secretome of smooth muscle cells isolated from segments of arteries from patients undergoing coronary artery bypass grafting was published recently (Dupont et al. 2005). Reference 2-DE maps of secreted and intracellular proteins were built, leading to the identification of 83 intracellular and 18 secreted proteins. Most of these proteins appeared to be cytoskeletal, but others were involved in a wide range of functions such as protein biosynthesis and proteolysis, cellular defence or metabolic pathways. This confirmed the complex pathogenesis of heart disease.

To establish if specific patterns of proteins are associated with acute coronary syndrome, Mateos-Caceres et al. (2004) analysed plasma protein modifications by 2-DE. They compared plasma samples from 11 patients with acute myocardial infarction (AMI), eight with unstable angina and nine age-matched volunteers as controls. They highlighted differential expression of three protein isoforms (α1-antitrypsin, fibrinogen γ-chain and apolipoprotein-A1) between acute myocardial infarction patients , unstable angina patients and the controls. These results suggest that post-translational modifications could be used to characterise heart disease subtypes.

The molecular mechanisms underlying other vascular diseases are also largely unknown and proteomics is a key tool to improve our knowledge. For instance, to identify phosphoproteins that may act as protectors in cardiac allograft vasculopathy, De Souza et al. (2005) investigated 22 patients within the first 2 weeks and then 9 years after transplant. Among them, 11 had developed early cardiac allograft vasculopathy after transplant. Using cardiac biopsies, 2-DE revealed a diphosphorylated form of HSP27 as being associated with normal blood vessels. In another example, Brennan et al. (2004) reported intermolecular protein disulfide formation during cardiac oxidative stress.

9.4.4 Application of Proteomics to Cerebrovascular Disease

A number of groups have used proteomics with cellular and animal models to study ischemia of the brain. For instance, Schrattenholz et al. (2005) recently reported a combination of affinity enrichment of phosphoproteins, isotopic labelling, 2-DE and MS to investigate chemical ischemia on neural embryonic stem cells. They identified differential isoforms of seven proteins from neurons exposed to chemical ischemia. Some of these are known to be responsive to oxidative stress and others are part of the immediate molecular response to chemical ischemia induced in cultured cells. Animal models of brain damage and ischemia have also been developed. Murine transient or permanent middle cerebral artery occlusion (MCAO) is the most widely used animal model of focal ischemia. An original approach combining SELDI-TOF-MS, 2-DE and peptide mass fingerprinting allowed Suzumaya et al. (2004) to discover CSF protein patterns after transient brain ischemia. They highlighted the rapid increase of monomeric transthyretin in rat CSF after transient MCAO. Brain damage has also been studied using spontaneously hypertensive stroke-prone rats, an inbred animal model of cerebrovascular pathology that resembles human stroke. Unlike the MCAO model, the SHSPR model displays a vasogenic ischemia (breakdown of the cerebral vasculature's autoregulatory system or the blood-brain barrier) with no evidence of cytotoxic oedema (transfer of water from the extracellular space into the cells) (Guerrini et al. 2002). To investigate the pathogenic mechanism leading to cerebrovascular disease in SHSPR, Gianazza and Sironi (2004) monitored the appearance of brain damage and the alteration

of the permeability of the blood–brain barrier by MRI and 2-DE. They highlighted an impairment of the blood–brain barrier simultaneously with the detection of brain abnormalities by MRI, followed by the appearance of plasma proteins in the CSF. These proteins include thiostatin, a marker of inflammation, and high molecular weight proteins that indicate gross alteration in the permeability of the barrier.

Body fluids, such as cerebral fluids, provide good opportunities for the investigation of brain injury. Maurer et al. (2003) analysed the proteome of human brain microdialysate, a fluid that allows neurotransmitter release and metabolic products to be monitored in stroke patients. The microdialysate obtained from the non-infarcted hemisphere of three stroke patients was analysed by 2-DE. They identified 27 proteins, among them 17 already previously described in the CSF and ten uniquely found in the microdialysate. These last ten proteins could potentially be used, after validation in more easily obtained plasma, as biomarkers for diagnosis, prognosis or follow-up of stroke patients.

In our laboratory, we have developed proteomic strategies for the identification and validation of new diagnostic markers of stroke. A first approach relied on SELDI-TOF protein profiling. Seven peaks were found differentially expressed ($p<0.05$) and were identified as apolipoprotein CI, apolipoprotein CIII, serum amyloid A and a fragment of antithrombin-III. Assessment of apolipoprotein CI and apolipoprotein CIII levels in stroke plasma samples using a sandwich ELISA enabled the discrimination of an ischemic from a hemorrhagic event with a high sensitivity and specificity ($p<0.005$) (Allard et al. 2004). In another approach, the concept of post-mortem CSF as a model of massive and global brain insult was explored (Allard et al. 2004; Lescuyer et al. 2004). Supporting this view, Zimmerman-Ivol et al. (2004) demonstrated that H-FABP identified in post-mortem CSF may be reliably used as a diagnostic marker of stroke. In a more recent study, we presented RNA-binding protein regulatory subunit (RNA-BP, also called PARK7) and nucleoside diphosphate kinase A (NDKA) as diagnostic biomarkers of stroke. They were also found overexpressed in post-mortem CSF compared with ante-mortem CSF and were validated by ELISA in plasma samples in three independent retrospective studies encompassing 165 control patients and 622 stroke patients (Allard et al. 2005).

9.4.5 Conclusion

Vascular diseases are complex, involve many different mechanisms and lead to diverse pathological events. Consequently, a vast variety of models have been developed to investigate different aspects of such diseases. In this context, proteomic analysis is proving to be a powerful tool to study normal versus pathological conditions, although the outcomes from most studies still require validation. A further complication is that the division between

normal and disease states is not always clear, which can make interpretation of results difficult.

9.5 Neurodegenerative Disorders

The prevalence of neurodegenerative disorders is growing exponentially in developed countries, as the result of an increasing life span in the general population. Neurodegenerative disorders include a wide range of abnormalities, such as Alzheimer's disease (AD), Parkinson's disease (PD), Creutzfeldt–Jakob disease (CJD), Huntington's disease (HD), amyotrophic lateral sclerosis (ALS) and various neurodegenerative dementias. These diseases generally begin late in life and slowly but inexorably cause progressive neuronal degeneration and result in disability or death. Their diagnosis is difficult for several reasons. These include:

1. The progression of the disease is insidious and the symptoms appear at an advanced stage of the neurodegenerative process (DeKosky and Marek 2003). The diagnosis is then made when the brain lesions are fully constituted.
2. Almost no biological marker is currently available for the routine diagnosis of these abnormalities. Diagnosis relies mainly on clinical examination and neuroimaging techniques (DeKosky and Marek 2003).
3. Differential diagnosis is difficult since there is a considerable overlap between the clinicopathological features of many disorders (Armstrong et al. 2005).
4. The definite diagnosis of many of these conditions, notably CJD, is possible only by post-mortem examination of brain tissue.

The role that proteomics can play in the field of research on neurodegenerative disorders is well described in the goals of the Human Proteome Brain Project launched by the Human Proteome Organisation – HUPO (Hamacher et al. 2004).

9.5.1 Brain Proteome

The systematic proteomic analysis of human or animal brains is of great relevance to neurological diseases. Such studies provide a detailed understanding of the normal and healthy state, and allow comparisons with affected patients or animal models of disease. The first published proteomic studies used 2-DE to produce reference maps of human, mouse and rat brain proteomes. For example, more than 400 proteins were identified from mouse brain (Klose et al. 2002) and more than 300 proteins were identified from human brain (Lubec et al. 2003). The proteins identified corresponded to various functional

categories: structural proteins, energy metabolism, protein synthesis, protein degradation, RNA transcription, chaperones, oxidative stress response, signal transduction and synaptic proteins. However, it is likely that the 2-DE analysis of total extracts from whole brain or brain compartments led mainly to the detection of high-abundance proteins. Furthermore, the proteins identified were derived from various cell types: neurons, astrocytes, oligodendrocytes, microglia, blood vessel walls and blood cells.

In recent publications, gel-free technologies were reported for the profiling of the brain proteome. Moreover, efforts were focused on particular subfractions, enabling deeper and more specialised analysis of the proteome. Nielsen et al. (2005) analysed brain plasma membrane proteins using an original fractionation protocol. This procedure was based on high-speed shearing of tissues in solution for removal of soluble proteins, followed by density-gradient enrichment of membrane fractions. Proteins were then digested on-membrane with endoproteinase Lys-C. The resulting peptides were fractionated by reversed-phase chromatography, digested by trypsin and analysed by liquid chromatography (LC)– tandem MS (MS/MS). One advantage of this method is that it requires much lower amounts of tissue than classic membrane-enrichment approaches based on subcellular fractionation. It is thus applicable to small brain compartments and to clinical samples. Use of this method allowed 862 proteins to be identified from 150 mg of mouse brain cortex. After further development and miniaturisation, the authors identified 1,685 proteins from 15 mg of mouse hippocampus. Of the proteins with assigned subcellular localisation, more than 60% were annotated as membrane proteins, including several classes of ion channels and neurotransmitter receptors. In another study, synaptosomal proteins were investigated using ICAT labelling and LC-MS/MS (Schrimpf et al. 2005). Synaptosomes are a subcellular fraction of brain tissue corresponding to isolated synapses. They contain the complete presynaptic terminal and portions of the postsynaptic side. A total of 1,131 proteins were identified, including synaptic adhesion molecules, postsynaptic scaffolding proteins, postsynaptic receptors, postsynaptic receptor-ligands, and proteins involved in synaptic vesicle trafficking or signal transduction cascades.

9.5.2 Proteomic Profiling of Neurodegenerative Disorders

The goal of proteomic studies investigating neurological disorders is to detect differences in protein expression and in protein post-translational modifications that are associated with the disease. The identification of these changes can yield insights into potential molecular mechanisms of neurodegeneration and lead to the identification of potential diagnostic markers or therapeutic targets. For example, the detection of oxidised proteins is of particular interest since oxidative damage is thought to play an important role in the pathogenesis and progression of most, if not all, neurodegenerative disorders.

AD is the major dementia affecting the elderly; thus, it is logical to find a large number of proteomic studies devoted to the investigation of this disease (Butterfield et al. 2003). Proteomic analyses of different brain regions from AD patients and non-demented controls led to the identification of various proteins with altered expression levels. These changes in protein expression reflect the cascade of alterations on multiple pathways within the brain of AD patients: decreased energy metabolism, increased oxidative stress, dysregulation of apoptosis, protein misfolding, neurotransmitter imbalance and decreased levels of proteins involved in neuronal cell proliferation, neurite outgrowth or synaptic plasticity. Furthermore, numerous proteins were identified as targets of protein oxidation, which shows the importance of oxidative damage in the progression of the disease (Butterfield et al. 2003). As observed for proteins with altered expression levels, oxidised proteins belong to various functional classes such as energy metabolism, the ubiquitin–proteasome system or neuronal development.

One hallmark of neurodegenerative disorders is the presence of protein deposits in the nervous system. These exist either in the form of extracellular plaques – amyloid plaques in AD or prion protein deposits in CJD – or in the form of intracellular filamentous inclusions – neurofibrillary tangles in AD or Lewy bodies in PD) (Armstrong et al. 2005). Recently, two papers describing the proteomic profiling of such structures after isolation by laser-capture microdissection were published. In the first study, 488 proteins were identified by LC-MS/MS in amyloid plaques isolated from post-mortem AD brain tissues (Liao et al. 2004). Moreover, 26 proteins were found enriched at least twofold in the plaques by quantitative comparison with the adjacent non-plaque regions of the cortex. This approach allowed the discovery of dynein heavy chain in amyloid plaques, a protein involved in intracellular motility of vesicles and organelles along microtubules. The colocalisation of dynein heavy chain with amyloid plaques was corroborated by immunofluorescence confocal microscopy in human AD cortex and in a transgenic mouse model of AD. In the second study, a similar approach was used for the proteomic profiling of neurofibrillary tangles isolated from pyramidal neurons of AD patients (Wang et al. 2005). A total of 155 proteins were tentatively identified, although the majority of these with only one peptide. Of the 72 proteins identified with multiple unique peptides, 63 were not known to be associated with neurofibrillary tangles. Immunohistochemistry experiments confirmed the colocalisation of one of these proteins, glyceraldehyde-3-phosphate dehydrogenase, with neurofibrillary tangles. Further investigations also showed that this protein immunoprecipitates with phosphorylated tau, the major protein component of neurofibrillary tangles, and that it is one of the few proteins known to undergo conversion to a detergent-insoluble form in AD.

Several transgenic animal and cell line models have been developed for the study of neurodegenerative disorders such as AD, PD, CJD, ALS and HD. The R6/2 transgenic mouse model of HD was used in two proteomic studies to identify differentially expressed proteins and oxidative modifications associated

with the disease. HD is a hereditary disease caused by a well-characterised genetic defect: the expansion of CAG repeats in exon 1 of the Huntingtin gene. In contrast, the mechanisms by which the mutation causes the disease are not fully understood. The 2-DE analysis of striatum from R6/2 mice and age-matched controls showed that the expression of α-enolase was increased in HD mice (Perluigi et al. 2005). Furthermore, the protein levels of succinyl *S*-transferase and aspartate aminotransferase were found to increase over the course of the disease, while expression of pyruvate dehydrogenase decreased. In addition, measurement of carbonyl levels led to the detection of six proteins that were oxidised in old versus young R6/2 mice: α-enolase, γ-enolase, aconitase, VDAC1, Hsp90 and creatine kinase. Another 2-DE study on R6/2 mice identified two other proteins whose expression decreased in the brain over the course of the disease: α1-antitrypsin and the chaperone αB-crystallin (Zabel et al. 2002). Disease progression was also found to be associated with reduced α1-antitrypsin levels in liver and testes, indicating that the disease also exerts its influence outside the brain.

Proteomic techniques were also used to investigate a cell line model of a subset of familial ALS, linked to mutations in the Cu–Zn superoxide dismutase SOD1 (Fukada et al. 2004). ALS is a fatal neurodegenerative disorder characterised by progressive motor neuron death. One hypothesis regarding the consequence of SOD1 mutations is the dysregulation of mitochondrial functions and the activation of apoptosis. To explore this hypothesis, mitochondrial proteins from wild-type NSC34 motor-neuron-like cell lines or those expressing the G93A-SOD1 ALS-causing mutation were analysed by 2-DE. Forty proteins were found to have altered expression in the mutant cell line, including chaperones, subunits of the mitochondrial respiratory chain complexes and two mitochondrial outer-membrane proteins, VDAC1 and VDAC2, both potentially involved in apoptosis.

9.5.3 Cerebrospinal Fluid Protein Markers

CSF is the body fluid that surrounds the central nervous system. Exchanges between the CSF and the blood are regulated by a diffusion barrier called the blood–brain barrier (Ballabh et al. 2004). Owing to its close proximity to the brain, CSF is considered to be the sample of choice for the discovery of markers of brain damage. Up to now, more than 300 proteins have been identified in CSF samples from healthy subjects using 2-DE or gel-free proteomics (Finehout et al. 2004; Maccarrone et al. 2004; Zhang et al. 2005b). However, it seems clear that this number will increase quickly owing to the development of complex proteomic strategies combining depletion of high-abundance proteins, multiple isoelectric focusing or LC fractionation steps, and highly sensitive MS detection.

A number of proteomic studies have sought to identify protein markers of neurodegenerative disorders by comparing diseased and control CSF samples.

A clear example of the value of this approach was the identification of the 14-3-3 protein as a CSF diagnostic marker of CJD using 2-DE gels (Hsich et al. 1996; Zerr et al. 1996). In another 2-DE study, Puchades et al. (2003) identified eight CSF proteins with altered expression in AD patients compared with controls. In particular, they detected two acidic spots decreased in AD patients, which correspond to an abundant brain-derived glycoprotein called beta-trace or prostaglandin D2 synthase. Experiments performed by our laboratory using a method combining isoelectric focusing and immunoblotting further showed that neurological disorders and particularly neurodegenerative dementia are associated with complex alterations of the beta-trace post-translational modification pattern (Lescuyer et al. 2005). Recently, a study was published describing the use of ICAT labelling and LC-MS/MS to quantify the relative changes in the CSF proteome of AD patients and age-matched controls (Zhang et al. 2005a). Of the 163 proteins quantified in this study, 85 were increased or decreased by 20% or more in the CSF of AD patients; however, the very high number of proteins showing altered expression, which might be explained by the low cut-off value chosen to represent a biologically important difference, makes the interpretation of these results difficult.

Another approach described by our group was the use of post-mortem CSF as a model for the identification of potential markers of brain damage (Lescuyer et al. 2004). The hypothesis was that proteins leaking from the brain owing to the massive necrosis following blood-flow arrest and global brain anoxia could represent markers of neurodegeneration. Comparative 2-DE analysis of ante-morten and post-mortem CSF samples led to the identification of 13 differentially expressed proteins. ELISA experiments showed that CSF and serum levels of one of these proteins, H-FABP, were increased in CJD and other neurodegenerative dementias (Guillaume et al. 2003; Steinacker et al. 2004).

9.6 Proteomics and Cancer

Cancer is a major public health problem worldwide and a leading cause of death in developed countries. Recently, collaborative work between the World Health Organization and the International Agency for Research on Cancer reported that more than ten million new cancers were diagnosed in 2000 and that there were over six million deaths from the disease (Stewart and Kleihues 2003). Nowadays, the incidence of cancer continues to increase slightly. However, the mortality rate from all cancers decreases by about 1% every year in developed countries (Jemal et al. 2005), essentially due to early detection.

Cancers are heterogeneous abnormalities that develop through various well-identified early genetic changes. These drive the progressive, multistep transformation of normal cells into malignant, cancerous cells. It is suggested

that neoplastic cells from most of the different cancers acquire the same set of functional capabilities during their development. These include self-sufficiency in growth signals, insensitivity to antigrowth signals, evading apoptosis, sustained angiogenesis, limitless replicative potential and tissue invasion and metastasis (Hanahan and Weinberg 2000). All these events lead to modifications of the proteome that are due to the reprogramming of cancer cells and also to the body's response to cancer.

Proteomics can contribute to cancer research at several levels, namely for diagnosis, prognosis, monitoring treatment response and the identification of targets for cancer prevention and treatment. Depending on their interests, investigators can use different proteomic approaches to answer their biological questions and to highlight proteins that are present in greater or lesser quantities. Investigations often focus on comparing control tissue versus cancerous tissue, discovery of proteins specific for one state and changes in protein post-translational modifications between states. Nowadays, two main methods are used in cancer proteomics: biomarker discovery and proteomic profiling.

9.6.1 Biomarker Discovery in Cancer Proteomics

In biomarker discovery, the focus is on identifying proteins whose expression levels or modifications are specific for a disease state. Historically, investigators have used 2-DE as the primary tool, but in recent years, new proteomic tools have been developed for biomarker discovery in tissues and cell lines.

9.6.1.1 Tissues

The comparison of proteomes from normal versus malignant tissues has been widely used to highlight proteins that could be involved in disease establishment and progression. For tissue-based studies, sample preparation is of particular importance to avoid protein degradation prior to proteomic analysis. Numerous studies have successfully examined the whole proteome of cancer tissues. However, tumours are heterogeneous samples containing various proportions of cell types such as fibroblasts, endothelial cells, normal epithelial cells, immune cells and others. This heterogeneity has a major impact on comparative studies. This is why different techniques have been set up to selectively enrich samples for cells of interest, allowing the analysis of individual cell types.

Franzen et al. (1995) addressed this critical issue in several publications and demonstrated that non-enzymatic methods for the preparation of tumoral cells, including fine-needle aspiration, scraping or squeezing tissue biopsies, had advantages over methods using enzymatic extraction of cells. Non-enzymatic methods were shown to be rapid and to reduce loss of high molecular weight proteins. These methods did not require the separation of viable and nonviable

cells by Percoll gradient centrifugation. They also analysed qualitative aspects of tissue preparation in relation to the histopathology of lung cancer, and examined the relationship between histopathological findings and 2-DE gel quality. They concluded that histopathological features, such as a local homogeneity, and the amounts of connective tissue and serum proteins were critical factors for the successful preparation of the sample and the high quality of overall protein separation and analysis. They clearly overcame some major technical difficulties. As a result of their work, clear guidelines are now available for sample preparation of patient cells and biopsies (Franzen et al. 1995).

Page et al. (1999) used immunoselection of cells before proteomic analysis. They used a double antibody magnetic affinity cell sorting technique to purify normal human luminal and myoepithelial breast cells from the reduction mammoplasties of ten premenopausal women. For this, two antibodies were used, one produced in rat and directed to the luminal epithelial marker EMA and the other produced in mouse and directed to the myoepithelial antigen CD-10. The use of antirat and antimouse magnetic beads allowed the separation of EMA and CD-10 expressing cells. Myoepithelial cells were then purified a second time using anti-CD-10 and anti-FAP antibodies to free them from F-19 positive fibroblasts. Purified luminal and myoepithelial cells were then subjected to proteomic analyses using 2-DE. One hundred and seventy different proteins were found to be differentially expressed between the two breast cell types and 51 of them were identified using MS/MS. This work forms the basis for future studies of purified breast cancer cells.

Another technique, called laser-capture microdissection, is widely used in studies of cancer tissues (Jain 2002). This technique employs a pulsed infrared laser to activate a transfer film placed over the tissue of interest, causing the film to become fused to the cells. With a diameter measuring only few microns, laser-capture microdissection allows single cells to be extracted from heterogeneous tumour samples. Interestingly, this technique permits tumour cells and normal cells to be isolated from the same biopsy without using chemicals or physical agents that might modify the proteome. An example of this technique is the work of Li et al. (2004) and their study of hepatocellular carcinoma (HCC). They used laser-capture microdissection to isolate HCC and non-HCC hepatocytes, and compared them using cleavable ICAT and 2-D LC-MS/MS. A total of 644 proteins were identified and 261 differentially expressed proteins were described. These results provided a new basis for understanding the mechanism of HCC, and identified potential markers and drug targets that could be useful for disease diagnosis and treatment.

9.6.1.2 Primary and Established Cell Lines

Cell lines have been extensively used for the proteomic analysis of cancer. In several respects, they are much more useful than biopsies. In contrast to tissues, the quantity of cells is unlimited. This allows extensive experiments.

Cell lines also represent a pure cell population and allow investigators to manipulate growth conditions. However, it is important to mention that in vitro cell lines are not identical to the corresponding cells in vivo since they are removed from their native environment. For example, Ornstein et al. (2000) compared the proteome of in vivo prostate cancer cells with in vitro prostate cell lines and found that less than 20% of proteins were shared by both cell types.

Cell lines have proven to be very useful for the unravelling of pathways perturbed by particular genetic alterations. For example, new insights into the development of nervous system tumours in patients bearing the NF1 tumour-predisposition syndrome have been recently published (Dasgupta et al. 2005). For this, the investigators generated transgenic mice producing NF1–/– astrocytes. The proteome of these astrocytes was compared with that of NF1+/+ astrocytes using 2-DE. The MS identification of differentially expressed proteins between these two cell lines revealed that the mammalian target of the rapamycin pathway is hyperactivated in NF1–/– astrocytes. Furthermore, the inhibition of this pathway in NF1–/– astrocytes restores a normal proliferative rate, suggesting that this pathway could represent a target for therapy of brain tumours in patients bearing the NF1 tumour-predisposition syndrome.

Resistance to therapy is the main cause of therapeutic failure and death in cancer patients. Cell lines represent a powerful tool for the study of therapy-resistant cells. Brown and Fenselau (2004) have addressed the question of doxorubicin resistance in breast cancer cells. For this, they used shotgun proteomics with proteolytic ^{18}O labelling to compare the cytosolic proteome of a doxorubicin resistant MCF-7 cell line with the drug-sensitive MCF-7 cell line. This study identified several proteins with altered expression levels in the drug-resistant cell line. A number of these proteins may represent key factors explaining doxorubicin resistance of breast cancer cells.

An advantage of cell lines for cancer proteomics, as compared with tissue samples, is the opportunity for efficient subcellular fractionation. Subcellular fractionation allows the direct investigation of functional cell compartments instead of complete proteomes. This enriches low-abundance proteins, and allows the focus on subcellular compartments with potential value in tumorigenesis and cancer progression. This includes the plasma membrane (Zhao et al. 2004) or nucleoli (Scherl et al. 2002).

Biomarker discovery using proteomics has proven to be a valuable means to identify markers that could become useful in diagnosis, prognosis or treatment of cancer. To date, dozens of markers have been identified using proteomic strategies. The most promising ones are now undergoing validation.

9.6.2 Proteomic Profiling in Oncology

Advances in proteomic technologies have given rise to proteomic profiling – a new approach used in cancer proteomics. In contrast to biomarker discovery,

this approach does not rely on protein identification but instead compares the protein patterns of tumours, cancer cells, subcellular compartments or body fluids from normal and diseased states.

9.6.2.1 Surface-Enhanced Laser Desorption/Ionisation Time-of-Flight Mass Spectrometry

SELDI-TOF MS has become widely used in the past 5 years. It combines sample fractionation on a chip with MALDI-TOF MS analysis, and has become very popular thanks to its simplicity, ease-of-use and automation. The processing of several samples using SELDI-TOF generates proteomic patterns (one peak per polypeptide at a given *m*/*z* value). These patterns are analysed with software able to classify peaks and to segregate proteomic patterns of different samples. SELDI-TOF MS has been used for the analysis of many types of samples, such as tumours, cancer cells isolated by laser-capture microdissection, cell lines or body fluids as starting materials (Xiao et al. 2005). Researchers using this technique have discovered proteomic patterns with better specificity and sensitivity than the single protein biomarkers markers currently in use. However, it must be emphasised that this technique is in its infancy and needs further validation (Diamandis 2004).

A complementary technique was also described that uses coated magnetic beads (Villanueva et al. 2004) instead of chip arrays for sample fractionation. This technique is of interest as it frees investigators from using the SELDI interface and enables proteomic pattern acquisition on numerous mass spectrometers.

9.6.2.2 Protein Microarrays

The development and use of protein microarrays is an intensive area of research. They consist of an array of protein samples (reverse-phase arrays), or of protein baits such as antibodies (forward-phase arrays) immobilised on a solid phase. Reverse-phase arrays are probed with antibodies and forward-phase arrays with cell or tumour lysates. The probe can then be detected using colorimetric, fluorescent or chemiluminescent means. This technique remains under development and faces limitations such as low sensitivity and high variability (Liotta et al. 2003). If successful, however, protein microarrays should permit the high-throughput study of changes in protein expression or modification.

9.6.2.3 Tissue Profiling by Matrix-Assisted Laser Desorption/Ionisation Mass Spectrometry Imaging

First described in 1999 for the analysis of proteins from intact tissues (Chaurand et al. 1999), MALDI MS imaging has the aim of determining the

relative expression levels of proteins and also their spatial distribution. It has been applied to the study of several cancers and has shown promising results, even if imaging MS needs some improvements in sensitivity, speed and bio-computational aspects (Chaurand et al. 2004; see Chap. 6).

9.6.3 Use of Proteomics To Define the Tissue of Origin

The identification of the tissue of origin is often required to define a diagnosis as well as a prognosis and choice of treatment. Even with currently available immunochemistry methods, pathologists sometimes have difficulty to identify a tissue under a microscope or the tissue of origin. For example, an adenocarcinoma may be classified adenocarcinoma of unknown primary or unknown origin. The following example highlights the potential of proteomics to overcome this difficulty. A patient was found to have an abdominal tumour. The mass was surgically resected, but the physicians had difficulties distinguishing between duodenal and pancreatic cancer. Therefore, samples of the tumour, normal duodenal mucosa and normal pancreatic tissue from the same patient were compared by 2-D gel electrophoresis. The tumour pattern resembled the normal duodenal mucosa more closely than pancreatic tissue, suggesting that the tumour was of duodenal origin (Isoda et al. 1990).

9.6.4 Conclusion

Proteomics has provided new avenues for the understanding of cancer. It is showing promise for the identification of potential drug targets and for the discovery of proteins of diagnostic or prognostic value. Proteomic profiling is also of promise and may become a central part of clinical chemistry laboratories in the future, notably as a tool used for diagnosis and prognosis of cancer, but also for the development of personalised cancer therapy.

9.7 Toxicopharmacology: the Example of Type 2 Diabetes

The simultaneous identification, characterisation and quantitation of numerous gene products and their post-translational modifications is required to evaluate how multiple overlapping pathways are affected by drug treatment or toxins. In the past, several authors used 2-DE as the core technique for pharmaceutical and toxicological studies. The aim of this section is to demonstrate how proteomic approaches have already been of help in improving our understanding of the pharmacology and toxicology of drugs in type 2 diabetes.

9.7.1 Introduction to Diabetes

The identification of new molecular targets that could lead to new treatments for diabetes is an area where proteomics could have a key role. Epidemiological data from the late nineteenth century described diabetes mellitus (from the Greek for 'pass through' and the Latin for 'sweet as honey') as a rather frequent disorder in man, in obese people above 50 years old, in cities and in western countries (Pavy 1894). Even at that time, diabetes was seen as a disease of urban life.

There are two main types of diabetes. Type 1 diabetes, also known as insulin-dependent diabetes mellitus, is an autoimmune disorder associated with MHC genes (Campbell and Milner 1993). Type 2 diabetes, known as non-insulin-dependent diabetes mellitus, is a complex multifactorial disease. Type 2 diabetes accounts for 90% of the diabetic population, affects nearly 130 million people and is expected to reach epidemic proportions in the next 10 years (Zimmet and McCarthy 1995; Zimmet 1995). This dramatic increase in the prevalence of type 2 diabetes was predicted 10 years ago when urbanisation, industrialisation and western habits became increasingly widespread in developing countries. This means that type 2 diabetes is a truly global health problem. It is of high cost and suffering primarily owing to its long-term complications.

9.7.2 Pathogenesis of Type 2 Diabetes

Type 2 diabetes is characterised by an abnormal glucose homeostasis leading to hyperglycaemia. Diagnosis often occurs after many years of a prediabetic state accompanied by an absence of pathognomic symptoms. When diabetes is suspected, the standard test is to measure glucose concentration in the plasma after an overnight fast. The diagnostic value is considered to be around 7.1 mmol/L and has to be measured on two different occasions. The oral glucose tolerance test or the intravenous glucose tolerance test also have the ability to diagnose diabetes but are used less widely. Another test used by clinicians is the measurement of glycosylated haemoglobin. It reflects the antecedent glycaemia over a period of 3 months under real-life conditions. The glucose homeostasis deregulation is mainly due to a combination of insulin resistance and defects in insulin secretion.

The pathophysiology of type 2 diabetes is multifactorial and complex, involving interactions between genetic and environmental factors. Many candidate genes have been reported to be associated with the disease, however none of them account for the majority of patients affected by type 2 diabetes. Factors including diet, stress, exercise, aging and obesity seem to play a major role in its development. Considerable research has been undertaken to determine if islet dysfunction or peripheral insulin resistance provokes the other or at least precedes the other. Scientists who support the theory of insulin resistance propose

that in the very early stages of the prediabetic state, insulin resistance is already present and that β cells simply increase insulin secretion to maintain glucose levels. Overt diabetes occurs when the β cells are exhausted by this compensatory mechanism (DeFronzo 1988). For those researchers who support the theory of impaired insulin secretion, they propose that the prediabetic state is first characterised by a subclinical defect in insulin secretion, but this secretion is still sufficient to maintain glucose levels. Impaired glucose tolerance develops because of the secondary superimposition of insulin resistance, and finally overt diabetes occurs because of the worsening of either insulin resistance or insulin secretion (Pimenta et al. 1995). At present it impossible to confirm either theory. One can only state that both defects are required for the appearance of clinical diabetes (Weir 1982).

9.7.3 Treatments of Type 2 Diabetes

In treating type 2 diabetes, the goal of patient treatment is to decrease the metabolic abnormalities associated with the disease and thus prevent further complications. Drugs are only part of the management of the disease, as patient education, diet and exercise also play a major role in maintaining health and extending life. Within type 2 diabetes patients there is a wide spectrum in the degree of impaired insulin secretion and action, and in the degree of obesity. These factors affect individual therapeutic strategies. Type 2 diabetes patients can be divided into six groups. They have to be first separated into obese and non-obese groups because weight loss will be crucial. They can then be divided into three main categories. Firstly, patients who still have sufficient β-cell mass and insulin sensitivity to maintain normal glucose levels through the control of energy intake and expenditure, secondly, patients who require additional help through an oral antidiabetic molecule, and thirdly, patients who absolutely need exogenous insulin to control their glycaemia. The treatments currently used in clinical practice include diet, exercise, sulphonylureas, biguanides, α-glucosidase inhibitors, thiazolidenediones and insulin. They all have different mechanisms and sites of action. They can be used individually or in combination in a stepwise mode according to the type of patient who has to be treated (Riddle 2000). Major efforts are currently under way in pharmaceutical companies to find and develop novel therapeutic approaches at different possible sites of action to control glucose level, delay disease progression and reduce late complications.

9.7.4 Proteomics for the Discovery of Treatment Targets for Type 2 Diabetes

There is a close relationship between drug treatment, protein expression and resulting physiological effects. Most of the time, pharmacological

intervention results in the regulation or modulation of gene-product expression, in a similar way that complex disease processes alter global protein expression. From this, we can assert that an ideal drug is one that restores global protein expression of a disturbed system (in our case, a metabolic disorder) to a normal state; however, it is quite unusual that a drug only modulates gene products implicated in the disorder. Most of the time, drugs also cause perturbations in the expression of proteins that are not involved in the disease process. This leads to side effects.

Quantitative protein expression changes due to drugs can be measured accurately enough to detect overlapping pathways. An example of this was presented by Anderson et al. (1996), who investigated the effects of five peroxisome proliferators on proteins in the liver of rats. They demonstrated that the peroxisome proliferators produced effects on protein abundance over wide time and dose ranges (Anderson et al. 1996). The commercialised thiazolidinediones, which include troglitazone (Rezulin™), pioglitazone (Actos™) and rosiglitazone (Avandia™), are a class of drugs that decrease insulin resistance. They promote peripheral glucose uptake and decrease hepatic glucose release. They also produce a fall in plasma insulin and triglyceride levels. They seem to restore the expression and translocation of GLUT-4 in adipocytes. However, the cellular mechanism by which they increase glucose uptake in muscle is unknown (Ciaraldi and Henry 1997). The detailed mechanism of action of this class of drugs is far from completely unravelled. However, they are known to act as activators of a nuclear receptor, the peroxisome proliferator activated receptor γ (PPARγ) which is mainly present in adipose tissue. This activation process leads to an increase in the sensitivity of tissues to insulin. Edvardsson et al. (1999a, b) have investigated the effect of PPAR agonists on type 2 diabetes models using 2-DE and protein identification. They studied liver proteins from ob/ob mice treated either with rosiglitazone or WY14,643 (PPARα agonist). They found that acylcoenzyme A oxidase, peroxisome bifunctional enzyme and 3-ketoacyl thiolase were upregulated by WY14,643 and to a lesser extent by rosiglitazone. These three proteins are involved in peroxisome fatty acid β-oxidation and are known to be markers of PPARγ activation. More recently, the same group demonstrated differences at the hepatic proteome level between lean and obese diabetic mice, to identify metabolic pathways modulated by WY14,643 and rosiglitazone (PPARγ agonist), and to discriminate their effects triggered to the closely related PPARα and PPARγ receptors. Both compounds upregulated enzymes implicated in lipogenesis. WY14,643 enhanced the level of the ketogenic hydroxymethylglutarylcoenzyme A synthase and normalised the expression of several enzymes involved in glycolysis, gluconeogenesis and amino acid metabolism. Rosiglitazone partially normalised the expression of enzymes involved in amino acid metabolism.

In our laboratory, we have developed proteomic strategies for the identification and validation of markers associated with type 2 diabetes and rosiglitazone. We investigated liver, white and brown adipose tissue, islets and

muscle proteins from genetically obese C57 Bl/6J lep/lep mice. This is an animal model of obesity, hyperinsulinemia, insulin resistance and mild type 2 diabetes. The identification of differentially expressed proteins provided evidence that modulation of actin binding, fatty acid and carbohydrate metabolism protein expression is an important characteristic of rosiglitazone mechanism of action. It also highlighted molecules of unknown function modulated by rosiglitazone, suggesting that these might be potential independent effects of rosiglitazone that contribute to improved insulin sensitivity (Sanchez et al. 2002). Finally, rosiglitazone increased carboxypeptidase B precursor protein expression in both lep/lep and normal islets, suggesting that this might be an independent effect of rosiglitazone that contributes to improve insulin processing (Sanchez et al. 2003).

In conclusion, the information extracted from the studies described above and others is crucial for the potential development of new pharmacological molecules, as it might guide in selecting the compounds with the best therapeutic profile, and/or with the lowest toxicity. Several new hypotheses were also generated through this work which may be addressed in future studies using other technologies.

9.8 Current Limitations and Future Directions of Proteomics for Medicine

Proteomics still faces many challenges in the foreseeable future. Particular limitations involve preanalytical, analytical and postanalytical proteomic workflows. These should be addressed to enhance the future of proteomics in medicine.

9.8.1 Preanalytical Issues

Preanalytical issues can be divided into at least six components

1. *The sample 'inside the patient' or intrinsic factors*. Samples should always be obtained at the same time during the day, in the same position, with the same food regimen or before meals, and in the same cycle of rest. Many internal factors modify the composition of any protein-containing samples; therefore, these factors should be controlled as much as possible to allow biochemically meaningful findings to be made from the results.
2. *The sample outside the patient or extrinsic factors*. When outside the patient, samples struggle to survive. They are forced to live in anaerobic conditions and modify their metabolism very rapidly. The overall protein and metabolite compositions are changed accordingly. Degradation occurs at a fast rate and, depending upon the sample container, other physiological or non-physiological phenomena can occur.

3. *The sample container*. For specimens such as blood, the sample container has a tremendous influence on the final sample. For example, glass containers with no anticoagulants induce blood coagulation via the activity of many peptidases, bringing about massive sample degradation. Plastic containers, by contrast, tend to absorb numerous peptides in low amounts.
4. *The sample transport and storage*. Fast sample transport in hospital vacuum systems can often break cells and induce haemolysis. Cooling the blood samples for transportation and storage can induce platelet degranulation, resulting in dramatic modification of the protein content of the sample. Even when stored at −20 or −80°C, samples continue to degrade or can be modified. Tissue biopsies frozen at −80°C get progressively dehydrated or lyophilised and, in time, become unsuitable for analysis with proteomic methods.
5. *Sample complexity*. This is one of the major challenges of proteomics. The chemical diversity of proteins is huge. Proteins can be very acidic or very basic, very hydrophilic or very hydrophobic, very small or very large. Their concentration varies from millimolar to femtomolar levels, more than 12 logs of difference. Some have a long half-life, some a very short one. Many of their modifications are labile. Careful prefractionation is increasingly required to undertake in-depth proteomic work.
6. *Sample preparation*. Sample preparation must balance the need for an analytical pipeline with the requirement to preserve the original composition of the proteins or peptides. It relates to the scientific questions that one is trying to solve. Unfortunately, there is and will be no unique or universal method for sample preparation. Instead, sample preparation is usually a compromise, which favours the analysis of one or another aspect of the proteome (see Chap. 2). Salts keep proteins in solution but are deleterious for electrophoresis and MS analysis. Lipids and sugars are protein post-translational modifications but add such tremendous complexity that they might need to be removed from some proteins to facilitate their analysis.

9.8.2 Analytical Aspects

The complexity of biological samples remains a key issue. As stated already, the tremendous chemical diversity of proteins and their enormous range of abundance are impediments to rapid progress in proteomics. The proteomic workflow requires massive parallel fractionation to analyse proteomes in more depth. It therefore remains a challenge to obtain reproducible, qualitative and quantitative, accurate and sensitive results. Further miniaturisation and automation should ultimately provide the analytical speed and robustness that is required.

9.8.3 Postanalytical Aspects

Proteomic workflows often produce an overwhelming amount of data. They can also yield information that cannot be humanly analysed or interpreted. As a result, much of it is lost in archives or placed in databases that become data cemeteries. To address this, software development is absolutely critical to help mine this vast amount of data and to retrieve relevant information (see Chap. 7). A very interesting consideration is that a lot of proteomic data can be correlated; this should be exploited to increase the signal-to-noise ratio in large data sets. Uncertainty should also be measured at all levels using appropriate probability calculations and statistics. Careful study designs must be used, as selection of proper negative and positive controls ultimately determines the value of clinical proteomic studies and assists biological and clinical interpretation. In a proteomic workflow to discover biomarkers or drug targets, relatively large clinical trials are required to obtain statistically significant data. In this case, access to patient samples and the costs of a large proteomic study can be a severe limitation.

9.9 Present and Future Directions

Proteomics has many roles to play in medicine. For discovery, proteomics is one of the best tools to study protein expression, post-translational modifications, protein–protein interactions and biochemical pathways. In clinical medicine, it will certainly uncover many important new biomarkers. It will highlight prognostic biomarkers, drug targets and even therapeutic molecules. It will certainly be used for studies of drug toxicity, for quality assurance to ensure sample integrity. In a routine mode, proteomics may also be used to establish multiple immunoassays on chips. If chromatographic separation and online MS/MS becomes fully automated, cheap and robust proteomics may be used to analyse biopsies and body fluids in routine clinical laboratories. Standardised and exchangeable software and databases are a prerequisite for its success, similarly to the field of therapeutic drug monitoring where multidimensional chromatography and MS have become routine practice.

References

Allard L, Lescuyer P, Burgess J, Leung KY, Ward M, Walter N, Burkhard PR, Corthals G, Hochstrasser DF, Sanchez JC (2004) ApoC-I and ApoC-III as potential plasmatic markers to distinguish between ischemic and hemorrhagic stroke. Proteomics 4:2242–2251

Allard L, Burkhard PR, Lescuyer P, Burgess J, Walter N, Hochstrasser DF, Sanchez JC (2005) PARK7 and NDKA as plasmatic markers for the early diagnosis of stroke. Clin Chem 51:2043–2051

Anderson NL, Esquer-Blasco R, Richardson F, Foxworthy P, Eacho P (1996) The effects of peroxisome proliferators on protein abundances in mouse liver. Toxicol Appl Pharmacol 137:75–89

Anonymous (1979) Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. National Diabetes Data Group. Diabetes 28:1039–1057

Armstrong, RA, Lantos PL, Cairns NJ (2005) Overlap between neurodegenerative disorders. Neuropathology 25:111–124

Ballabh P, Braun A, Nedergaard M (2004) The blood-brain barrier: an overview: structure, regulation, and clinical implications. Neurobiol Dis 16:1–13

Brennan JP, Wait R, Begum S, Bell JR, Dunn MJ, Eaton P (2004) Detection and mapping of widespread intermolecular protein disulfide formation during cardiac oxidative stress using proteomics with diagonal electrophoresis. J Biol Chem 279:41352–41360

Brown KJ, Fenselau C (2004) Investigation of doxorubicin resistance in MCF-7 breast cancer cells using shot-gun comparative proteomics with proteolytic ^{18}O labeling. J Proteome Res 3:455–462

Bruneel A, Labas V, Mailloux A, Sharma S, Vinh R, Vaubourdolle M, Baudin B (2003) Proteomic study of human umbilical vein endothelial cells in culture. Proteomics 3:714–723

Butterfield DA, Boyd-Kimball D, Castegna A (2003) Proteomics in Alzheimer's disease: insights into potential mechanisms of neurodegeneration. J Neurochem 86:1313–1327

Campbell RD, Milner CM (1993) MHC genes in autoimmunity. Curr Opin Immunol 5:887–893

Chaurand P, Stoeckli M, Caprioli RM (1999) Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. Anal Chem 71:5263–5270

Chaurand P, Sanders ME, Jensen RA, Caprioli RM (2004) Proteomics in diagnostic pathology: profiling and imaging proteins directly in tissue sections. Am J Pathol 165:1057–1068

Ciaraldi T, Henry RR (1997) Thiazolidinediones and their effects on glucose transporters. Eur J Endocrinol 137:610–612

Dasgupta B, Yi Y, Chen DY, Weber JD, Gutmann DH (2005) Proteomic analysis reveals hyperactivation of the mammalian target of rapamycin pathway in neurofibromatosis 1-associated human and mouse brain tumors. Cancer Res 65:2755–2760

DeFronzo RA (1988) Lilly lecture 1987. The triumvirate: beta-cell, muscle, liver. A collusion responsible for NIDDM. Diabetes 37:667–687

DeKosky ST, Marek K (2003) Looking backward to move forward: early detection of neurodegenerative disorders. Science 302:830–834

De Souza AI, Wait R, Mitchell AG, Banner NR, Dunn MJ, Rose ML (2005) Heat shock protein 27 is associated with freedom from graft vasculopathy after human cardiac transplantation. Circ Res 97:192–198

Diamandis EP (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol Cell Proteomics 3:367–378

Donners MM, Verluyten MJ, Bouwman FG, Mariman EC, Devreese B, Vanrobaeys F, van Beeumen J, van den Akker LH, Daemen MJ, Heeneman S (2005) Proteomic analysis of differential protein expression in human atherosclerotic plaque progression. J Pathol 206:39–45

Dupont A, Corseaux D, Dekeyzer O, Drobecq H, Guihot AL, Susen S, Vincentelli A, Amouyel P, Jude B, Pinet F (2005) The proteome and secretome of human arterial smooth muscle cells. Proteomics 5:585–596

Duran MC, Mas S, Martin-Ventura JL, Meilhac O, Michel JB, Gallego-Delgado J, Lazaro A, Tunon J, Egido J, Vivanco F (2003) Proteomic analysis of human vessels: application to atherosclerotic plaques. Proteomics 3:973–978

Edvardsson, U, Alexandersson M, Brockenhuus von Lowenhielm H, Nystrom AC, Ljung B, Nilsson F, Dahllof B (1999a) A proteome analysis of livers from obese (ob/ob) mice treated with the peroxisome proliferator WY14,643. Electrophoresis 20:935–942

Edvardsson U, Bergstrom M, Alexandersson M, Bamberg K, Ljung B, Dahllof B (1999b) Rosiglitazone (BRL49653), a PPARgamma-selective agonist, causes peroxisome proliferator-like liver effects in obese mice. J Lipid Res 40:1177–1184

Fach EM, Garulacan LA, Gao J, Xiao Q, Storm SM, Dubaquie YP, Hefta SA, Opiteck GJ (2004) *In vitro* biomarker discovery for atherosclerosis by proteomics. Mol Cell Proteomics 3:1200–1210

Finehout EJ, Franck Z, Lee KH (2004) Towards two-dimensional electrophoresis mapping of the cerebrospinal fluid proteome from a single individual. Electrophoresis 25:2564–2575

Franzen B, Hirano T, Okuzawa K, Uryu K, Alaiya AA, Linder S, Auer G (1995) Sample preparation of human tumors prior to two-dimensional electrophoresis of proteins. Electrophoresis 16:1087–1089

Fukada K, Zhang F, Vien A, Cashman NR, Zhu H (2004) Mitochondrial proteomic analysis of a cell line model of familial amyotrophic lateral sclerosis. Mol Cell Proteomics 3:1211–1223

Gianazza E, Sironi L (2004) Vasculature, vascular disease and atherosclerosis. In: Sanchez JC, Corthals GL, Hochstrasser DF (eds) Biomedical applications of proteomics. Wiley-VCH, Weinheim, pp 39-55

Guerrini U, Sironi L, Tremoli E, Cimino M, Pollo B, Calvio AM, Paoletti R, Asdente M (2002) New insights into brain damage in stroke-prone rats: a nuclear magnetic imaging study. Stroke 33:825–830

Guillaume E, Zimmermann C, Burkhard PR, Hochstrasser DF, Sanchez JC (2003) A potential cerebrospinal fluid and plasmatic marker for the diagnosis of Creutzfeldt-Jakob disease. Proteomics 3:1495–1499

Hamacher M, Klose J, Rossier J, Marcus K, Meyer HE (2004) Does understanding the brain need proteomics and does understanding proteomics need brains? Second HUPO HBPP Workshop hosted in Paris. Proteomics 4:1932–1934

Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100:57–70

Hsich G, Kenney K, Gibbs CJ, Lee KH, Harrington MG (1996) The 14-3-3 brain protein in cerebrospinal fluid as a marker for transmissible spongiform encephalopathies. N Engl J Med 335:924–930

Isoda N, Kajii E, Ikemoto S, Kimura K (1990) Two-dimensional polyacrylamide gel electrophoretic pattern of duodenal tumour proteins. J Chromatogr 534:47–55

Jain KK (2002) Recent advances in oncoproteomics. Curr Opin Mol Ther 4:203–209

Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ (2005) Cancer statistics, 2005. Cancer J Clin 55:10–30

Klose J, Nock C, Herrmann M, Stuhler K, Marcus K, Bluggel M, Krause E, Schalkwyk LC, Rastan S, Brown SD, Bussow K, Himmelbauer H, Lehrach H (2002) Genetic analysis of the mouse brain proteome. Nat Genet 30:385–393

Lescuyer P, Allard L, Zimmermann-Ivol CG, Burgess JA, Hughes-Frutiger S, Burkhard PR, Sanchez JC, Hochstrasser DF (2004) Identification of post-mortem cerebrospinal fluid proteins as potential biomarkers of ischemia and neurodegeneration. Proteomics 4:2234–2241

Lescuyer P, Gandini A, Burkhard PR, Hochstrasser DF, Sanchez JC (2005) Prostaglandin D2 synthase and its post-translational modifications in neurological disorders. Electrophoresis 26:4563–4570

Li C, Hong Y, Tan YX, Zhou H, Ai JH, Li SJ, Zhang L, Xia QV, Wu JR, Wang HY, Zeng R (2004) Accurate qualitative and quantitative proteomic analysis of clinical hepatocellular carcinoma using laser capture microdissection coupled with isotope-coded affinity tag and two-dimensional liquid chromatography mass spectrometry. Mol Cell Proteomics 3:399–409

Liao L, Cheng D, Wang J, Duong DM, Losik TG, Gearing M, Rees HD, Lah JJ, Levey AI, Peng J (2004) Proteomic characterization of postmortem amyloid plaques isolated by laser capture microdissection. J Biol Chem 279:37061–37068

Liotta LA, Espina V, Mehta AI, Calvert V, Rosenblatt K, Geho D, Munson PJ, Young L, Wulfkuhle J, Petricoin EF 3rd (2003) Protein microarrays: meeting analytical challenges for clinical applications. Cancer Cell 3:317–325

Lubec G, Krapfenbauer K, Fountoulakis M (2003) Proteomics in brain research: potentials and limitations. Prog Neurobiol 69:193–211

Lusis AJ (2000) Atherosclerosis. Nature 407:233–241

Maccarrone G, Milfay D, Birg I, Rosenhagen M, Holsboer F, Grimm R, Bailey J, Zolotarjova N, Turck CW (2004) Mining the human cerebrospinal fluid proteome by immunodepletion and shotgun mass spectrometry. Electrophoresis 25:2402–2412

Mateos-Caceres PJ, Garcia-Mendez A, Lopez Farre A, Macaya C, Nunez A, Gomez J, Alonso-Orgaz S, Carrasco C, Burgos ME, de Andres R, Granizo JJ, Farre J, Rico LA (2004) Proteomic analysis of plasma from patients during an acute coronary syndrome. J Am Coll Cardiol 44:1578–1583

Matuszek MA, Aristoteli LP, Bannon PG, Hendel PN, Hughes CF, Jessup W, Dean RT, Kritharides L (2003) Haptoglobin elutes from human atherosclerotic coronary arteries – a potential marker of arterial pathology. Atherosclerosis 168:389–396

Maurer MH, Berger C, Wolf M, Futterer CD, Feldmann RE Jr, Schwab S, Kuschinsky W (2003) The proteome of human brain microdialysate. Proteome Sci 1:7

Mayr M, Mayr U, Chung YL, Yin X, Griffiths JR, Xu Q (2004) Vascular proteomics: linking proteomic and metabolomic changes. Proteomics 4:3751–3761

Minino AM, Arias E, Kochanek KD, Murphy SL, Smith BL (2002) Deaths: final data for 2000. Natl Vital Stat Rep 50:1–119

Nielsen PA, Olsen JV, Podtelejnikov AV, Andersen JR, Mann M, Wisniewski JR (2005) Proteomic mapping of brain plasma membrane proteins. Mol Cell Proteomics 4:402–408

Ornstein DK, Gillespie JW, Paweletz CP, Duray PH, Herring J, Vocke CD, Topalian SL, Bostwick DG, Linehan WM, Petricoin EF 3rd, Emmert-Buck MR (2000) Proteomic analysis of laser capture microdissected human prostate cancer and in vitro prostate cell lines. Electrophoresis 21:2235–2242

Page MJ, Amess B, Townsend RR, Parekh R, Herath A, Brusten L, Zvelebil MJ, Stein RC, Waterfield MD, Davies SC, O'Hare MJ (1999) Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties. Proc Natl Acad Sci USA 96:12589–12594

Pavy FW (1894) Br Med J 23:1349–1350

Perluigi M, Poon HF, Maragos W, Pierce WM, Klein JB, Calabrese V, Cini C, De Marco C, Butterfield DA (2005) Proteomic analysis of protein expression and oxidative modification in R6/2 transgenic mice – a model of Huntington's disease. Mol Cell Proteomics 4:1849–1861

Pimenta W, Korytkowski M, Mitrakou A, Jenssen T, Yki-Jarvinen H, Evron W, Dailey G, Gerich J (1995) Pancreatic beta-cell dysfunction as the primary genetic lesion in NIDDM. Evidence from studies in normal glucose-tolerant individuals with a first-degree NIDDM relative. JAMA 273:1855–1861

Puchades M, Hansson SF, Nilsson CL, Andreasen N, Blennow K, Davidsson P (2003) Proteomic studies of potential cerebrospinal fluid protein markers for Alzheimer's disease. Brain Res Mol Brain Res 118:140–146

Riddle M (2000) Combining sulfonylureas and other oral agents. Am J Med 108(Suppl 6):15S–22S

Sanchez JC, Converset V, Nolan A, Schmid G, Wang S, Heller M, Sennitt MV, Hochstrasser DF, Cawthorne MA (2002) Effect of rosiglitazone on the differential expression of diabetes-associated proteins in pancreatic islets of C57Bl/6 lep/lep mice. Mol Cell Proteomics 1:509–516

Sanchez JC, Converset V, Nolan A, Schmid G, Wang S, Heller M, Sennitt MV, Hochstrasser DF, Cawthorne MA (2003) Effect of rosiglitazone on the differential expression of obesity and insulin resistance associated proteins in lep/lep mice. Proteomics 3:1500–1520

Scherl A, Coute Y, Deon C, Calle A, Kindbeiter K, Sanchez JC, Greco A, Hochstrasser DF, Diaz JJ (2002) Functional proteomic analysis of human nucleolus. Mol Biol Cell 13:4100–4109

Schrattenholz A, Wozny W, Klemm M, Schroer K, Stegmann W, Cahill MA (2005) Differential and quantitative molecular analysis of ischemia complexity reduction by isotopic labeling of proteins using a neural embryonic stem cell model. J Neurol Sci 229:261–267

Schrimpf SP, Meskenaite V, Brunner E, Rutishauser D, Walther P, Eng J, Aebersold R, Sonderegger P (2005) Proteomic analysis of synaptosomes using isotope-coded affinity tags and mass spectrometry. Proteomics 5:2531–2541

Steinacker P, Mollenhauer B, Bibl M, Cepek L, Esselmann H, Brechlin P, Lewczuk P, Poser S, Kretzschmar HA, Wiltfang J, Trenkwalder C, Otto M (2004) Heart fatty acid binding protein as a potential diagnostic marker for neurodegenerative diseases. Neurosci Lett 370:36–39

Stewart BW, Kleihues P (eds) (2003) World cancer report. IARC, Lyons, pp 1–352

Suzuyama K, Shiraishi T, Oishi T, Ueda S, Okamoto H, Furuta M, Mineta T, Tabuchi K (2004) Combined proteomic approach with SELDI-TOF-MS and peptide mass fingerprinting identified the rapid increase of monomeric transthyretin in rat cerebrospinal fluid after transient focal cerebral ischemia. Brain Res Mol Brain Res 129:44–53

Villanueva J, Philip J, Entenberg D, Chaparro CA, Tanwar MK, Holland EC, Tempst P (2004) Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. Anal Chem 76:1560–1570

Wang Q, Woltjer RL, Cimino PJ, Pan C, Montine KS, Zhang J, Montine TJ (2005) Proteomic analysis of neurofibrillary tangles in Alzheimer disease identifies GAPDH as a detergent-insoluble paired helical filament tau binding protein. FASEB J 19:869–871

Weir GC (1982) Non-insulin-dependent diabetes mellitus: interplay between B-cell inadequacy and insulin resistance [editorial]. Am J Med 73:461–464

Xiao Z, Prieto D, Conrads TP, Veenstra TD, Issaq HJ (2005) Proteomic patterns: their potential for disease diagnosis. Mol Cell Endocrinol 230:95–106

You SA, Archacki SR, Angheloiu G, Moravec CS, Rao S, Kinter M, Topol EJ, Wang Q (2003) Proteomic approach to coronary atherosclerosis shows ferritin light chain as a significant marker: evidence consistent with iron hypothesis in atherosclerosis. Physiol Genomics 13:25–30

Zabel C, Chamrad DC, Priller J, Woodman B, Meyer HE, Bates GP, Klose J (2002) Alterations in the mouse and human proteome caused by Huntington's disease. Mol Cell Proteomics 1:366–375

Zerr I, Bodemer M, Otto M, Poser S, Windl O, Kretzschmar HA, Gefeller O, Weber T (1996) Diagnosis of Creutzfeldt-Jakob disease by two-dimensional gel electrophoresis of cerebrospinal fluid. Lancet 348:846–849

Zhang J, Goodlett DR, Peskind ER, Quinn JF, Zhou Y, Wang Q, Pan C, Yi E, Eng J, Aebersold R, Montine TJ (2005a) Quantitative proteomic analysis of age-related changes in human cerebrospinal fluid. Neurobiol Aging 26:207–227

Zhang J, Goodlett DR, Quinn JF, Peskind E, Kaye JA, Zhou Y, Pan C, Yi E, Eng J, Wang Q, Aebersold R, Montine TJ (2005b) Quantitative proteomics of cerebrospinal fluid from patients with Alzheimer disease. J Alzheimers Dis 7:125–133

Zhao Y, Zhang W, Kho Y (2004) Proteomic analysis of integral plasma membrane proteins. Anal Chem 76:1817–1823

Zimmermann-Ivol CG, Burkhard PR, Le Floch-Rohr J, Allard L, Hochstrasser DF, Sanchez JC (2004) Fatty acid binding protein as a serum marker for the early diagnosis of stroke: a pilot study. Mol Cell Proteomics 3:66–72

Zimmet P, McCarthy D (1995) The NIDDM epidemic – global estimates and projections. IDF Bull 40:8–16

Zimmet PZ (1995) The pathogenesis and prevention of diabetes in adults: genes, autoimmunity, and demography. Diabetes Care 18:1050–1064

10 Proteomics: Where to Next?

Keith L. Williams and Denis F. Hochstrasser

Abstract

In this chapter we review the changes to proteomics over the last decade and suggest where the field will go over the next 10 years. In the early days of proteomics, the focus was on large-scale protein identification either in a whole organisms or tissue. Changes in protein expression were often compared with emergence of disease etc. Today, the focus of basic research is moving towards protein characterisation (with emphasis on post-translational modifications) and protein–protein interaction mapping, with new informatics needed. Increasingly this leads to studies on function and towards practical applications of new discoveries. We suggest that proteomics is now close to providing the markers for new-generation diagnostic and prognostic tests, as well as the basis for finding new-generation drugs.

10.1 Introduction

When we wrote the first book on proteomics (Wilkins et al. 1997), the genomics revolution was maturing but the full sequence of the human genome was still a far-off dream. The significance of genomics was still not fully understood, and proteomics was merely a gleam in the eye and the depth of proteome complexity was not fully anticipated.

In 2006, the human genome sequencing was essentially complete and hundreds of genomic sequences are publicly available (Ensembl 2007). Analysis of gene expression using microarrays is becoming increasingly reliable and has started to be used in clinical diagnostics. Databases of normal and diseased tissues, developmental time courses, etc. are becoming accessible not only for human data (SymAtlas 2007), but also for an ever-increasing library of organisms.

Transcriptomic studies provide vast amounts of expression data for most human tissues and new technologies such as massively parallel signature sequencing (Jongeneel et al. 2005) are emerging. As might be expected, at this early stage there is some ambiguity in the data as a result of different

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

approaches. An immediate need is to produce better-quality genomics data and also to systematise it.

Gene silencing through RNA interference has become a potent way of studying the role of the many proteins that can be produced from a single gene product. Craig Venter's (2004) exploration of the genomes in the oceans is demonstrating how little we know of the microorganisms on planet earth. All of the above are contributing to ever-greater challenges in mining vast amounts of data. Indeed, it may be argued that currently we are becoming more confused than ever because of our inability to make sense of this explosion of information.

Proteomics is now seen as a technology to help dissect useful information from the masses of DNA-based microarray data, because with proteomics one sees the outcomes from gene expression that have been translated into protein product. Hence, it is one step closer to function.

The late 1990s saw the rise of genomics companies and a landgrab of patenting human genes. At one stage, Incyte had more than 1,000 gene patents. Hundreds of millions of dollars were raised in building these genomics companies and a huge amount of DNA sequencing was done, gene families were fleshed out and attempts were made to protect intellectual property. However, it became clear that it is difficult to sustain a business without products and all of the genomics companies have now shifted their focus towards the development of drugs. Proteomics followed a similar path, with the formation of several companies that raised hundreds of millions of dollars to unravel the human proteome. Unlike the genomic revolution where sequencing the human genome was a clear goal that was greatly accelerated by the corporate drive from Craig Venter and Celera, the human proteome is a much more elusive goal and, indeed, proteomics is an enabling technology rather than an end in itself. By 2006, high-profile proteomics companies including Oxford GlycoSciences, GeneProt, Myriad Proteomics, MDS Proteomics, Proteome Systems and Large Scale Biology Corporation had either closed their operations or moved in the direction of drug development. Unlike genomics, where there was a clear goal that was achieved, but which is yet to translate to successful business activity, the challenges of proteomics have been more in the execution and the complexity of the proteome. We are still learning how proteomics can be best addressed, although it is becoming clear that proteomics is most useful when applied to solving particular problems.

10.2 The Relevance of -omics to Biology

Looking back over the last decade it is interesting to think about the changes that have occurred in the way biology is viewed and, more importantly, conducted. The definition of genomics and proteomics has had a big impact

on how biology is done. It is now accepted that genomic and proteomic technologies are generic, breaking down previous barriers across the whole of biology (from medicine to microbiology to agricultural research). Institutions now expect previously small groups to interact and decide on how to best expend precious resources, and granting bodies require consolidation so that funding can be effective. The challenge has been to support 'big science' initiatives, while at the same time leaving open the opportunity for the small creative group to disrupt how the world is viewed.

So in some respects, genomics and proteomics have served as words to herald new ways of approaching complex problems in biology. They led to many new '-omics' terms (e.g. metabolomics, phenomics, glycomics) and all of the above have been subsumed under a bigger banner of 'systems biology'. Of course, like all fashions, things wax and wane. Currently much proteomics research comes under the banner of 'biomarker discovery'. Genomics (primary information), transcriptomics (secondary information), proteomics (functional molecules) and metabolomics (small molecules often reflecting protein pathways) most probably have a long-term future as they investigate critical aspects of biology that are central to modern research and biological outcomes.

More significantly, proteomics is the technology that enables moving beyond protein expression to understanding the actual form of the protein produced and its interacting partners. This is important because it is now understood that the form of the protein produced is critical to cell location and function. Form and function result not just from the protein produced, but from how it is produced. How is the gene spliced in making the protein? Is the protein co- or post-translationally modified? If so, what modifications are present (glycosylation, phosphorylation or others)?

10.3 Technological Developments in Proteomics

While the importance of protein post-translational modifications is now increasingly well understood, there is still a major focus on protein identification in its general sense as opposed to focusing on how the protein is modified (see Chap. 3). This, in part, reflects the maturation of proteomic technologies. In the 1990s the focus was on parallel protein identification using mass spectrometry techniques and much of our first proteomics book addressed this problem. Indeed, many of the global technologies address protein identification (notably tandem mass spectrometry identification of proteins based on the fragmentation of peptides and protein chip applications based on the capture of proteins by arrays of antibodies). While both of these technologies can produce very large numbers of identifications at the gene level, by attributing a protein to a specific gene product, they have less to say about the actual final form of the protein in terms of how the gene has been spliced or how the protein is modified.

10.3.1 Characterising Modifications

Today the focus is shifting to technologies for sensitively identifying not only the proteins but also their modifications (see Chap. 5). There are several ways that this can be approached. Firstly, one can use capture techniques for specific modifications, such as lectins for glycosylation or metal-affinity resins for phosphorylation. These technologies help simplify the problem, but are restricting in that you need to know what you are looking for.

An emerging area involves studying whole proteins: so called 'top-down' proteomics (see Chap. 3). While the technology is making possible accurate mass measurement of large protein molecules, the issue remains whether detailed characterisation is possible on the basis of a single mass measurement. Given the large number of possible modifications to a single protein, the chances of fully characterising a protein solely on the basis of measurement of its intact mass will remain challenging for some time.

Another approach is to purify protein isoforms and seek to fully characterise each isoform. Here, well-established 2-D gel technology is useful, because this technique arrays proteins and is excellent for the separation of isoforms. When coupled to solid-phase analysis, such as chemical printer technology (Cooley et al. 2005), the opportunities for detailed characterisation are near-term. A variation on this theme involves stretching out separation of proteins solely on the basis of their isoelectric point (see Chap. 2).

Yet another approach is to study specific peptides for many relevant proteins in a defined biological sample and to monitor them quantitatively with high sensitivity and specificity using specific or multiple reaction monitoring technology. This technology was developed in the late 1970s but it has been refreshed with new instrumentation (Anderson and Hunter 2006). The development of the means for systematising peptide data is critical for this work (Deutsch et al. 2005).

10.3.2 Global Tissue Analysis

Genomics and proteomics were initially used to understand the whole organism or its organs. There are now extensive libraries of microarray data for different tissues and there are several proteomics initiatives co-ordinated by the Human Proteome Organisation including human plasma, the liver and the brain (Human Proteome Organisation 2007). These are bold cataloguing initiatives that are still largely focused on protein identification rather than studying spliced forms of proteins or their modifications.

Recently, by marrying *in situ* hybridisation of tissue sections with large-scale genomics analysis, ambitious studies have been commenced to look at specific organisation of gene expression within tissues. For example, The Allen Institute for Brain Science has a programme for cataloguing gene expression of the whole mouse brain. In this study the brain is sectioned and

these sections are probed against a very large library of genes to get a clear understanding of the expression of particular genes regionally throughout the brain. This is a major undertaking on 56-day-old male mice of strain C57BL/6J. Once this description has been assembled, the next step will be to study different mutants. Results of this project are available in an online database (Allen Brain Atlas 2007).

The next step will be to develop a proteomic approach to similar studies. The Human Protein Atlas is one such study (see Chap. 7). There are many challenges in this work, not the least of which is the numbers game of finding sufficiently sensitive technologies to be able to detect the interesting rare proteins, and thus transcend the small group of abundant proteins at the cellular level in tissue sections. There are, however, interesting new technologies under development whereby tissue sections are being treated and proteins or peptides are desorbed and analysed by mass spectrometry. These include the molecular scanner (Bienvenut et al. 1999) and chemical inkjet printing (Sloane et al. 2002), discussed in Chap. 6. This is a new horizon and it may well be another decade before technologies are sufficiently sensitive to do more than a cursory examination of the spatial localisation of proteins.

In summary, despite proteomics being almost 10 years old, at least by name, we still remain at an early stage of the proteomics revolution. Clearly proteins can be screened and identified on a mass scale, and increasingly sophisticated protein interaction maps are being constructed. However, we are still just beginning to catalogue the modified forms of proteins and interaction studies based on authentic proteins are rare, being largely undertaken with yeast two-hybrid approaches. There is much technology development to do before the hard work of providing a catalogue of a native proteome is possible. This is likely to occupy a large number of researchers in the next decade. Then and only then will it be possible to begin to develop the interaction maps for authentic (as opposed to gene product based) structural proteomics.

10.4 The Next Steps for Proteomics: Diagnostics and Drugs

Cataloguing remains a major focus for fundamental research in biology because you have to know what is there before you can understand what it does. However, the ultimate focus of biological research is to produce useful outcomes for medical, agricultural and industrial activities. Proteomic technology is sufficiently developed to begin making useful discoveries. The most obvious of these involves (1) using proteomics to further the development of diagnostic and prognostic technologies for identifying and monitoring treatment of disease and (2) use of proteomics as the basis for development of new drugs.

10.4.1 Diagnostics

Disease, whether it is of environmental (infectious or toxic) or genetic origin, leads to organismal change. The challenge is to find evidence of such changes at an early stage, because early identification invariably leads to more effective treatment. Also, having the capacity to monitor the effectiveness of treatment allows one to fine-tune it.

There are two kinds of changes that happen in relation to disease. Firstly, there are molecules that are specific to the disease usually present in small quantities. In the case of tuberculosis, it is possible using proteomics to identify proteins originating from *Mycobacterium tuberculosis* in blood or sputum of patients with active disease. These proteins can be used as the basis of an antigen-based test to quantify the level of infection and monitor treatment. In the case of cancer, various studies are in progress to identify a small group of specific markers or a pattern of changes that reflect the cancer. In situations where disease needs to be monitored, it may be possible to use more abundant proteins to monitor the progression of the disease in an individual. In the case of cystic fibrosis, this may be certain inflammatory proteins. This may have major advantages for the treatment of the individual.

One important point to understand about diagnostics is that whilst proteomic technology used to make fundamental discoveries of the important biomarkers is complex, the goal for diagnostics should be simplicity. Ideally we need fast, simple, cheap devices that allow diagnosis at a distance (home, office, bedside) but which also have informational connectivity so that results can be assembled and reviewed at a distance.

10.4.2 Drugs

A clear goal for proteomics has been to identify targets for new-generation drugs. Almost all drug targets are proteins, so the outcomes of proteomics research represent the bread and butter of drug development. For a time, it was hoped that drug development could be done by a combination of genomics and *in silico* protein reconstruction. It is now clear that the world is not so simple and in most cases proteins need to be studied in order to develop drugs. While there have been dramatic advances in analysis of protein structure by NMR and X-ray crystallography, these fields are still largely limited to analysis of proteins without their modifications. There remains a major challenge to generate structures of authentic proteins complete with their glycosylation and other modifications. This is important because with many proteins, the modifications can occupy a large part of the molecule and, hence, are relevant if one is developing molecules to target the protein. Also, since proteins often exist in different forms in different tissues, the ability to specifically target modified forms of the protein of interest is likely to lead to fewer problems with side effects. In addition, proteomics can be used to

monitor drug development and to seek evidence of toxic side effects early in the drug-development process by analysing proteomic outcomes of treatment with a potential drug at an early stage (see Chap. 9).

Traditionally, drug development has involved seeking small molecules to target the proteins of interest; however, this situation has changed dramatically in the last 5 years, such that a large number of new drugs under development are proteins rather than small molecules. For example, it is now well recognised that antibodies can act as effective therapeutic agents. Whether in the longer term the drug industry will become an industry largely using proteins as the drugs or whether we are going through an interim stage before we get back to small-molecule drugs is a question that the next decade will begin to answer.

10.5 Conclusions

This brief attempt at crystal-balling is done with a much bigger palette than was available when we wrote the first version of this book. The challenges are there and proteomics remains an exciting area for young people entering the life sciences, whether it be with a technological bent, from a basic research perspective, or with a view to solving practical problems in medicine and agriculture.

References

Allen Brain Atlas (2007) Neuroscience gateway. Allen Institute for Brain Science, Seattle, and Nature Publishing Group, New York. http://www.brainatlas.org/aba/. Cited 14 Mar 2007

Anderson L, Hunter CL (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics 5:573–588

Bienvenut W, Sanchez JC, Karmime A, Rouge V, Rose K, Binz PA, Hochstrasser DF (1999) Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot. Anal Chem 71:4800–4807

Cooley PW, Joss JL, Hopwood FG, Wilson NL, Gooley AA (2005) The *in situ* characterisation of membrane-immobilized 2-D PAGE-separated proteins using ink-jet technology. In: Walker JM (ed) The Proteomics protocols handbook. Humana, Totowa, pp 341–354

Deutsch EW, Eng JK, Zhang H, King NL, Nesvizhskii AI, Lin B, Lee H, Yi EC, Ossola R, Aebersold R (2005) Human Plasma PeptideAtlas. Proteomics 5:3497–3500

Ensembl (2007) EMBL-EBI and The Sanger Institute, Cambridge. http://www.ensembl.org. Cited 14 Mar 2007

Human Proteome Organisation (2007) HUPO Website. http://www.hupo.org. Cited 14 Mar 2007

Jongeneel CV, Delorenzi M, Iseli C, Zhou D, Haudenschild CD, Khrebtukova I, Kuznetsov D, Stevenson BJ, Strausberg RL, Simpson AJG, Vasicek TJ (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). Genome Res 15:1007–1014

Sloane AJ, Duff JL, Wilson NL, Gandhi PS, Hill CJ, Hopwood FG, Smith PE, Thomas ML, Cole RA, Packer NH, Breen EJ, Cooley PW, Wallace DB, Williams KL, Gooley AA (2002) High throughput peptide mass fingerprinting and protein macroarray analysis using chemical printing strategies. Mol Cell Proteomics 1:490–499

SymAtlas (2007) Genomics Institute of the Novartis Research Foundation, San Diego. http://symatlas.gnf.org/SymAtlas/. Cited 14 Mar 2007

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Wilkins MR, Williams KL, Appel RD, Hochstrasser DF (eds) (1997) Proteome research: new frontiers in functional genomics. Springer, Berlin, pp 1–243

Index