

Dmitrij Frishman
Alfonso Valencia
Editors

Modern Genome Annotation

The BioSapiens Network

 SpringerWienNewYork

 SpringerWienNewYork

Dmitrij Frishman
Alfonso Valencia
Editors

Modern Genome Annotation

The Biosapiens Network

SpringerWienNewYork

Prof. Dmitrij Frishman

TU München, Wissenschaftszentrum Weihenstephan, Freising, Germany

Alfonso Valencia

Spanish National Cancer Research Centre, Structural and Computational Programme, Madrid, Spain

This work is subject to copyright.

All rights are reserved, whether the whole or part of the material is concerned, specifically those of translation, reprinting, re-use of illustrations, broadcasting, reproduction by photocopying machines or similar means, and storage in data banks.

Product Liability: The publisher can give no guarantee for all the information contained in this book. This does also refer to information about drug dosage and application thereof. In every individual case the respective user must check its accuracy by consulting other pharmaceutical literature. The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

© 2008 Springer-Verlag/Wien

Printed in Austria

SpringerWienNewYork is part of
Springer Science+Business Media
springer.at

Typesetting: Thomson Press (India) Ltd., Chennai, India

Printing: Holzhausen Druck und Neue Medien GmbH, 1140 Wien, Austria

Printed on acid-free and chlorine-free bleached paper

SPIN: 12121748

With numerous (partly coloured) Figures

Library of Congress Control Number: 2008934450

ISBN 978-3-211-75122-0 SpringerWienNewYork

BioSapiens Partners



CENTRO
RBIOLGI
CALSEQU
ENCEANA
LYSIS CBS



EMBL-EBI



EMBL



UNIVERSITÉ
DE GENÈVE
FACULTÉ DE MÉDECINE



UNIVERSITY OF HELSINKI



INSTITUT PASTEUR

HelmholtzZentrum münchen

German Research Center for Environmental Health



האוניברסיטה העברית בירושלים
The Hebrew University of Jerusalem



max planck institut
informatik



MPIMG



SAPIENZA
UNIVERSITÀ DI ROMA



Stockholms
universitet



CONTENTS

Introduction

BIOSAPIENS: A European Network of Excellence

to develop genome annotation resources (*D. Frishman, A. Valencia*) 1

SECTION 1: Gene definition

Chapter 1.1

State of the art in eukaryotic gene prediction (*T. Alioto, R. Guigó*) 7

- 1 Introduction 7
- 2 Classes of information 10
 - 2.1 Extrinsic information 10
 - 2.2 Intrinsic information 11
 - 2.2.1 Signals 11
 - 2.2.2 Content 14
 - 2.3 Conservation 15
- 3 Frameworks for integration of information 17
 - 3.1 Exon-chaining 17
 - 3.2 Generative models: Hidden Markov models 18
 - 3.2.1 Basic hidden Markov models 18
 - 3.2.2 Generalized hidden Markov models 21
 - 3.2.3 Generalized pair HMMs 21
 - 3.2.4 Phylo-HMMs or evolutionary HMMs 23
 - 3.3 Discriminative learning 24
 - 3.3.1 Support vector machines 24
 - 3.3.2 Semi-Markov conditional random fields 25
 - 3.4 Combiners 25
- 4 Training 26
- 5 Evaluation of gene prediction methods 27
 - 5.1 The basic tools 27
 - 5.2 Systematic evaluation 28

5.3	The community experiments	29
5.3.1	GASP	29
5.3.2	EGASP	30
5.3.3	NGASP	30
6	Discussion	32
6.1	Genome datasets	32
6.2	Atypical genes	32
6.3	Outstanding challenges to gene annotation	33
6.4	What is the right gene prediction strategy?	34

Chapter 1.2

Quality control of gene predictions (*A. Nagy, H. Hegyi, K. Farkas, H. Tordai, E. Kozma, L. Bányai, L. Patthy*) 41

1	Introduction	41
2	Quality control of gene predictions	42
2.1	Principles of quality control	42
2.1.1	Violation of some generally valid rules about proteins	42
2.1.1.1	Conflict between the presence of extracellular Pfam-A domain(s) in a protein and the absence of appropriate sequence signals	43
2.1.1.2	Conflict between the presence of extracellular and cytoplasmic Pfam-A domains in a protein and the absence of transmembrane segments	43
2.1.1.3	Co-occurrence of nuclear and extracellular domains in a predicted multidomain protein	43
2.1.1.4	Domain size deviation	44
2.1.2	Violation of some generally valid rules of protein-coding genes	44
2.1.2.1	Chimeric proteins parts of which are encoded by exons located on different chromosomes	44
3	Results	44
3.1	Validation of the MisPred pipeline	44
3.2	Errors detected by the MisPred tools in public databases	46
3.2.1	Analysis of the TrEMBL section of UniProtKB	46
3.2.2	Analysis of sequences predicted by the Ensembl and GNOMON gene prediction pipelines	47
4	Alternative interpretations of the results of MisPred analyses	50
4.1	MisPred has a low false positive rate	50
4.2	MisPred detects errors in gene prediction	50

- 4.3 MisPred detects “errors” of biological processes 51
- 4.4 MisPred discovers exceptions to generally valid rules 51
- 5 Conclusions 52

SECTION 2: Gene regulation and expression

Chapter 2.1

Evaluating the prediction of cis-acting regulatory elements in genome sequences (*O. Sand, J. Valéry Turatsinze, J. van Helden*) 55

- 1 Introduction 55
- 2 Transcription factor binding sites and motifs 58
- 3 Scanning a sequence with a position-specific scoring matrix 59
 - 3.1 Background probability 61
 - 3.2 Probability of a sequence segment given the motif 62
 - 3.3 Scanning profiles 63
- 4 Evaluating pattern matching results 64
 - 4.1 Evaluation statistics 64
 - 4.2 Accuracy profiles 66
 - 4.3 Avoiding circularity in the evaluation 67
 - 4.4 Why the statistics involving TN should be avoided 67
 - 4.5 Difficulties for the evaluation of pattern matching 68
- 5 Discovering motifs in promoter sequences 69
 - 5.1 Example of pattern discovery result 70
 - 5.2 Evaluation statistics 72
 - 5.3 Correctness of predicted motifs for a collection of annotated regulons 73
 - 5.4 Distributions of motif scores in positive and negative testing sets 77
 - 5.5 The Receiver Operating Characteristics (ROC) curve 81
 - 5.6 Using ROC curves to find optimal parameters 83
- 6 Methodological issues for evaluating pattern discovery 83
- 7 Good practices for evaluating predictive tools 84
 - 7.1 Use comprehensive data sets 85
 - 7.2 Think about your negative control 85
 - 7.3 Ensure neutrality 85
- 8 What has not been covered in this chapter 86
- 9 Materials 87

Chapter 2.2

A biophysical approach to large-scale protein-DNA binding data

(*T. Manke, H. Roeder, M. Vingron*) 91

- 1 Binding site predictions 92
- 2 Affinity model {XE “affinity model, TRAP”} 95
- 3 Affinity statistics {XE “affinity statistics”} 99
- 4 Applications 101
- 5 Summary 102

Chapter 2.3

From gene expression profiling to gene regulation (*R. Coulson, T. Manke, K. Palin, H. Roeder, O. Sand, J. van Helden, E. Ukkonen, M. Vingron, A. Brazma*) 105

- 1 Introduction 105
- 2 Generating sets of co-expressed genes 106
- 3 Finding putative regulatory regions using comparative genomics 109
- 4 Detecting common transcription factors for co-expressed gene sets 111
- 5 Combining transcription factor information 114
- 6 “*De novo*” prediction of transcription factor binding motifs 115

SECTION 3: Annotation and genetics

Chapter 3

Annotation, genetics and transcriptomics (*R. Mott*) 123

- 1 Introduction 123
- 2 Genetics and gene function 125
 - 2.1 Genetic association studies in humans 125
- 3 Use of animal models 128
- 4 Transcriptomics: gene expression microarrays 130
- 5 Gene annotation 132

SECTION 4: Functional annotation of proteins

Chapter 4.1

Resources for functional annotation (*A. J. Bridge, A. Lise Veuthey, N. J. Mulder*) 139

- 1 Introduction 139
- 2 Resources for functional annotation – protein sequence databases 140

3	UniProt – The Universal Protein Resource	141
4	The UniProt Knowledgebase (UniProtKB)	142
4.1	UniProtKB/Swiss-Prot	142
4.1.1	Sequence curation in UniProtKB/Swiss-Prot	145
4.1.2	Computational sequence annotation in UniProtKB/Swiss-Prot	147
4.1.3	Functional annotation in UniProtKB/Swiss-Prot	147
4.1.4	Annotation of protein structure in UniProtKB/Swiss-Prot	149
4.1.5	Annotation of post-translational modifications in UniProtKB/Swiss-Prot	149
4.1.6	Annotation of protein interactions and pathways in UniProtKB/Swiss-Prot	150
4.1.7	Annotation of human sequence variants and diseases in UniProtKB/Swiss-Prot	150
4.2	UniProtKB/TrEMBL	151
5	Protein family classification for functional annotation	152
5.1	Protein signature methods and databases	152
5.1.1	Regular expressions and PROSITE	152
5.1.2	Profiles and the PRINTS database	153
5.1.3	Hidden Markov Models (HMM) and HMM databases	153
5.1.4	Structure-based protein signature databases	154
5.1.5	ProDom sequence clustering method	154
5.2	InterPro – integration of protein signature databases	154
5.3	Using InterProScan for sequence classification and functional annotation	155
5.3.1	InterProScan	155
5.3.2	Interpreting InterProScan results	156
5.3.3	Large-scale automatic annotation	159
6	From genes and proteins to genomes and proteomes	160
7	Summary	161

Chapter 4.2

Annotating bacterial genomes (*C. Médigue, A. Danchin*) 165

1	Background	165
2	Global sequence properties	170
3	Identifying genomic objects	172
4	Functional annotation	174
5	A recursive view of genome annotation	176

- 6 Improving annotation: parallel analysis and comparison of multiple bacterial genomes 178
- 7 Perspectives: new developments for the construction of genome databases, metagenome analyses and user-friendly platforms 180
- 8 Annex: databases and platforms for annotating bacterial genomes 182

Chapter 4.3

Data mining in genome annotation (*I. Artamonova, S. Kramer, D. Frishman*) 191

- 1 Introduction 191
- 2 An overview of large biological databases 193
 - 2.1 Manually curated vs. automatic databases 193
 - 2.2 Manually curated databases: the Swiss-Prot example 196
 - 2.3 Automatically generated databases: the PEDANT example 198
- 3 Data mining in genome annotation 200
 - 3.1 General remarks 200
 - 3.2 Supervised learning 201
 - 3.3 Unsupervised learning 201
 - 3.4 Clustering 202
 - 3.5 Association rule mining 203
- 4 Applying association rule mining to the Swiss-Prot database 205
- 5 Applying association rule mining to the PEDANT database 207
- 6 Conclusion 210

Chapter 4.4

Modern genome annotation: the BioSapiens network (*C. Yeats, C. Orengo, A. Lise Veuthey, B. Boeckmann, L. Juhl Jensen, A. Valencia, A. Rausell, P. Bork*) 213

- 1 Homologous and non-homologous sequence methods for assigning protein functions 213
 - 1.1 Introduction 213
 - 1.2 Homologs, orthologs, paralogs 216
 - 1.3 The HAMAP resource for the annotation of prokaryotic protein sequences and their orthologues 219
 - 1.4 CATH, Gene3D & GeMMA 222
 - 1.5 From SMART to STRING and STITCH: diverse tools for deducing function from sequence 228

- 1.6 General approaches for inheriting functions between homologous proteins 230
- 1.7 Non-homologous methods for predicting protein function from sequence 234

Chapter 4.5

Structure to function (*J. D. Watson, J. M. Thornton, M. L. Tress, G. Lopez, A. Valencia, O. Redfern, C. A. Orengo, I. Sommer, F. S. Domingues*) 239

- 1 Introduction to protein structure and function 239
- 2 FireDB and firestar – the prediction of functionally important residues 241
 - 2.1 Introduction 241
 - 2.2 FireDB 242
 - 2.3 Firestar 244
- 3 Modelling local function conservation in sequence and structure space for predicting molecular function 246
 - 3.1 Introduction 246
 - 3.2 Method 246
 - 3.3 Application 247
- 4 Structural templates for functional characterization 249
 - 4.1 Introduction 249
 - 4.2 Predicting protein function using structural templates 249
 - 4.3 FLORA method 250
- 5 An integrated pipeline for functional prediction 252
 - 5.1 Introduction 252
 - 5.2 The ProFunc server 253
 - 5.2.1 Sequence-based searches 254
 - 5.2.2 Structure-based searches 255
 - 5.3 Case studies 257
 - 5.3.1 Case study 1: published function identified 257
 - 5.3.2 Case study 2: function unclear 259
 - 5.4 Conclusion 259

Chapter 4.6

Harvesting the information from a family of proteins

(*B. Vroiling, G. Vriend*) 263

- 1 Introduction 263
 - 1.1 Information transfer 264
- 2 Molecular class-specific information systems 265
 - 2.1 G-protein-coupled receptors 266

- 3 Extracting information from sequences 267
 - 3.1 Correlated mutation analysis 268
- 4 Correlation studies on GPCRs 269
 - 4.1 Evolutionary trace method 271
 - 4.2 Entropy-variability analysis 273
 - 4.3 Sequence harmony 274
- 5 Discussion 274

SECTION 5: Protein structure prediction

Chapter 5.1

Structure prediction of globular proteins (*A. Tramontano, D. Jones, L. Rychlewski, R. Casadio, P. Martelli, D. Raimondo and A. Giorgetti*) 283

- 1 The folding problem 283
- 2 The evolution of protein structures and its implications for protein structure prediction 286
- 3 Template based modelling 287
 - 3.1 Homology-based selection of the template 288
 - 3.2 Fold recognition 288
 - 3.3 Using sequence based tools for selecting the template 289
 - 3.4 Completing and refining the model 291
 - 3.5 Current state of the art in template based methods 292
- 4 Template-free protein structure prediction 293
 - 4.1 Energy functions for protein structure prediction 296
 - 4.2 Lattice methods 297
 - 4.3 Fragment assembly methods 298
 - 4.4 Practical considerations 299
- 5 Automated structure prediction 300
 - 5.1 Practical lessons from benchmarking experiments 302
- 6 Conclusions and future outlook 304

Chapter 5.2

The state of the art of membrane protein structure prediction: from sequence to 3D structure (*R. Casadio, P. Fariselli, P. L. Martelli, A. Pierleoni, I. Rossi, G. von Heijne*) 309

- 1 Why membrane proteins? 309
- 2 Many functions 311
- 3 Bioinformatics and membrane proteins: is it feasible to predict the 3D structure of a membrane protein? 311

-
- 4 Predicting the topology of membrane proteins 312
 - 5 How many methods to predict membrane protein topology? 314
 - 5.1 From theory to practice 314
 - 6 Benchmarking the predictors of transmembrane topology 316
 - 6.1 Testing on membrane proteins of known structure and topology 316
 - 6.2 Topological experimental data 317
 - 6.3 Validation towards experimental data 318
 - 7 How many membrane proteins in the Human genome? 319
 - 8 Membrane proteins and genetic diseases: PhD-SNP at work 320
 - 9 Last but not least: 3D MODELLING of membrane proteins 322
 - 10 What can currently be done in practice? 323
 - 11 Can we improve? 324

SECTION 6: Protein–protein complexes, pathways and networks

Chapter 6.1

Computational analysis of metabolic networks (*P.-Y. Bourguignon, J. van Helden, C. Ouzounis, V. Schächter*) 329

- 1 Introduction 329
- 2 Computational resources on metabolism 331
 - 2.1 Databases 331
 - 2.1.1 KEGG 331
 - 2.1.2 BioCyc 332
 - 2.1.3 Reactome 332
 - 2.1.4 Querying and exporting data 333
 - 2.2 Reconstruction of metabolic networks 333
 - 2.2.1 From annotated genomes to metabolic networks 334
 - 2.2.2 Filling gaps 334
- 3 Basic notions of graph theory 335
 - 3.1 Metabolic networks as bipartite graphs 335
 - 3.2 Node degree 336
 - 3.3 Paths and distances 336
- 4 Topological analysis of metabolic networks 336
 - 4.1 Node degree distribution 337
 - 4.1.1 Robustness to random deletions and targeted attacks 339
 - 4.1.2 Generative models for power-law networks 340
 - 4.2 Paths and distances in metabolic networks 341
- 5 Assessing reconstructed metabolic networks against physiological data 342

5.1	Constraints-based models of metabolism	343
5.1.1	The flux balance hypothesis	343
5.1.2	Modelling the growth medium	344
5.1.3	Biomass function	345
5.2	Predicting metabolic capabilities	345
5.2.1	Predicting growth on a defined medium	345
5.2.2	Predicting gene essentiality	346
5.3	Assessing and correcting models using experimental data	347
5.4	Structural properties of the flux cone	347
5.5	Working with constraints-based models	348
6	Conclusion	348

Chapter 6.2

Protein–protein interactions: analysis and prediction (*D. Frishman, M. Albrecht, H. Blankenburg, P. Bork, E. D. Harrington, H. Hermjakob, L. Juhl Jensen, D. A. Juan, T. Lengauer, P. Pagel, V. Schächter, A. Valencia*) 353

1	Introduction	353
2	Experimental methods	354
3	Protein interaction databases	356
4	Data standards for molecular interactions	356
5	The IntAct molecular interaction database	360
6	Interaction networks	362
7	Visualization software for molecular networks	365
8	Estimates of the number of protein interactions	371
9	Multi-protein complexes	372
10	Network modules	373
11	Diseases and protein interaction networks	376
12	Sequence-based prediction of protein interactions	380
12.1	Phylogenetic profiling	381
12.2	Similarity of phylogenetic trees	383
12.3	Gene neighbourhood conservation	384
12.4	Gene fusion	385
13	Integration of experimentally determined and predicted interactions	385
14	Domain–domain interactions	389
15	Biomolecular docking	395
15.1	Protein–ligand docking	396
15.2	Protein–protein docking	398

SECTION 7: Infrastructure for distributed protein annotation**Chapter 7****Infrastructure for distributed protein annotation** (*G. A. Reeves, A. Prlic, R. C. Jimenez, E. Kulesha, H. Hermjakob*) 413

- 1 Introduction 413
- 2 The Distributed Annotation System (DAS) 415
- 3 DAS infrastructure 415
 - 3.1 DASTY2 – a protein sequence-oriented DAS client 418
 - 3.2 SPICE – a protein structure-oriented DAS client 418
 - 3.3 Ensembl 420
 - 3.4 DAS servers 422
- 4 The protein feature ontology 422
- 5 Conclusion 425

SECTION 8: Applications**Chapter 8.1****Viral bioinformatics** (*B. Adams, A. Carolyn McHardy, C. Lundegaard, T. Lengauer*) 429

- 1 Introduction 429
- 2 Viral evolution in the human population 430
 - 2.1 Biology and genetics 430
 - 2.2 Vaccine strain selection for endemic influenza 431
 - 2.3 Pandemic influenza 433
 - 2.4 Conclusion 434
- 3 Interaction between the virus and the human immune system 434
 - 3.1 Introduction to the human immune system 434
 - 3.2 Epitopes 436
 - 3.3 Prediction of epitopes 437
 - 3.4 Epitope prediction in viral pathogens in a vaccine perspective 441
- 4 Viral evolution in the human host 442
 - 4.1 Introduction 442
 - 4.2 Replication cycle of HIV 443
 - 4.3 Targets for antiviral drug therapy 444
 - 4.4 Manual selection of antiretroviral combination drug therapies 444
 - 4.5 Data sets for learning viral resistance 445
 - 4.6 Computational procedures for predicting resistance 446

- 4.7 Clinical impact of bioinformatical resistance testing 449
- 4.8 Bioinformatical support for applying coreceptor inhibitors 450
- 5 Perspectives 450

Chapter 8.2

Alternative splicing in the ENCODE protein complement (*M. L. Tress, R. Casadio, A. Giorgetti, P. F. Hallin, A. S. Juncker, E. Kulberkyte, P. Martelli, D. Raimondo, G. A. Reeves, J. M. Thornton, A. Tramontano, K. Wang, J.-J. Wesselink, A. Valencia*) 453

- 1 Introduction 453
- 2 Prediction of variant location 455
- 3 Prediction of variant function – analysis of the role of alternative splicing in changing function by modulation of functional residues 458
 - 3.1 Functions associated with alternative splicing 458
 - 3.2 Functional adaptation through alternative splicing 458
 - 3.2.1 Tafazzin 459
 - 3.2.2 Phosphoribosylglycinamide formyltransferase (GARS-AIRS-GART) 461
 - 3.3 Analysis across the ENCODE dataset 462
- 4 Prediction of variant structure 463
- 5 Summary of effects of alternative splicing 467
- 6 Prediction of principal isoforms 472
 - 6.1 A series of automatic methods for predicting the principal isoform 473
 - 6.1.1 Methods 474
 - 6.1.2 Evaluation of pipeline definitions 474
- 7 The ENCODE pipeline – an automated workflow for analysis of human splice isoforms 477
 - 7.1 Behind EPipe 477
 - 7.2 Example workflow: IFN alpha/beta receptor protein 479
 - 7.3 Future perspectives 480

INTRODUCTION

BIOSAPIENS: A European Network of Excellence to develop genome annotation resources

BioSapiens is a Network of Excellence, funded by the European Union's 6th Framework Programme, and made up of bioinformatics researchers from 26 institutions based in 15 countries throughout Europe. The objective of this network is to stimulate the development of bioinformatics resources to provide automated, validated and distributed annotation of genome data, with particular emphasis on the human genome.

The genome projects have revealed for the first time the 'blue-print' of life. The first genome of a free-standing organism – the bacterium *Haemophilus influenzae* – was sequenced in 1995, rapidly followed by another bacterium, *Escherichia coli* and baker's yeast. The first draft of the human sequence was published in 2001, and several other genomes of vertebrate species, including the mouse and rat, followed recently. In total there are now over 700 completed and 3000 draft genome sequences in the public domain and many more are planned. In addition there are many other complete and partial genome sequences in the private sector. This explosion in genomic information has been achieved in a remarkably short period of time, especially when we consider that the three-dimensional structure of DNA was discovered only 50 years ago. It is now possible to sequence a whole bacterium in a few days and the flood of new sequence data with the emergence of the new generation of sequencing technologies will certainly continue for the next decade. However, DNA sequences must be interpreted in terms of the RNA and proteins that they encode and the promoter and regulatory regions that control transcription and translation. Genomic sequences also provide a convenient 'coordinate' reference frame onto which functional information can be mapped.

Annotation can be described as the process of 'defining the biological role of a molecule in all its complexity' and mapping this knowledge onto the relevant gene products encoded by genomes. This involves both experimental and computational approaches and, indeed, absolutely requires their integration. As such, in one sense, this effort will occupy the majority of biologists (experimental and theoretical) for most of this century. The mission of the BioSapiens Network is to provide the necessary expertise and European infrastructure to allow distributed annotation, from both

computational and experimental laboratories. These expert annotations are being made available for everyone over the web using a specific technology for the integration of distributed annotations, towards which the network has made a substantial contribution. BioSapiens has also made a concerted effort to form a new generation of biologists capable of using these tools with the development of a series of courses as part of the Permanent European School of Bioinformatics.

The structure of this book logically follows the work packages of the BioSapiens project. Genome annotation starts by defining the positions of genes along the sequence, and by identifying their coding regions, regulatory sequences and promoters (Chapters 1.1, 1.2, 2.1, 2.2, 2.3). Once the proteins (and RNAs) and their localisation have been defined, secondary annotation to provide identification of biochemical and biological function is needed (Chapter 4.1, 4.2, 4.6). Errors in genome annotation can be identified using data mining algorithms (Chapter 4.3).

Protein families provide a powerful route to improved protein annotation (Chapter 4.4). The three-dimensional structures of proteins provide detailed knowledge of residue locations and probable functional sites. Indeed during the course of BioSapiens the emergence of new structures solved by the structural genomics consortia have made very clear the demand for programs able to predict function based on structural characteristics (Chapter 4.5). Predicting the structure of both globular (Chapter 5.1) and membrane (Chapter 5.2) proteins still remains a largely unsolved problem. Predictions about the functions of gene products can be made through sequence analysis combined, where possible, with analysis of where and when a gene product is produced and through its interactions with other proteins. The identification of relevant protein–protein interactions provides further clues for functional characterisation (Chapter 6.2), as well as the knowledge of the pathways and networks in which they participate (Chapter 6.1). Tools for comparative genomics, to map interactions and networks from one organism to another, are of critical importance. In addition, for humans, sequence variation among individuals is particularly important, especially in the context of disease and inherited disorders (Chapter 3). On the more technical side, BioSapiens has concentrated on laying the methodological foundations for systematic annotation with the development of a complete infrastructure based on the DAS (Distributed Annotation System) technology and a number of interfaces and web servers to make the information accessible to the final users (Chapter 7).

In parallel BioSapiens have intensively worked on applying annotation methods to relevant biological problems. Chapter 8.1 discusses recent advances in viral bioinformatics and describes computational approaches to analyzing host-pathogen interactions. The final chapter (8.2) presents the analysis of the splice variants identified in the genes experimentally studied by the pilot phase of the ENCODE project.

We do not want to close this introduction without acknowledging the stimulating criticisms of the BioSapiens Scientific Advisory Committee, as well as the continuous

support of Dr. Fred Marcus, Scientific officer of the BioSapiens project. This book and the work it describes would have not been possible without the scientific guidance of Prof. Janet Thornton, the BioSapiens coordinator. Finally, we would like to thank Kerstin Nyberg, BioSapiens project manager, and Stephen Soehnlén from SpringerWienNewYork for help in creating this book.

Dmitrij Frishman and Alfonso Valencia
Editors

SECTION 1

Gene defintion

CHAPTER 1.1

State of the art in eukaryotic gene prediction

T. Alioto and R. Guigó

Center for Genomic Regulation, Barcelona, Spain

1 Introduction

Computational gene prediction is the cornerstone upon which a genome annotation is built, as gene prediction is usually the first step taken toward the annotation of a newly sequenced genome. This is largely due to the fact that computational identification of the entire repertoire of genes in a genome is vastly more economical than the experimental identification of each and every gene, or for that matter, even a single gene. Apart from the economic driving force behind the development of the gene prediction field, there also exists a fundamental scientific or intellectual driving force: in order to precisely delineate the gene structures within anonymous genomic sequences, we must be able to accurately model, and therefore understand, individually and collectively the mechanisms of transcription, splicing, mRNA maturation, nonsense-mediated decay, translation and even non-coding RNA regulatory circuits.

The interplay between prediction and experimentation should be seen as hypothesis driven, not data driven, biological research. Each gene prediction is a hypothesis waiting to be tested and the results of testing then inform our next set of hypotheses. It is really no different than the early days of gene-finding. Ever since genes were defined as the hereditary units that confer traits or phenotypic traits to organisms, their study has been essential to the study of biology. The discoveries that genes reside in deoxyribonucleic acid (DNA), are transcribed into ribonucleic acid (RNA) and (in many cases) then translated into polypeptides spurred the rapid development of molecular biology, which revolves around trying to understand the function of genes at the molecular level. Thus it has become requisite that their coding sequences and, by necessity their physical locations within the genome and intron-exon structures, be determined.

Methods for finding genes have evolved since the early days of genetics. In the pre-genomic age, genetic maps were constructed by analysis of phenotypic segregation (either natural traits or mutant phenotypes) in large pedigrees or through series of

Corresponding author: Tyler Alioto, Center for Genomic Regulation, calle Dr. Aiguader, 88, 08003 Barcelona, Spain (e-mail: tyler.alioto@crg.es)

genetic crosses. In the post-genomic age, the gene finding problem has largely turned into a computational one. The task can now be stated as follows: given a DNA sequence, perhaps a chromosome or entire genome, what are the precise boundaries and exonic structures of all of the genes?

In prokaryotes and some simple eukaryotes, the computational solution is a relatively simple task: to identify long open reading frames (ORFs) that, due to their length, are likely to code for proteins. The precise start codon can often be identified using simple rules such as choosing the ATG that maximizes the length of the ORF. The presence of other signals such as a Pribnow box (TATAAT consensus), the -35

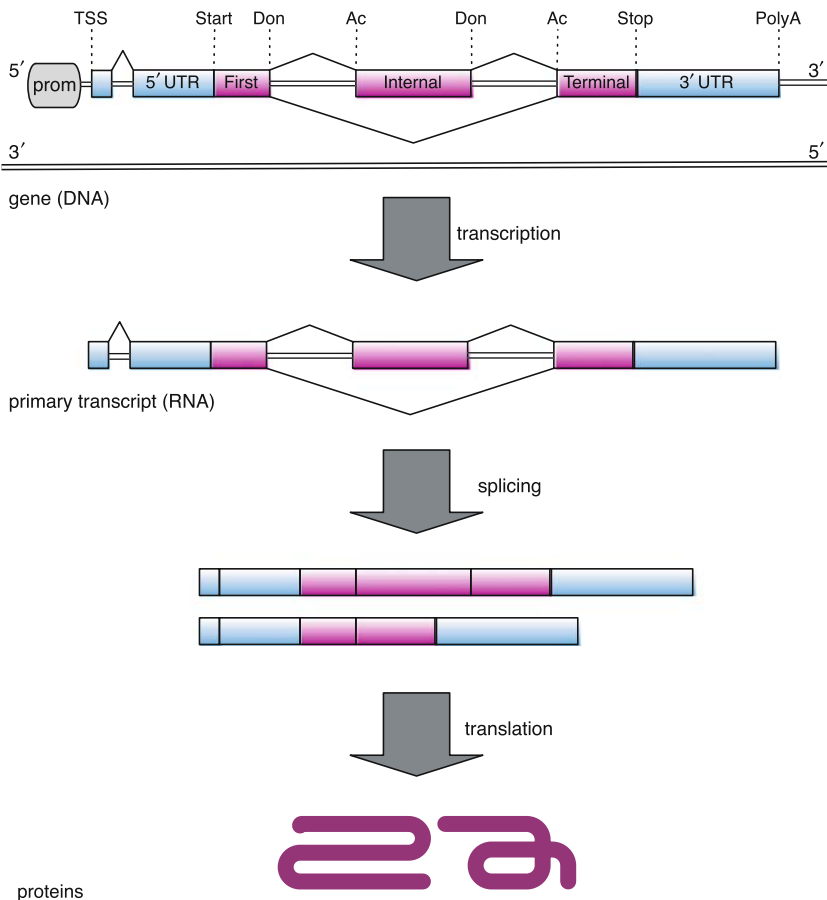


Fig. 1 Typical eukaryotic gene structure. Protein-coding genes are typically interrupted by non-coding sequences called introns, which are spliced out of the primary transcript (sometimes alternatively) to produce one or more mature messenger RNA products, which are then translated starting at the start codon and ending at the first in-frame stop codon

sequence or ribosomal binding sites can be used to refine the prediction of the transcriptional and translational start sites. Furthermore, codon bias is often used to deduce the correct frame for overlapping ORFs. The accuracy of prokaryotic gene finders is upwards of 90% for both sensitivity and specificity. GLIMMER (Salzberg et al. 1998) is perhaps one of the most accurate prokaryotic *ab initio* gene finders. It uses an interpolated Markov model (IMM), discussed later in this chapter. GeneMark (Borodovsky and McIninch 1993) is another successful prokaryotic (and now also eukaryotic) gene finder which pioneered the use of the 3-periodic Markov model for exon recognition that forms the basis of almost all modern gene predictors.

Eukaryotes, on the other hand, are more complex and pose a much greater challenge. First of all, their genomes can be orders of magnitude larger, and much of their DNA sequence does not code for proteins. For instance, only 3% of the human genome codes for proteins. Second, genes are almost always split into smaller coding sequences (exons) by intervening non-coding sequences (introns) which are spliced out of pre-messenger RNA by a ribonucleoprotein complex called the spliceosome to form a mature mRNA (see Fig. 1). Introns can sometimes be very large (>100 kb), making the search for exons like trying to find a needle in a haystack. Not to mention the fact that due to alternative splicing, multiple mature transcripts can be derived from one pre-mRNA. Alternative transcription start sites are also quite common. Genes can also be interleaved, overlapping, or nested, adding to the complexity.

Thus, for simplicity's sake, gene finding efforts to date have mainly focused on finding the genomic coordinates corresponding to a single protein-coding sequence per non-overlapping genomic locus. UTRs have largely been ignored as well as non-canonical splice sites (including U12 introns). That said, we must take note that this operational definition of a gene may have to be modified as our understanding of the transcriptional activity of the genome increases. A large proportion of the transcriptional activity in eukaryotic genomes, according to the results of new experimental techniques, appears not to code for proteins. These transcripts of unknown function, polyadenylated and non-polyadenylated, sense and antisense, overlapping and interleaved with protein coding genes, are distorting what once seemed to be a clear concept of a "gene" (Gingeras 2007).

For clarity, we will assume the operational definition but, where possible, highlight cases in which some of the complexities of transcription, RNA processing and translation are starting to be addressed. Even with these simplifying assumptions, gene finding programs exhibit far from perfect performance, thus we will refer to computational gene finding as "gene prediction" reflecting the still-necessary step of validating the gene models predicted by these programs.

In the next section we will introduce the basic principles of gene prediction, namely signal and content detection, and in the following section, we will illustrate how they are incorporated into modern eukaryotic gene finders. We will also discuss the development of more sophisticated frameworks for combining signal and content sensors with

diverse sources of information such as phylogenetic conservation and genomic alignments of expressed sequences.

2 Classes of information

We will begin by introducing the main sources of information that have been traditionally used to find genes. Then in Sect. 3 we will outline how this information can be captured and incorporated into gene model predictions. Information can be divided logically into two main categories, extrinsic and intrinsic, based on whether or not the information can be derived solely from the target genome sequence.

2.1 Extrinsic information

Extrinsic information includes any source of evidence that is not itself a genome sequence. In general, we refer to expressed sequence such as cDNAs, expressed sequence tags (ESTs) or the sequence of their protein products as extrinsic information. Gene prediction methods which do not use this information are referred to as *de novo* methods.

Homology information can be used in several ways according to quality and completeness. If the homologous sequence is derived from the same species and locus as the target sequence, then a spliced alignment approach often suffices to accurately map the region of homology. If the homologous sequence is full-length, such as a full-length cDNA sequence, and the boundaries of the transcript coincide with canonical splice sites, the coordinates of the genomic alignment represent the gold standard of gene annotation to which all other methods are compared. Determination of the start and stop codons then usually entails finding the longest open reading frame, although on occasion the true start codon is not the first methionine codon encountered. The presence of a Kozak consensus sequence ([A/G]XXAUGG) (Kozak 1981) can help distinguish true start codons from other potential start codons nearby.

Although BLAST (Altschul et al. 1990) is often used to roughly locate a gene within a genomic sequence using a homologous sequence, precise mapping of homologous sequences to the genome is ideally performed by programs specifically designed to perform spliced alignments. Procrustes (Gelfand et al. 1996), EST GENOME (Mott 1997), sim4 (Florea et al. 1998), BLAT (Kent 2002), GMAP (Wu and Watanabe 2005), and Exonerate (Slater and Birney 2005) are a few such examples. Genewise (Birney and Durbin 2000; Birney et al. 2004) is another program that aligns proteins to the genome. All such spliced aligners use either a basic model (terminal dinucleotide consensi) or more sophisticated models (such as position weight matrices/arrays) of splice junctions and introns.

If the region of homology is incomplete or of lower quality, then the preferred approach is to extend the spliced alignment with *ab initio* gene prediction. This approach is generally implemented as a stepwise pipeline such as in ENSEMBL (Hubbard et al. 2002) or UCSC genes (Hsu et al. 2006). However, EST and cDNA alignments may also be incorporated directly into gene predictions through extensions to gene predictors like Twinscan (Wei and Brent 2006), or, as is becoming more common, by “combiner” programs. At low levels of identity, BLAST high-scoring-pairs (HSPs) can either be used to weight predicted exons in a non-probabilistic way or may be incorporated into gene prediction probabilistically using pair hidden Markov models (see below).

2.2 Intrinsic information

De novo gene predictors are programs that predict the exon-intron structures of genes using the sequences of one or more genomes as their only input. The term *ab initio* is used strictly for *de novo* gene predictors that do not use informant genomes, and more or less means “from first principles”. The most “*ab initio*” of gene prediction programs would be a program that simulates the transcription, splicing and postprocessing of a transcript using only the information available to the cell. Such a simulator, if successful, would truly demonstrate our understanding of the molecular mechanisms and dynamics of gene expression. However, our understanding at this point is at best rudimentary and we must rely on metrics derived from many examples of genes with known exonic structures. These informative metrics can be categorized as either signal sensors or content sensors.

2.2.1 Signals

The signals in which we are interested are nucleic acid sequence motifs that are recognized by the cellular machinery responsible for transcribing, processing and translating messenger RNA molecules. The minimal set of signals that describes the structure of a coding sequence (CDS) include the start and stop codons and, if there is more than one exon in the coding part of the transcript, the donor and acceptor splice sites for each intron present. The acceptor site may be sometimes be defined as a composite of branch site, poly-pyrimidine tract and the acceptor junction signals. Additional signals that may be taken into consideration are splicing enhancer and silencer elements, transcription start and termination sites, polyadenylation signals, and even proximal and distal promoter sequences.

Many of these signals can be modeled as simple position weight matrices, or PWMs (alternatively known as position specific scoring matrices or position specific probability matrices). PWMs attempts to capture the intrinsic variability characteristic of sequence patterns and are usually derived from a set of aligned sequences which are

functionally related. PWMs simply tabulate the frequency with which each nucleotide is observed at each position. Formally, from a set S of n aligned sequences of length l , s_1, \dots, s_n , where $s_k = s_{k1}, \dots, s_{kl}$ (the s_{kj} being one of A, C, G, T in the case of DNA sequences) a Position Weight Matrix, $M_{4 \times l}$ is derived as

$$M_{ij} = \frac{1}{n} \sum_{k=1}^n I_i(S_{kj})$$

$$i \in [A, C, G, T]$$

$$j = 1 \cdots n$$

where $I_i(q) = \begin{cases} 1 & \text{if } i = q, \\ 0 & \text{otherwise.} \end{cases}$

This matrix is usually converted to a frequency or probability matrix with the sum of each column equal to one. A novel sequence can now be searched for this motif by moving a window the size of the motif across the sequence and for each position of the matrix summing the frequencies corresponding to each nucleotide observed. A score is obtained where the higher the score the better the match. However, scores from different matrices are difficult to compare and selecting a proper threshold becomes rather empirical. The solution to this problem is to use a “back-ground” model. Background frequencies could be equiprobable nucleotide frequencies with 0.25 for each A, C, G and T, or the frequencies may be derived from the true genome-wide nucleotide frequencies or perhaps from the local context of the true sites. The likelihood of a sequence belonging to the category of the motif becomes the product of the probabilities of the observed nucleotides occurring in each position of the motif divided by the product of the probabilities of the background nucleotides in each position of the motif. If we then take the log of this ratio, called the log-likelihood ratio, then sequences with scores above zero can be interpreted as being more likely to be an instance of the motif, while those that score below zero are not likely to be. If we store the log likelihood ratio for each position of the motif in the matrix, then we may simply take the sum of these ratios at each position to be the score of the entire motif. This method is illustrated in Fig. 2 using the U12 branch point PWM as an example (U12 introns, which comprise only a fraction of a percent of all human introns, are spliced by the minor U12 snRNP-containing spliceosome).

Dependencies between adjacent positions can be captured in a weight array matrix (WAM) model. The probabilities in the matrix are now calculated as conditional probabilities, where the probability of a sequence $S = s_1 \cdots s_n$ being an instance of a particular motif is

$$P(S) = P(s_1)P(s_2|s_1)P(s_3|s_2)P(s_4|s_3)P(s_5|s_4) \cdots P(s_n|s_{n-1})$$

where $P(s_i|s_j)$ is the probability of nucleotide s_j in position k given that nucleotides s_i is at position $k - 1$. Log-likelihood ratio scores can also be computed, by calculating the probability of the sequence S under some background model.

This type of dependency, where the state at one position is conditioned only on the state immediately preceding it (in space or time) fulfills the Markov assumption. Thus, these models can also be thought of as 0-order and 1st-order Markov Chains,

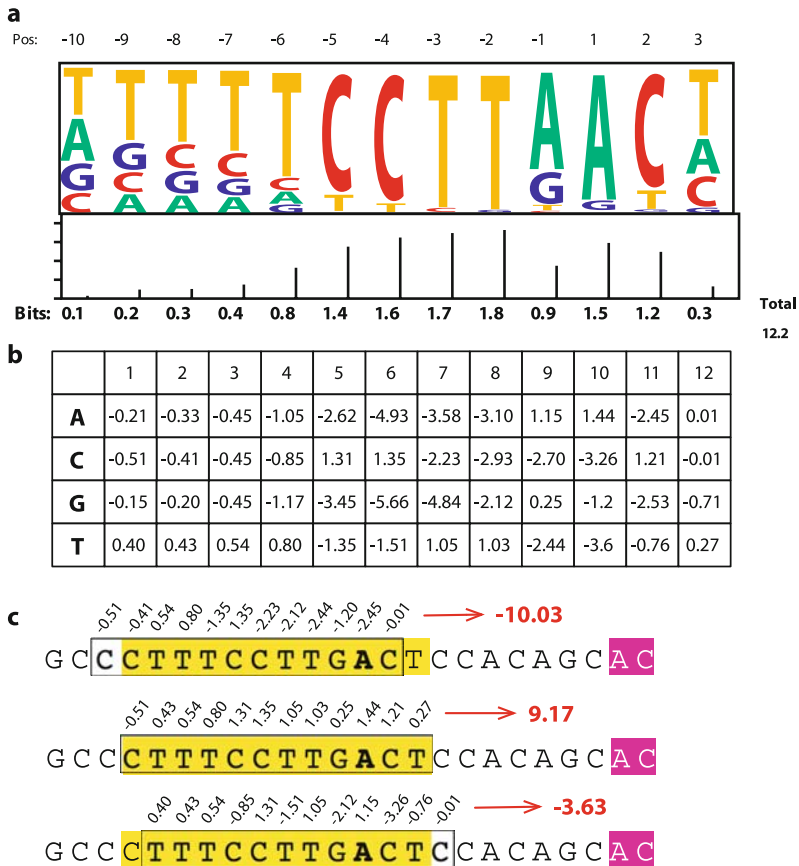


Fig. 2 Searching for signals. A position weight matrix (PWM) was calculated from known U12 branch point sequences. (a) The sequence logo shows the information content of the U12 branch point for human U12-dependent introns. (b) The PWM contains the log likelihood ratios (signal/background) for each base at each position of the 12 bp profile. (c) A 12 bp window is advanced one base pair at a time over the genomic sequence and the log ratios are summed over each position to give the branch point score. The result of scoring the positions immediately before, exactly over and immediately after the branch point are shown. The branch adenosine is shown in bold and the profile-matching bases are highlighted in yellow

respectively, where the order refers to the number of immediately preceding nucleotides on which the probability of observing a particular base is conditioned.

Donor splice sites, for example, are often modeled as 1st or 2nd order Markov chains. In fact, so are acceptor splice sites, branch points, polypyrimidine tracts, and start sites, among other signals.

Sometimes, however, non-adjacent positions exhibit dependencies, for example in the donor site motif. Several methods have been developed to capture these dependencies. Maximal dependence decomposition (MDD), which is used by Genscan (Burge and Karlin 1997), uses a decision tree to select one of several WAMs for scoring the site. Inclusion-driven learned Bayesian Networks (idlBNs) have also been tried (Castelo and Guigó 2004). These methods outperform PWMs and first-order Markov models when predicting individual sites, but the improvements tend to vanish when considered in the overall framework of a gene finding program. Support vector machines (SVMs) trained with sequence features local to the splice site have also shown promise (Sun et al. 2003; Zhang et al. 2003; Degroevé et al. 2005; Baten et al. 2006; Rättsch et al. 2006), however, it is unclear to what extent other features such as codon usage (usually detected separately from the splice site) influence their success. When used alone (not in a gene prediction context), they perform substantially better than the PWM or first-order Markov model (WAM).

2.2.2 Content

In theory, the signals on their own should completely specify the intron-exon structure of a transcript. However, proper classification of all potential start codons and splice sites in a genomic sequence is still a challenge. Properly detecting the start and end of transcription is also a major challenge. This suggests that either our models of these signals are inadequate, or we have yet to identify additional signals involved (such as cis-acting enhancer or silencer elements affecting splice site choice), or our models of the mechanisms of transcription and/or splicing are deficient or a combination of all of the above. Therefore, most gene prediction strategies also take advantage of the statistical properties of coding sequences. We call such content-based coding versus non-coding measures “coding statistics”.

Indeed, protein coding regions exhibit characteristic DNA sequence composition bias, which is absent from non-coding regions (see Fig. 3). The bias is a consequence of (1) the uneven usage of the amino acids in real proteins, and (2) of the uneven usage of synonymous codons. To discriminate protein coding from non-coding regions, a number of content measures can be computed to detect this bias (Fickett and Tung 1992; Gelfand 1995; Guigó et al. 2000). Such coding statistics can be defined as functions that compute a real number related to the likelihood that a given DNA sequence codes for a protein (or a fragment of a protein). Most coding statistics measure directly or indirectly either codon or di-codon usage bias, base compositional bias between codon positions, or periodicity

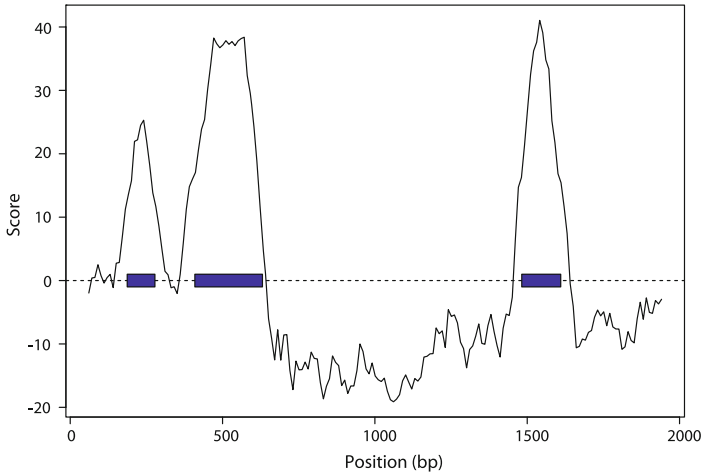


Fig. 3 Coding potential calculated using a fifth-order Markov model over the human beta globin gene locus. Annotated exons are shown in blue

in base occurrence (or a mixture of them all). Since the early eighties, a great number of coding statistics have been published in the literature. Hexamer frequencies usually in the form of codon position dependent 5th-order Markov models (Borodovsky and McIninch 1993) appear to offer the maximum discriminative power, and are at the core of most popular gene finders today. In practice it is implemented as a three-periodic inhomogeneous Markov model, with one Markov chain corresponding to each position of a codon. GRAIL (Uberbacher and Mural 1991; Xu et al. 1994), an earlier gene finding method, popular in the early nineties, used a neural networks to determine the optimal combination of a variety of coding statistics for predicting coding regions.

2.3 Conservation

When one or more informant genomes are available, it is possible to detect the characteristic conservation pattern of coding sequence and use it as an orthogonal measure of coding potential. Over the past few years, several programs have been developed that exploit sequence conservation between two genomes to predict genes. A wide variety of strategies have been explored. In one such strategy (Alexandersson et al. 2003) (further discussed below), alignment of the genomic sequence and gene prediction are performed simultaneously. In the “informant genome” approach (e.g. SGP2 (Parra et al. 2003) and TWINSKAN (Korf et al. 2001) alignments are performed first using standard tools such as TBLASTX or BLASTN and these alignments are used to inform prediction. More recently methods that use multiple alignments among several genomes have been developed.

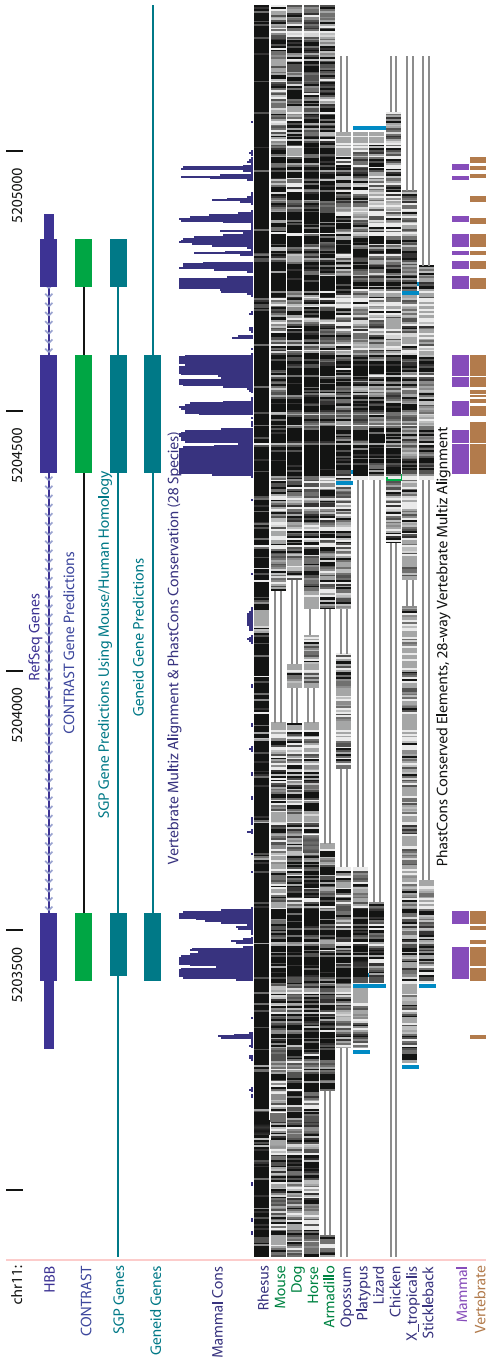


Fig. 4 Coding sequences are more conserved than non-coding sequences. Conservation within mammals at the human beta globin gene locus is shown. Gene prediction programs that utilize conservation (CONTRAST and SGP2) perform better than those that do not (GeneID)

To illustrate this point, in Fig. 4 we display the human beta globin gene locus on the UCSC genome browser. The definitive annotation is represented by the aligned RefSeq sequence at the top, while the conservation track at the bottom shows the evolutionary conservation as determined by a phylo-HMM. In between are various gene predictions which use 0 (GeneID), 1 (SGP2) or 27 (CONTRAST) aligned genomes.

3 Frameworks for integration of information

As we have seen, genomic and extra-genomic information of many different forms (sequence motifs, coding nucleotide composition, evolutionary conservation) can contribute to the prediction of the intron-exon structure of protein-coding transcripts. Successful gene prediction, however, depends on more than the sum of its parts; accurate and efficient integration of this information is critical. In this section we will look at gene prediction from the perspective of integration, outlining the various frameworks that have been developed and elaborated over the years.

3.1 Exon-chaining

Once exons are predicted, explicitly or implicitly, along a genomic sequence, exons need to be chained into gene predictions. Exon-chaining, therefore, is actually something that every gene predictor does, at least conceptually. The main difficulty in exon assembly is the combinatorial explosion problem: the number of ways N candidate exons may be combined grows exponentially with N . The key idea of computational feasibility comes from dynamic programming (DP), which allows finding the “optimal assembly” quickly without having to enumerate all possibilities (Gelfand and Roytberg 1993). Exon chaining DP (Guigó 1998) is implicit to several currently available gene predictors such as Fgenesh (Solovyev et al. 1995) and GeneID (Guigó et al. 1992; Parra et al. 2000). In GeneID, gene prediction is done hierarchically. First, splice sites, start and stop codons are predicted and scored on the query sequence. From these sites, all potential protein coding exons are built. The exons are scored as a function of the scores of the exon defining sites, and the score of a fifth-order Markov model which evaluates the coding bias of the predicted exon sequence. Because in GeneID all scores are log-likelihood ratios, the score of the exons is simply the sum of individual scores. Finally, exons are assembled into gene structures, so that the final assembly is the one maximizing the sum of the assembled exons.

The advantages of the hierarchical approach is that the gene finding problem can be tackled in discrete steps and analyzed at intermediate stages. It is also very fast and can analyze large mammalian genomes in only a few hours. It also allows for a quite flexible scoring approach, since exons can be re-scored, using *ad-hoc* procedures, depending on their conservation in other genome(s) or their similarity to known protein or cDNA sequences. However, a number of shortcomings are apparent, especially when com-

pared to the more recent crop of HMM and CRF-based gene predictors (see below): exon and intron length distributions are not very well modeled (only minimum and maximum lengths can be specified), and scores are not truly probabilistic.

3.2 Generative models: Hidden Markov models

A novel advance in eukaryotic gene prediction methodologies was the application of generalized Hidden Markov Models (HMMs), initially implemented in the Genie algorithm (Kulp et al. 1996). (HMMs were first used in a bacterium gene finder by Krogh et al. (1994) after its success in protein modeling.) Soon after, it was implemented in the Genscan algorithm (Burge and Karlin 1997) to predict multiple genes. Several other HMM-based gene prediction programs were developed later: Veil (Henderson et al. 1997), HMMgene (Krogh 1997) and Fgenesh (Salamov and Solovyev 2000).

In the HMM approach, different types of structure components (such as exons or introns) are characterized by a state, and the gene model is thought to be generated by a state machine: starting from 5' to 3', each base-pair is generated by an “emission probability” conditioned on the current state (and if using a higher order Markov model, a limited number of preceding bases), and the transition from one state to another is governed by a “transition probability” which obeys a number of constraints (e.g. an intron can only follow an exon, reading frames of two adjacent exons must be compatible, etc.). All of the parameters of the emission probabilities and the (Markov) transition probabilities are learned (pre-computed) from some training data. Since the states are unknown (“hidden”), an efficient algorithm (called the Viterbi algorithm, similar to DP) may be used to select the best set of consecutive states (called a “parse”), which has the highest overall probability of any possible parse for the given genomic sequence without actually having to enumerate all possible parses (see (Rabiner 1989) for a tutorial on HMMs).

The reason these fully probabilistic state models have become preferable is that all scores are probabilities themselves and the weighting problem becomes only a matter of counting relative observed state frequencies. It is easy to introduce more states (such as intergenic regions, promoters, UTRs, etc.) and transitions into HMM-based models to accommodate partial genes, intronless genes, even multiple genes or genes on different strands. These features are essential when annotating genomes or large contigs in an automated fashion.

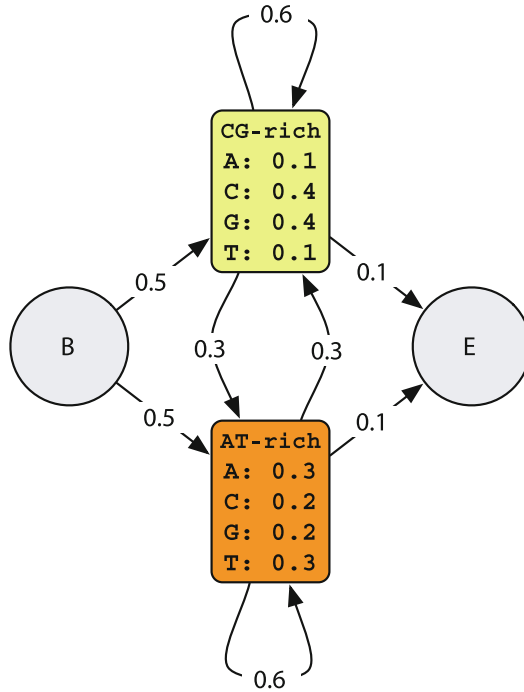
In the following sections, we will outline the various “flavors” of HMMs that have been applied to the problem of gene prediction, starting with the basic HMM.

3.2.1 Basic hidden Markov models

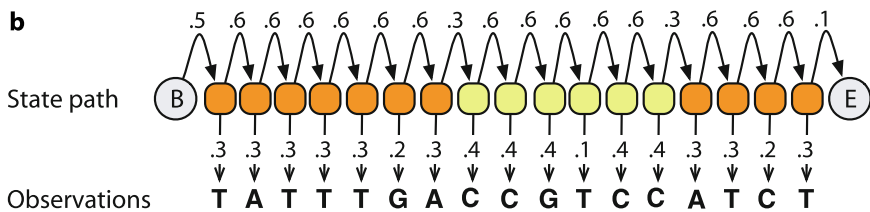
The first HMM-based gene predictors such as Genie were designed around a basic hidden Markov model, which is described by a set of possible states (e.g. start, exon,

donor, intron, acceptor, stop, intergenic, etc.), a set of possible observations (e.g. the set of nucleotides A, C, G and T), a transition probability matrix, an emission probability matrix, and the initial state probabilities. Transition probabilities govern the chance of

a



b



$$\Pr(\text{Sequence, State Path} \mid \text{Model}) = 2.8e-15$$

Fig. 5 A simple HMM for detecting regions of high GC content. (a) The state diagram exhibits two states which “emit” sequence according to different nucleotide probabilities. The begin (“B”) and end (“E”) states are silent, i.e. they do not emit sequence. Transition probabilities are shown with arrows. Emission probabilities are shown as tables in each of the two states. (b) The calculation of the joint probability $\Pr(x, y)$ of a sequence x and a particular state path or “parse” y is shown, and is simply the product of the transition and emission probabilities that were visited while traversing the path. The “true” sequence of states is “hidden”

moving from one state to any of the other states (or even back to the same state), for example from an exon to a donor site, from a donor site to an intron, etc. Emission probabilities correspond to the frequencies of nucleotides occurring in particular states (similar to a PWM model).

For an example of a simple hidden Markov model that illustrates the concept of states, transition probabilities and emission probabilities, please refer to Fig. 5, in which we show how one might design an HMM for detecting regions of high GC content. With this model, one can solve the following problems associated with an HMM:

1. *Evaluation.* Find the probability of the sequence given the model and its parameters. This would be the sum of all possible state paths through the sequence. The probability of one such path is shown in Fig. 5b. To enumerate all possible paths and sum their probabilities is generally an intractable problem, however fortunately there exists a dynamic programming algorithm, the “forward” algorithm, that can solve it efficiently.
2. *Decoding.* Find the most likely state path (i.e. sequence of AT-rich and GC-rich regions) given the model and a particular sequence. This is solved by the Viterbi algorithm.
3. *Learning.* Adjust the parameters (initial, transition and emission probabilities) to maximize the likelihood of the sequence given the model. In the example in Fig. 5, this would correspond to learning the probabilities of emitting the nucleotides A, C, G and T in each of the two states, AT-rich and GC-rich, and learning the probabilities of switching between the two states given a set of training sequences. If, however, the training sequences are already annotated with AT-rich regions, the learning step can be bypassed and the transition and emission probabilities set to the frequencies and base composition corresponding to the annotation.

Hidden Markov models for gene prediction, on the other hand, are necessarily more complex than the example in Fig. 5 due to the larger number of states and possible transitions needed to model gene structures. The first step in gene finding using an HMM is to learn the parameters from either labeled data (i.e. known genes) or unlabeled data. If the annotation is trusted, the transition and emission probabilities can simply be set to the frequencies observed in the annotated genes. Likewise, the weight array matrices for the various signals and content sensor sub-models that we described above are simply set by obtaining count frequencies. This procedure is called maximum likelihood estimation. In some cases, however, the optimal states are unknown, for example the ancestral evolutionary states in a phylo-HMM (described below). In these cases, the probabilistic basis of HMMs allows the parameters to be systematically learned from the data by maximum likelihood using the Baum-Welch algorithm (Baum et al. 1970), which is a special case of the Expectation Maximization algorithm (Dempster et al. 1977).

Once the model is trained, the software can be run on genomic sequences. Given the DNA sequence and the HMM model, a dynamic programming algorithm called the Viterbi algorithm can be used to find the optimal parse (i.e. the most likely sequence of exons and introns), or in other words annotate the sequence.

For gene finding, the probability of a sequence given an HMM is rarely solved for explicitly, although once an optimal path (in this case, a sequence of exons and introns) is predicted, its probability can give tell us something about how well it fits the model. The “Forward” and “Backward” algorithms are used to make this calculation.

3.2.2 Generalized hidden Markov models

One problem with the basic HMM is that the duration of a state can only be modeled as a transition back to itself with transition probability p . This in effect limits the duration of state to a geometric length distribution $E[l_X] = 1/(1 - p)$.

In a generalized HMM (GHMM), length distributions can be explicitly modeled, for example with a Poisson point process, which is a counting process that represents the total number of occurrences of discrete events during a temporal/spatial interval. An additional variable d is introduced into the HMM. Upon entering a state, a duration is chosen according to a particular probability distribution and then d number of characters are emitted according to the emission probabilities. The transition to the next state is made according to the transition probabilities. The advantage of this is that exon lengths and intron lengths can be explicitly modeled according to their estimated length distributions obtained from training. The disadvantage is an increase in computational complexity, thus often compromises are made. The program Augustus (Stanke et al. 2006), for example, reduces this computational cost by explicitly modeling short introns and using a geometric distribution for longer introns.

Another advantage of GHMMs is that they are modular. The states, in fact, can be represented by any suitable model and can be trained separately from the main model. For example, in Genscan, one of the first programs to utilize a GHMM, the donor site is modeled using maximal dependence decomposition (MDD) while the acceptor site is modeled by a standard Markov chain. Such modularity facilitates the design of the overall gene model, allowing one to easily incorporate additional states. A basic state diagram for gene prediction is shown in Fig. 6. There are usually separate models for each intron phase and exon frame, thus enabling proper frame consistency.

3.2.3 Generalized pair HMMs

As described above in Sect. 2.3, the availability of multiple fully sequenced genomes heralded the advent of multi-genome *de novo* gene predictors. SGP2 directly uses BLAST scores to modify the log odds that a particular candidate exon is coding. Twinscan modified the Genscan model to use an extended alphabet (8 characters) corresponding to

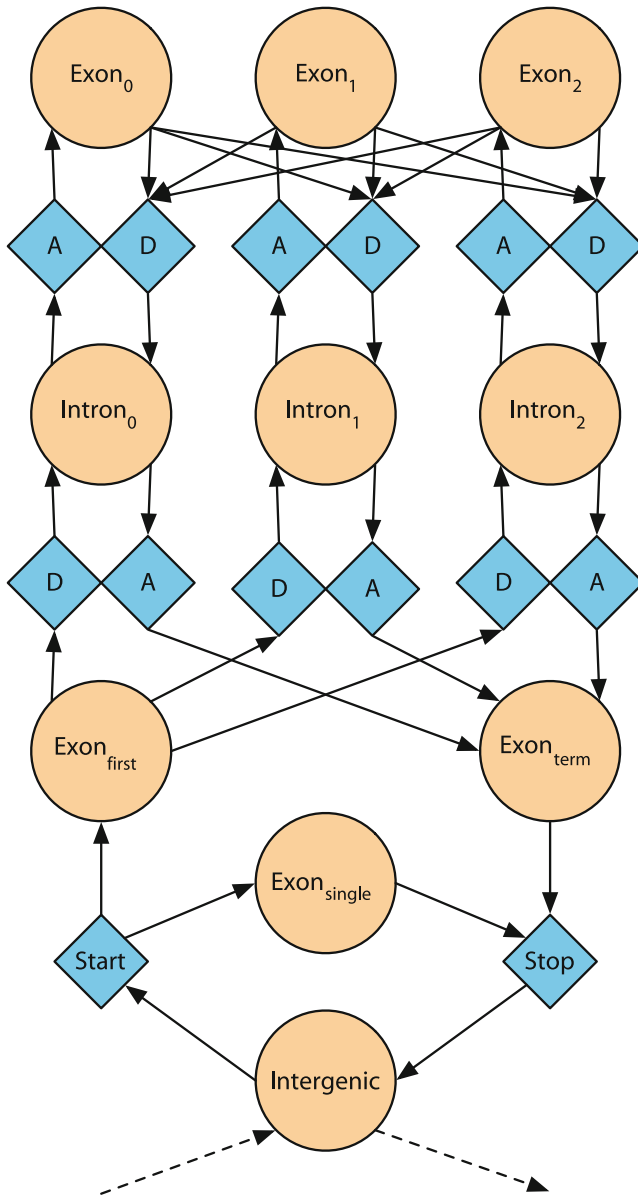


Fig. 6 A typical state diagram for a generalized hidden Markov model used for eukaryotic gene-finding. The three intron phases/exon frames are modeled by the separate intron and exon states 0, 1 and 2. Signal states donor (D), acceptor (A), start codon and stop codon (diamonds) mark the transitions between the variable-length content states introns, exons and intergenic regions (circles). Only the states for plus strand prediction are shown; simultaneous minus strand prediction are handled by a mirror image of the states linked through the intergenic state (not shown)

aligned and unaligned versions of the four bases, A, C, G and T. This represented a precursor to the next class of HMMs called generalized pair HMMs, pioneered by the program SLAM (Alexandersson et al. 2003) and also utilized by the program TWAIN (Majoros et al. 2005). Generalized pair HMMs (GPHMMs) represent a fully probabilistic comparative genomic approach that simultaneously produces both an alignment and annotation of two syntenic regions. Pair HMMs have traditionally been used in pairwise alignment algorithms and include match, insert and gap states. A generalized pair HMM is similar in that it emits gene features as aligned pairs (exon pairs or intron pairs, for example, one in each species). In addition to the set of parameters required by GHMMs, the GPHMM is additionally specified by a joint distribution of paired durations and a joint distribution of pair emission probabilities. A parse then becomes a series of states with paired durations. In general, exon insertion/deletions are not allowed, although Doublescan (Meyer and Durbin 2002), which uses a non-generalized pair HMM, does allow for indels.

The advantages of using GPHMMs are first of all, increased accuracy compared with methods that utilize only a single genome, and second you get two predictions for the price of one – gene predictions are made simultaneously in both genomic sequences. However variability in exon number is not tolerated, there are more parameters to estimate and the requirement for lengthy stretches of syntenic sequence is often difficult to meet, making their use in practice somewhat limited.

3.2.4 Phylo-HMMs or evolutionary HMMs

If a whole genome alignment of more than one genome is available, it is possible to integrate this information into a gene-finding HMM by explicitly modeling the evolutionary history of the DNA sequence. Phylo-HMMs (Siepel and Haussler 2004) (also called evolutionary HMMs (Pedersen and Hein 2003)) model a combination of two Markov processes operating in two different dimensions: space (along a genome, like in traditional GHMM gene finding) and time (along the branches of a phylogenetic tree). Basically, the columns of a multiple alignment are emitted according to a complex phylogenetic model such as the nucleotide substitution model of Hasegawa, Kishino and Yano (HKY) (Hasegawa et al. 1985), which is modeled using a continuous time Markov chain. The probability of mutation at a particular site is allowed to depend on the pattern of mutation at the previous few sites (obeying the Markov assumption) and the evolutionary rate in general differs according to biological function (coding versus non-coding, for example) and can also be allowed to vary from one region of the genome to another.

The UCSC conservation track is probably the best known example of a phylo-HMM. This model has also been successfully implemented in the gene prediction programs Shadower (McAuliffe et al. 2004) and N-SCAN (Gross and Brent 2006), a multi-genome version of Twinscan.

Phylo-HMMs represent a true advancement in the integration of multi-genome conservation and performance gains are seen over single- and dual-genome predictors. However, their use is restricted to cases where well-aligned genome sequences exist, and their computational cost is quite high.

3.3 Discriminative learning

Hidden Markov model based gene prediction has represented the state of the art of eukaryotic gene prediction for many years. More recently, however, we are beginning to see the application of new theoretical frameworks which may be best classified as discriminative in nature, as opposed to the generative nature of HMMs. In discriminative learning, the posterior probability $Pr(y|x)$ of hidden states (gene structure) given the observations (DNA sequence) is modeled directly. In generative learning (HMMs), a more general problem, estimation of the joint probability $Pr(x, y)$ of the states and observations from training data (as in Fig. 5b), is solved before calculating the posterior probability $Pr(y|x)$ according to Bayes rule (Ng and Jordan 2001), where x corresponds to the observations and y corresponds to the labels or state path.

The direct modelling of the probability of a gene annotation (a sequence of labeled segments, i.e. state path) given a sequence (the observations) lends itself to discriminative training, a training paradigm in which all parameters of the model are tuned or weighted in order to directly maximize the discriminatory power of the model. In the case of gene prediction, this means determining the weights of various model parameters in order to achieve maximum annotation accuracy according to defined measures of gene prediction accuracy (see Sect. 5). This type of training, in which the model is trained to maximize a conditional probability $Pr(x|y)$ versus a joint probability $Pr(x, y)$, is also called “conditional training”. Semi-Markov (or generalized) versions of support vector machines (SVMs) and conditional random fields (CRFs), both discriminative in nature, are promising newcomers to the field of gene prediction.

3.3.1 Support vector machines

Support vector machines (SVMs), a particular set of supervised learning methods, have rapidly become popular in biological research to solve classification problems. SVMs are designed to discriminate two classes, for example true splice sites from decoy sites, by separating them with a large margin. SVMs are trained by learning this margin, or boundary, from positively and negatively labeled training examples.

SVMs for gene prediction have been independently applied to the problems of splice site detection and exon content (coding versus non-coding) classification; however, more recently, the SVM framework has been generalized and applied to the

exon assembly problem, resulting in the programs mSplicer and mGene (Rätsch et al. 2007). Briefly the scores of the signal and content submodels (themselves learned by SVMs) are combined with segment length contributions and then given to piecewise linear weighting functions which have been trained to maximize the margin between the score of the best gene model and that of all false models.

3.3.2 Semi-Markov conditional random fields

Most recent on the scene of eukaryotic gene prediction are a set of programs based on semi-Markov conditional random fields (SM-CRFs). A SM-CRF on a sequence x outputs a segmentation of x in which labels are assigned to segments of the sequence (e.g. exon, intron, etc.) They are essentially “conditionally trained” semi-Markov chains, that is, they are designed to find the most likely set of labels (states) that the model has been trained to traverse given a set of observations (input sequence). SM-CRFs are analogous to GHMMs (or semi-HMMs) except that the probability of label-value pairs, the labels being conditioned on the values, is learned directly. The values or observations are examined and not “emitted” as they are in HMMs, which in many respects is more intuitive and more accurately reflects the problem that is trying to be solved. Some advantages of this framework are that (1) any discriminative feature corresponding to an arbitrary-length segment may be used, (2) it need not be probabilistic and (3) features may overlap – discriminative training will assign appropriate weights.

Recent examples of semi-Markov CRF implementations for gene prediction include:

- CRAIG (Bernal et al. 2007), which is trained globally on all input feature vectors using an online large-margin algorithm related to multiclass SVMs.
- CONRAD (DeCaprio et al. 2007), which is provided as a generic gene calling engine that promises to be highly customizable, although it has only been trained so far on fungal species.
- CONTRAST (Gross et al. 2007), a multi-genome predictor that is “phylogeny free” working directly with features extracting from whole genome multiple alignments.

The semi-Markov CRF framework would appear to hold much promise for the integration of multiple sources of information and may become the *de facto* model for such purpose.

3.4 Combiners

Programs that specifically aim to integrate the results of other gene callers have been dubbed “combiners”. Previous work has produced many such programs: GAZE (Howe et al. 2002), Jigsaw (Allen and Salzberg 2005), GLEAN (Elsik et al. 2007), Genomix

(Coghlan and Durbin 2007), and EuGène (Foissac and Schiex 2005) to name a few. The goal of such programs is to automate the task that faces human annotators: to produce an annotation when presented with the results of many different and potentially conflicting gene predictions.

While the combining functions differ among programs, the general principle on which they operate is that predictions should make uncorrelated errors which should tend to cancel each other out and increase the signal to noise ratio. This principle relies on the assumption that the input predictions are independent. However, this is often not the case due to the use of similar methods, training data or extrinsic evidence. This can be circumvented by careful choice of input methods or can be explicitly corrected for by the combining algorithm, as is done by the combiner GenePC, which we are developing.

In general, combiners perform better than any individual input, often dramatically improving on specificity measures at all levels. For this reason, they are becoming popular for the automated annotation of new genomes.

4 Training

In most gene prediction programs, there is a clear separation between the gene model itself and the parameters of the model. While the model is general, the parameters often need to be specifically estimated for different species, or taxonomic groups. Using the wrong parameters may lead to mispredictions. Typically, the parameters of the gene model define the characteristic of the sequence signals involved in gene specification (i.e. weight matrices for splice sites), the codon bias characteristic of coding exons (i.e. hexamer counts or Markov models for coding regions), and the relation between the exons when assembled into gene models (i.e. intron and exon lengths distributions, number of exons, etc.). These parameters are estimated from a set of annotated genomic sequences from the species of interest. If not enough annotated sequences are available, some programs, such as GLIMMER (Salzberg et al. 1998), allow for the use of Markov models of smaller order.

Depending on the framework, the exact training algorithms differ from program to program. However, almost all gene predictors end up being trained discriminatively as some point to fine tune the model parameters (both submodel and global parameters) in order to achieve maximum discrimination – and it seems that all programs are characterized by the presence of “fudge” factors that get manually tuned regardless of the training procedure used. For example, we have mentioned above that HMM-based predictors can be trained using the Baum-Welch EM algorithm; however, such maximum likelihood training is usually performed on each submodel separately and then the global model tuned afterwards, usually manually. It has been shown that further improvements are realized when formal discriminative training methods such

as generalized gradient ascent are used so as to maximize mutual information (MMI) on all the model parameters at once (Majoros and Salzberg 2004).

Because of all these and other reasons, training a gene prediction program for a new species or taxonomic group is not always a trivial exercise; it requires a lot of manual intervention, and very few applications, if any, offer automatic training protocols. Recently, however, methods have been developed to train gene-finding software even in the total absence of annotated genomic sequences of the organism under consideration (an increasingly common problem, when the sequencing of the genome of an organism is not followed by the sequencing of cDNAs from that organism) (Lomsadze et al. 2005; Korf 2004).

A limiting amount of training sequence available can also impinge on evaluation procedures (described in the next section). Of course it is desirable to train on as many known genes as possible to avoid overfitting; however, evaluation of the performance of a program should always be carried out on a clean set of genes on which the program's parameters were not estimated, in order not to bias the results. This is especially true when the model is trained to achieve maximum discriminative power. In this case one can perform an N-fold cross-validation or jackknife procedure, in which successive rounds of training and evaluation are performed with some of the data for training withheld and used for evaluation purposes. The results of all the rounds are then combined to give the final performance values.

5 Evaluation of gene prediction methods

5.1 The basic tools

Whether running gene prediction pipelines, or just running gene prediction programs on a locus of interest, it is important to compare the outputs of multiple runs of a predictor with different settings or to compare multiple predictions from different programs. The comparison should be able to tell you something about the quality of each prediction by graphically reflecting the confidence in each exon, and should be of sufficient resolution to compare alternative splice sites. Several solutions to this problem have emerged.

The program GFF2PS (Abril and Guigó 2000) is a highly customizable UNIX-based script for generating postscript figures from multiple prediction outputs or annotations in GFF format. GBROWSE is a database-driven application that performs a similar but web-based function. Perhaps the most easy to use online system, provided your genome is represented and you know the genomic coordinates of your annotations, is UCSC Genome Browser's custom track option. If you are an annotation group and provide annotation to the scientific community on a regular basis then the Distributed Annotation System (DAS) is the preferred approach. The most used DAS client for gene prediction annotations is ENSEMBL.

5.2 Systematic evaluation

In addition of having some clue on the accuracy of the predictions on particular cases, one would like to have an overall measure of the accuracy of the “*ab initio*” gene prediction programs. The accuracy of gene prediction programs is usually measured in controlled data sets. To evaluate the accuracy of a gene prediction program on a test sequence, the gene structure predicted by the program is compared with the actual gene structure of the sequence. The accuracy can be evaluated at different levels of resolution. Typically, these are the nucleotide, exon, and gene levels. These three levels offer complementary views of the accuracy of the program. At each level, there are two basic measures: Sensitivity (S_n) and Specificity (S_p), which essentially measure prediction errors of the first and second kind. Briefly, Sensitivity is the proportion of real elements (coding nucleotides, exons or genes) that have been correctly predicted, while Specificity is the proportion of predicted elements that are correct. More specifically, if TP are the total number of coding elements correctly predicted, TN , the number of correctly predicted non-coding elements, FP the number of non-coding elements predicted coding, and FN the number of coding elements predicted non-coding, then, in the gene finding literature, Sensitivity is defined as $S_n = TP/(TP + FN)$ and Specificity as $S_p = TP/(TP + FP)$. Both, Sensitivity and Specificity, take values from 0 to 1, with perfect prediction when both measures are equal to 1. Neither S_n nor S_p alone constitute good measures of global accuracy, since high sensitivity can be reached with little specificity and *vice versa*. It is desirable to use a single scalar value to summarize both of them. In the gene finding literature, the preferred such measure on the nucleotide level is the Correlation Coefficient defined as

$$CC = \frac{(TP \times TN) - (FN \times FP)}{(TP + FN) \times (TN \times FP) \times (TP + FT) \times (TN + FN)}$$

CC ranges from -1 to 1 , with 1 corresponding to a perfect prediction, and -1 to a prediction in which each coding nucleotide is predicted as non-coding and *vice versa*.

At the exon level, an exon is considered correctly predicted only if the predicted exon is identical to the true one, in particular both $5'$ and $3'$ exon boundaries have to be correct. A predicted exon is considered wrong (WE), if it has no overlap with any real exon, and a real exon is considered missed (ME) if it has no overlap with a predicted exon. A summary measure on the exon level is simply the average of sensitivity and specificity. At the gene level, a gene is correctly predicted if all of the coding exons are identified, every intron-exon boundary is correct, and all of the exons are included in the proper gene.

One of the first systematic evaluations of gene finders was produced by Burset and Guig (Burset and Guigó 1996). These authors evaluated seven programs in a

set of 570 vertebrate single gene genomic sequences. At that time, average exon prediction accuracy $((Sn + Sp)/2)$ ranged from 0.37 to 0.64. A few years later, Rogic et al. (2001) updated the analysis; the average exon accuracy of the tested programs increased to values between 0.43 to 0.76, illustrating the significant advances in computational gene finding that occurred during the nineties. (See Guigó and Wiehe (2003) for a review on the accuracy of gene prediction programs in the late nineties.)

The evaluations by Burset and Guigó, Rogic et al. and others suffered, however, from the same limitation: gene finders were tested in control led data sets made of short genomic sequences encoding a single gene with a simple gene structure. These data sets are not representative of the complete genome sequences being currently produced. To address this limitation, and in the context of large genome and annotation projects, more complex community evaluation experiments have been carried out to obtain a more realistic estimation of the actual accuracy of gene finding programs.

5.3 The community experiments

Community experiments – experiments on which many groups all over the world participate simultaneously – are becoming popular in Bioinformatics to comparatively benchmark the status of the prediction tools in a given area. One of the most well-known is CASP, which stands for Critical Assessment of Techniques for Protein Structure Prediction, and which takes place every two years since 1994. CASP provides the research community with an assessment of the state of the art in the field of protein structure prediction. Protein structures that are either expected to be solved shortly or that have been recently solved, but not yet discussed in public, are used as targets for the prediction. Predictions submitted by groups worldwide are then evaluated and compared.

5.3.1 GASP

GASP, the Genome Assessment Project, was inspired by CASP, and took place in 1999 in the context of the *Drosophila* Genome Project. In short, at GASP, a genomic region in *Drosophila melanogaster*, including auxiliary training data, was provided to the community and gene finding experts were invited to send the annotation files they had generated to the organizers before a fixed deadline. Then, a set of standards were developed to evaluate submissions against the later published annotations (Ashburner et al. 1999), which had been withheld until after the submission stage. Next, the evaluation results were assessed by an independent advisory team and publicly presented at a workshop at the Intelligent Systems in Molecular Biology (ISMB) 1999 meeting. This community experiment

was then published as a collection of methods and evaluation papers in *Genome Research* (Reese et al. 2000).

5.3.2 EGASP

Within the context of the pilot phase of the ENCODE project, the second GASP, the so-called ENCODE GASP (EGASP) took place. The 44 regions selected within the ENCODE project had been subjected to a detailed computational, experimental and manual inspection and a high quality gene annotation of the ENCODE regions had been produced – the so-called GENCODE annotation (Harrow et al. 2006). On January 15, 2005 the complete gene map for 13 of the 44 regions was released, and gene prediction groups worldwide were asked to submit predictions for the remaining 31 regions. Eighteen groups participated, submitting 30 prediction sets by the April 15. The annotation of the entire set of the ENCODE regions was then released, and on May 6 and 7, participants, organizers and a committee of external assessors met at the Sanger Institute to compare the GENCODE gene map with the gene maps predicted by the participating groups. As with GASP, results were published as a collection of papers in the journal *Genome Biology* (Guigó et al. 2006). Accuracy at the exon level for participating programs is shown in Fig. 7. At EGASP some programs reached average exon accuracies close to 0.85.

5.3.3 NGASP

Very recently, NGASP, the nematode genome annotation assessment project, has taken place. Since five *Caenorhabditis* nematode genomes are currently available, those of *C. remanei*, *C. japonica* and *C. brenneri*, *C. elegans* and *C. briggsae*, nGASP was launched with the implicit goal of promoting the usage of the comparative information across these five genomes. The explicit goal was to objectively assess the accuracy of the current state of the art for protein-encoding gene prediction algorithms in *C. elegans*, and to apply this knowledge to the annotation of the other *Caenorhabditis* genomes. A set of regions representing ~10% (10 Mb) of the *C. elegans* genome was selected to evaluate the performance of the participating gene predictors. As with previous genome annotation assessment projects, participation was open to all academic, private sector, and government researchers. A summary of the results will be submitted for publication.

These community experiments are an excellent exercise to focus a whole community on a certain problem task and motivate groups and individuals to participate and submit their best possible solutions. External assessment of the results is critical and standards and rules have to be laid out clearly at the beginning of the experiment. They have been received with enthusiasm within the gene prediction community and they have had a great impact in tool development.

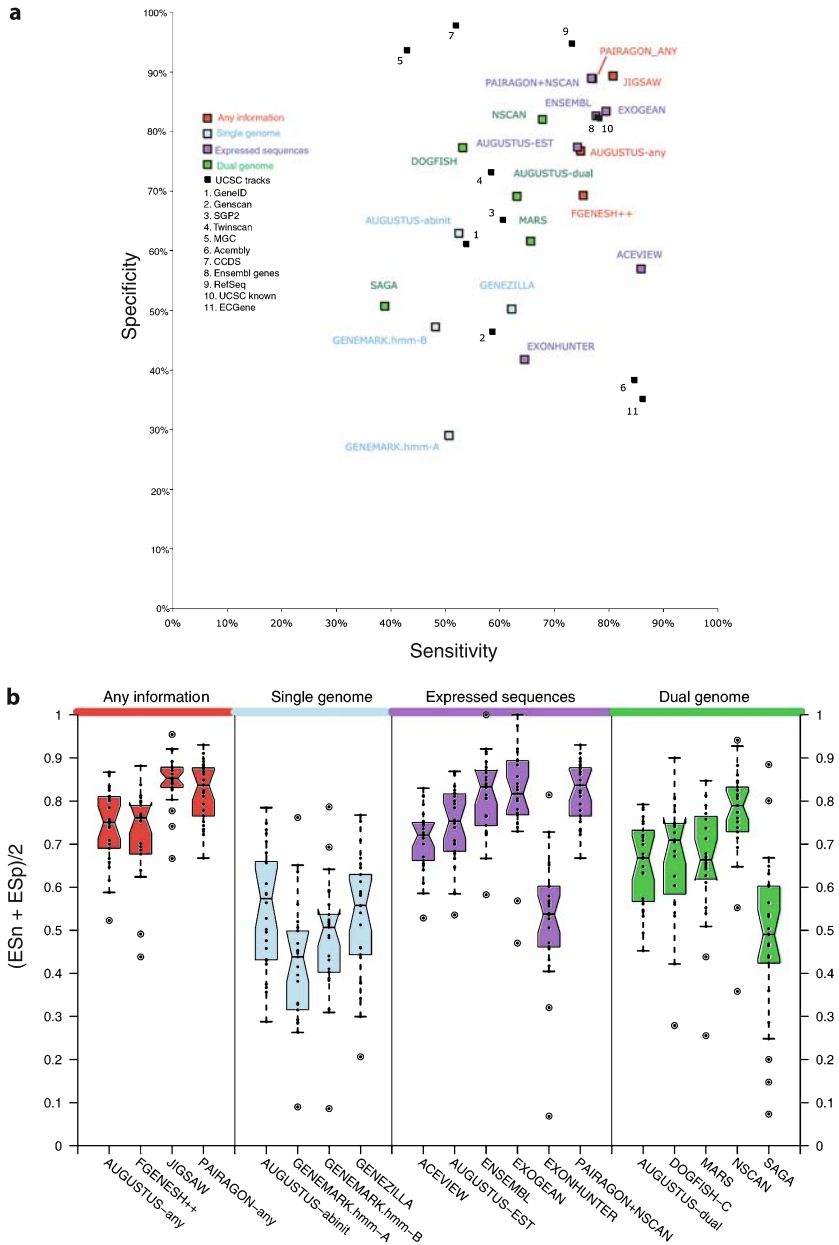


Fig. 7 Performance at the exon-level of various gene predictions submitted to the EGASP work-shop in 2005. **(a)** Sensitivity versus specificity on the 31 test regions for each program. **(b)** Boxplots of average sensitivity and specificity where each data point corresponds to the average in each of the test sequences for which a GENCODE annotation existed. Reproduced with permission from (Guigó et al. 2006) Fig. 6

6 Discussion

6.1 Genome datasets

There has been an effort to centralize all the information around the assembled sequences, and associated annotations, produced by the whole-genome sequencing projects. The best example are the three fully established whole-genome browsers: the NCBI Map Viewer (Wheeler et al. 2001), the UCSC Genome Browser (Karolchik et al. 2003) and the ENSEMBL browser at the Sanger Center (Hubbard et al. 2002), each of which present by default a set of contributed gene-finding predictions from different programs obtained for each new released assemblies. In addition, each site develops its own “in-house” gene set. These sets are based on mRNA evidence obtained from cDNA and EST sequences, augmented with computational predictions.

ENSEMBL human genes are generated automatically by the ENSEMBL gene builder. They are of three basic types: those having full-length cDNA or proteins, those having high homology to proteins in other organisms and those Genscan-predicted genes matching to proteins/vertebrate mRNA and UniGene clusters. The basic gene-annotator engine (using protein homology to construct gene structure) is Genewise (Birney and Durbin 2000). The “ENSEMBL genes” are regarded as being fairly conservative (with a low false positive rate), since they are all supported by experimental evidence of at least one form via sequence homology. Recently, ENSEMBL project has added spliced EST information for identification of alternative transcripts and to incorporate comparative genomics for getting orthologs and synteny relations. The basic annotator engine at the UCSC browser is BLAT (Kent 2002) which allows rapid alignment of primate DNAs/RNAs or land vertebrate proteins onto the human genome reliably, hence annotating the genome by similarities. Finally, NCBI LocusLink has a rule-based genome annotation pipeline. Known genes are identified by aligning RefSeq genes (<http://www.ncbi.nlm.nih.gov/RefSeq/>) and GenBank mRNAs to the genome using MegaBLAST (Zhang et al. 2000). Transcript models are reconstructed by attempting to settle disagreements between individual sequence alignments without using an a priori model (such as codon usage, initiation, or polyA signals). Genes (and corresponding transcript and protein features) are annotated on the contig if the defining transcript alignment is = 95% identity and the aligned region covers = 50% of the length, or at least 1000 bases. Finally, genes predicted by GenomeScan (Yeh et al. 2001), an extension of Genscan to include protein homology information, are annotated only if they do not overlap any model based on a mRNA alignment.

6.2 Atypical genes

Gene prediction efforts have been traditionally focused on predicting the “typical” gene. Genes with uncharacteristic features that do not appear with great frequency

tend to be ignored such as, for example, genes possessing U12-type introns, selenoprotein genes with in-frame UGA codons which code for selenocysteine, fast-evolving genes or genes with atypical codon usage. Progress has been made in a couple of these cases.

U12 introns, which comprise only a fraction of a percent of all introns, are spliced by the minor spliceosome, a low-abundance spliceosome with a different composition of snRNPs than the major U2-dependent spliceosome. It binds to donor and branch point sequences which are highly conserved across all species in which they are found, which includes most animals, plants and even a few fungi and protists. However, the splice sites do not conform to the regular U2 consensus – they are quite divergent and many of them have AT-AC terminal dinucleotides, making them invisible to most gene prediction software. By incorporating WAMs for the U12 splicing signals into the GeneID parameter file and making a few modifications to the dynamic programming routine, we have made the latest version of GeneID able to predict genes with U12 splice sites without a significant decrease in specificity. To aid in future genome annotation efforts, introns from a wide range of eukaryotic genomes that have been classified as U12-type are now stored in a specialized database called U12DB (Alioto 2007).

Selenoproteins pose an even greater challenge due to the presence of in-frame UGA codon(s) which are recognized by the selenocysteine tRNA in the presence of a SECIS element downstream, usually located in the 3' UTR. Yet these have also been systematically hunted down using a combination of *ab initio* gene prediction, RNA structure predictions and homology search (Kryukov et al. 2003). The selenoproteome is now catalogued in the SelenoDB (Castellano et al. 2008).

6.3 Outstanding challenges to gene annotation

Community assessment experiments have revealed that computational methods are not able to reproduce the accuracy in the annotation that a dedicated team of annotators, evaluating the individual evidence that exist for the transcripts mapping to a given genomic locus, can produce. For instance, EGASP revealed that the most accurate of the gene finding programs are able to predict correctly only about 40% of the full length transcripts in the GENCODE annotation. The GENCODE annotation heavily relies on human supervision (by the HAVANA team at the Sanger Institute (Harrow et al. 2006)) to solve the uncertainties arising from cDNA mapping onto the genome sequence, and it also includes computational predictions verified experimentally by RT-PCR and RACE. It is a much richer catalogue of the human transcriptome in the ENCODE regions than other existing gene sets. Indeed, the first release of the GENCODE annotation consisted of 2608 transcripts assigned to 487 loci, more than doubling the number of alternative transcripts per locus in ENSEMBL. It looks like, therefore, there is still room for improving gene finding software that can automatically reproduce the task being carried

out by human annotators when confronted with the complexity of transcription in the human genome.

This complexity, however, appears to be of a magnitude much higher than that implied by the GENCODE annotation. While extensive verification studies – including the EGASP community experiment (Guigó et al. 2006) – have demonstrated that the GENCODE is essentially complete with respect to existing cDNA sequences and computational predictions, recent research by a number of groups using a variety of technologies shows that many transcripts exist that are not annotated in GENCODE. Indeed, data from high-throughput tag sequencing of cDNA ends (Shiraki et al. 2003; Ng et al. 2005; Peters et al. 2007), from gene trapping in mouse embryonic stem cells (Roma et al. 2007) and from hybridization of RNA samples into high density tiling arrays (Kapranov et al. 2007; The ENCODE Consortium 2007) reveals many additional sites of transcription. Particularly relevant are the results of the so-called RACEarray experiments in which the products of RACE reactions originating from primers anchored in exons from GENCODE genes are hybridized onto genome tiling arrays. More than half of the sites of transcription detected in this way (the so-called RACEfrags), which are by construction specifically linked to annotated protein coding genes, do not correspond to GENCODE annotated exons (Denoëud et al. 2007). These results, therefore, are strongly indicative of the existence of a wealth of transcripts – including many alter-native transcript forms of protein coding genes, and other transcriptionally complex events – which had so far escaped detection through systematic sequencing of cDNA libraries. Computational gene prediction methods are generally based on computational models that capture our understanding of the way proteins are encoded in genomes. Modeling these other types of transcripts may be far more challenging than modeling the standard protein-coding ones, as they may lack the strong signatures characterizing the latter.

6.4 What is the right gene prediction strategy?

The answer to the question of which gene prediction program to use is “all”. As of yet, no one program is even close to perfect, so the best advice is to run a handful of the best and combine their results using a gene prediction combiner. And even then, the gene models produced should be regarded as hypotheses about the gene structures embedded within the chromosome. These models can and should be validated by RT-PCR and/or direct sequencing.

While the state of the art in eukaryotic gene finding has improved steadily over the last decade, there is still a long way to go before we can automatically produce high-quality gene models for an entire genome, even one as well studied as the human genome. Moreover, the plethora of eukaryotic genomes being sequenced now and in the future, and for which there is little transcriptional data, only increases the demand for better computational gene annotation methods.

References

- Abril JF, Guigó R (2000) gff2ps: visualizing genomic annotations. *Bioinformatics* (Oxford, England) 16: 743–744
- Alexandersson M, Cawley S, Pachter L (2003) Slam: Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res* 13: 496–502, doi: 10.1101/gr.424203
- Alioto T (2007) U12db: a database of orthologous u12-type spliceosomal introns. *Nucleic Acids Res* 35: 110–115, doi: 10.1093/nar/gkl796
- Allen J, Salzberg S (2005) Jigsaw: integration of multiple sources of evidence for gene prediction. *Bioinformatics* (Oxford, England) 21: 3596–3603, doi: 10.1093/bioinformatics/bti609
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410, doi: 10.1006/jmbi.1990.9999
- Ashburner M, Misra S, Roote J, Lewis SE, Blazej R, Davis T, Doyle C, Galle R, George R, Harris N, Hartzell G, Harvey D, Hong L, Houston K, Hoskins R, Johnson G, Martin C, Moshrefi A, Palazzolo M, Reese MG, Spradling A, Tsang G, Wan K, Whitelaw K, Celniker S (1999) An exploration of the sequence of a 2.9-mb region of the genome of *Drosophila melanogaster*: the *adh* region. *Genetics* 153: 179–219
- Baten AKMA, Chang BCH, Halgamuge SK, Li J (2006) Splice site identification using probabilistic parameters and svm classification. *BMC Bioinformatics* 7 Suppl 5: S15, doi: 10.1186/1471-2105-7-S5-S15
- Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat* 41(1): 164–171
- Bernal A, Crammer K, Hatzigeorgiou A, Pereira F (2007) Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput Biol* 3: e54, doi: 10.1371/journal.pcbi.0030054
- Birney E, Durbin R (2000) Using genewise in the *Drosophila* annotation experiment. *Genome Res* 10: 547–548
- Birney E, Clamp M, Durbin R (2004) Genewise and genomewise. *Genome Res* 14: 988–995, doi: 10.1101/gr.1865504
- Borodovsky M, McIninch J (1993) Genemark: parallel gene recognition for both dna strands. *Comput Chem* 17: 123–133
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic dna. *J Mol Biol* 268: 78–94, doi: 10.1006/jmbi.1997.0951
- Burset M, Guigó R (1996) Evaluation of gene structure prediction programs. *Genomics* 34: 353–367. 10.1006/geno.1996.0298
- Castellano S, Gladyshev VN, Guigó R, Berry MJ (2008) Selenodb 1.0: a database of selenoprotein genes, proteins and secis elements. *Nucleic Acids Res* 36: D332–D338, doi: 10.1093/nar/gkm731
- Castelo R, Guigó R (2004) Splice site identification by idlbn. *Bioinformatics* (Oxford, England) 20 Suppl 1: i69–i76, doi: 10.1093/bioinformatics/bth932
- Coghlan A, Durbin R (2007) Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron-exon structure. *Bioinformatics* (Oxford, England) 23: 1468–1475, doi: 10.1093/bioinformatics/btm133
- DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE (2007) Conrad: gene prediction using conditional random fields. *Genome Res* 17: 1389–6558107, doi: 10.1101/gr.6558107
- Degroeve S, Saeys Y, De Baets B, Rouzé P, Van de Peer Y (2005) Splicemachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* (Oxford, England) 21: 1332–1338, doi: 10.1093/bioinformatics/bti166
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the *em* algorithm. *J Roy Stat Soc B Met* 39(1): 1–38

- Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen C, Holroyd N, Dickson M, Taylor R, Hance Z, Foissac S, Myers R, Rogers J, Hubbard T, Harrow J, Guigo R, Gingeras T, Antonarakis S, Reymond A (2007) Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in encode regions. *Genome Res* 17: 746–759, doi: 10.1101/gr.5660607
- Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM (2007) Creating a honey bee consensus gene set. *Genome Biol* 8: R13, doi: 10.1186/gb-2007-8-1-r13
- Fickett JW, Tung CS (1992) Assessment of protein coding measures. *Nucleic Acids Res* 20: 6441–6450
- Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cdna sequence with a genomic dna sequence. *Genome Res* 8: 967–974
- Foissac S, Schiex T (2005) Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics* 6: 25, doi: 10.1186/1471-2105-6-25
- Gelfand MS (1995) Prediction of function in dna sequence analysis. *J Comput Biol: A J Comput Mole Cell Biol* 2: 87–115
- Gelfand MS, Roytberg MA (1993) Prediction of the exon-intron structure by a dynamic programming approach. *Bio Systems* 30: 173–182
- Gelfand MS, Mironov AA, Pevzner PA (1996) Gene recognition via spliced sequence alignment. *P Natl Acad Sci USA* 93: 9061–9066
- Gingeras T (2007) Origin of phenotypes: genes and transcripts. *Genome Res* 17: 682–690, doi: 10.1101/gr.6525007
- Gross S, Do C, Sirota M, Batzoglou S (2007) Contrast: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol* 8: R269, doi: 10.1186/gb-2007-8-12-r269
- Gross SS, Brent MR (2006) Using multiple alignments to improve gene prediction. *J Comput Biol: A J Comput Mol Cell Biol* 13: 379–393, doi: 10.1089/cmb.2006.13.379
- Guigó R (1998) Assembling genes from predicted exons in linear time with dynamic programming. *J Comput Biol: A J Comput Mol Cell Biol* 5: 681–702
- Guigó R, Wiehe T (2003) *Gene prediction accuracy in large DNA sequences*. Caister Academic Press, Norfolk
- Guigó R, Knudsen S, Drake N, Smith T (1992) Prediction of gene structure. *J Mol Biol* 226: 141–157
- Guigó R, Agarwal P, Abril JF, Burset M, Fickett JW (2000) An assessment of gene prediction accuracy in large dna sequences. *Genome Res* 10: 1631–1642
- Guigó R, Flicek P, Abril J, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic V, Birney E, Castelo R, Eyraas E, Ucla C, Gingeras T, Harrow J, Hubbard T, Lewis S, Reese M (2006) Egasp: the human encode genome annotation assessment project. *Genome Biol* 7 Suppl 1: 2–1, doi: 10.1186/gb-2006-7-s1-s2
- Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, Lagarde J, Gilbert J, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis S, Guigo R (2006) Gencode: producing a reference annotation for encode. *Genome Biol* 7 Suppl 1: 4–1, doi: 10.1186/gb-2006-7-s1-s4
- Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Molecular Evolution* 22: 160–174
- Henderson J, Salzberg S, Fasman KH (1997) Finding genes in dna with a hidden Markov model. *J Comput Biol: A J Comput Mole Cell Biol* 4: 127–141
- Howe K, Chothia T, Durbin R (2002) Gaze: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res* 12: 1418–1427, doi: 10.1101/gr.149502
- Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D (2006) The ucsc known genes. *Bioinformatics (Oxford, England)* 22: 1036–1046, doi: 10.1093/bioinformatics/btl048
- Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyraas E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp

- C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M (2002) The ensembl genome database project. *Nucleic Acids Res* 30: 38–41
- Kapranov P, Cheng J, Dike S, Nix DA, Dutttagupta R, Willingham AT, Stadler PF, Hertel J, Hacker-mueller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR (2007) Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science (New York, N.Y.)*, 316: 1138341–1488, doi: 10.1126/science.1138341
- Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ (2003) The ucsc genome browser database. *Nucleic Acids Res* 31: 51–54
- Kent WJ (2002) Blat—the blast-like alignment tool. *Genome Res* 12: 656–2292R, doi: 10.1101/gr.229202. Article published online before March 2002
- Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5: 59, doi: 10.1186/1471-2105-5-59
- Korf I, Flicek P, Duan D, Brent MR (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics (Oxford, England)* 17 Suppl 1: S140–S148
- Kozak M (1981) Possible role of flanking nucleotides in recognition of the aug initiator codon by eukaryotic ribosomes. *Nucleic Acids Res* 9: 5233–5252
- Krogh A (1997) Two methods for improving performance of an hmm and their application for gene finding. *Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. Int Conf Intell Syst Mol Biol* 5: 179–186
- Krogh A, Mian IS, Haussler D (1994) A hidden Markov model that finds genes in e. coli dna. *Nucleic Acids Res* 22: 4768–4778
- Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehrab O, Guigó R, Gladyshev VN (2003) Characterization of mammalian selenoproteomes. *Science (New York, N.Y.)* 300: 1439–1443, doi: 10.1126/science.1083516
- Kulp D, Haussler D, Reese MG, Eeckman FH (1996) A generalized hidden Markov model for the recognition of human genes in dna. *Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. Int Conf Intell Syst Mole Biol* 4: 134–142
- Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 33: 6494–6506, doi: 10.1093/nar/gki937
- Majoros WH, Salzberg SL (2004) An empirical analysis of training protocols for probabilistic gene finders. *BMC Bioinformatics* 5: 206, doi: 10.1186/1471-2105-5-206
- Majoros WH, Pertea M, Salzberg SL (2005) Efficient implementation of a generalized pair hidden Markov model for comparative gene finding. *Bioinformatics (Oxford, England)* 21: 1782–1788, doi: 10.1093/bioinformatics/bti297
- McAuliffe JD, Pachter L, Jordan MI (2004) Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics (Oxford, England)* 20: 1850–1860, doi: 10.1093/bioinformatics/bth153
- Meyer IM, Durbin R (2002) Comparative ab initio prediction of gene structures using pair hmms. *Bioinformatics (Oxford, England)* 18: 1309–1318
- Mott R (1997) Est genome: a program to align spliced dna sequences to unspliced genomic dna. *Computer applications in the biosciences: CABIOS* 13: 477–478
- Ng A, Jordan M (2001) On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. In *NIPS*, pp 841–848
- Ng P, Wei C-L, Sung W-K, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu E, Ruan Y (2005) Gene identification signature (gis) analysis for transcriptome characterization and genome annotation. *Nat Meth* 2: 105–111, doi: 10.1038/nmeth733

- Parra G, Blanco E, Guigó R (2000) Geneid in drosophila. *Genome Res* 10: 511–515
- Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigó R (2003) Comparative gene prediction in human and mouse. *Genome Res* 13: 108–117, doi: 10.1101/gr.871403
- Pedersen JS, Hein J (2003) Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* (Oxford, England) 19: 219–227
- Peters LM, Belyantseva IA, Lagziel A, Battey JF, Friedman TB, Morell RJ (2007) Signatures from tissue-specific mpss libraries identify transcripts preferentially expressed in the mouse inner ear. *Genomics* 89: 197–206, doi: 10.1016/j.ygeno.2006.09.006
- Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 77: 257–286
- Rätsch G, Sonnenburg S, Schäfer C (2006) Learning interpretable svms for biological sequence classification. *BMC Bioinformatics* 7 Suppl 1: S9, doi: 10.1186/1471-2105-7-S1-S9
- Rätsch G, Sonnenburg S, Srinivasan J, Witte H, Müller K-R, Sommer R-J, Schölkopf B (2007) Improving the caenorhabditis elegans genome annotation using machine learning. *PLoS Comput Biol* 3: e20, doi: 10.1371/journal.pcbi.0030020
- Reese M, Hartzell G, Harris N, Ohler U, Abril J, Lewis S (2000) Genome annotation assessment in drosophila melanogaster. *Genome Res* 10: 483–501
- Rogic S, Mackworth AK, Ouellette FB (2001) Evaluation of gene-finding programs on mammalian sequences. *Genome Res* 11: 817–832, doi: 10.1101/gr.147901
- Roma G, Cobellis G, Claudiani P, Maione F, Cruz P, Tripoli G, Sardiello M, Peluso I, Stupka E (2007) A novel view of the transcriptome revealed from gene trapping in mouse embryonic stem cells. *Genome Res* 17: 1051–5720807, doi: 10.1101/gr.5720807
- Salamov AA, Solovyev VV (2000) Ab initio gene finding in drosophila genomic dna. *Genome Res* 10: 516–522
- Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* 26: 544–548
- Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *P Natl Acad Sci USA* 100: 15776–15781, doi: 10.1073/pnas.2136655100
- Siepel A, Haussler D (2004) Combining phylogenetic and hidden Markov models in bio-sequence analysis. *J Comput Biol: A J Comput Mole Cell Biol* 11: 413–428. 10.1089/1066527041410472
- Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics [electronic resource]* 6: 31, doi: 10.1186/1471-2105-6-31
- Solovyev VV, Salamov AA, Lawrence CB (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. *Proceedings/. . . International Conference on Intelligent Systems for Molecular Biology; ISMB. Int Conf Intell Syst Mole Biol* 3: 367–375
- Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) Augustus: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 34: W435–W439, doi: 10.1093/nar/gkl200
- Sun Y-F, Fan X-D, Li Y-D (2003) Identifying splicing sites in eukaryotic rna: support vector machine approach. *Comput Biol Med* 33: 17–29
- The ENCODE Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature* 447: 799–816
- Ueberbacher EC, Mural RJ (1991) Locating protein-coding regions in human dna sequences by a multiple sensor-neural network approach. *P Natl Acad Sci USA* 88: 11261–11265
- Wei C, Brent MR (2006) Using ests to improve the accuracy of de novo gene prediction. *BMC Bioinformatics* 7: 327, doi: 10.1186/1471-2105-7-327

-
- Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA (2001) Database resources of the national center for biotechnology information. *Nucleic Acids Res* 29: 11–16
- Wu T, Watanabe C (2005) Gmap: a genomic mapping and alignment program for mrna and est sequences. *Bioinformatics (Oxford, England)* 21: 1859–1875, doi: 10.1093/bioinformatics/bti310
- Xu Y, Einstein JR, Mural RJ, Shah M, Uberbacher EC (1994) An improved system for exon recognition and gene modeling in human dna sequences. *Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. Int Conf Intell Syst Mole Biol* 2: 376–384
- Yeh RF, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. *Genome Res* 11: 803–816, doi: 10.1101/gr.175701
- Zhang XH-F, Heller KA, Hefter I, Leslie CS, Chasin LA (2003) Sequence information for the splicing of human pre-mrna identified by support vector machine classification. *Genome Res* 13: 2637–2650, doi: 10.1101/gr.1679003
- Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning dna sequences. *J Comput Biol: A J Comput Mole Cell Biol* 7: 203–214, doi: 10.1089/10665270050081478

CHAPTER 1.2

Quality control of gene predictions

A. Nagy, H. Hegyi, K. Farkas, H. Tordai, E. Kozma, L. Bányai and L. Patthy

Biological Research Center, Hungarian Academy of Sciences, Institute of Enzymology, Budapest, Hungary

1 Introduction

A recent study has systematically compared the performance of various computational methods to predict human protein-coding genes (Guigó et al. 2006). In this study a set of well annotated ENCODE sequences were blind-analyzed with different gene finding programs and the predictions obtained were compared with the annotations. Predictions were analyzed at the nucleotide, exon, transcript and gene levels to evaluate how well they were able to reproduce the annotation. These studies have revealed that none of the strategies produced perfect predictions but prediction methods that rely on mRNA and protein sequences and those that used combined information (including expressed sequence information) were generally the most accurate. The dual- or multiple genome methods were less accurate, although performing better than the single genome *ab initio* prediction methods. Importantly, at the nucleotide level no prediction method correctly identified greater than ~90% of nucleotides and at the transcript level (the most stringent criterion) no prediction method correctly identified greater than 45% of the coding transcripts.

Computational gene prediction pipelines, such as Ensembl (Hubbard et al. 2007) and NCBI's GNOMON (Gnomon description 2003) are automated techniques for large datasets, which means that the resultant sequences are not individually analyzed for possible errors. Mispredicted gene and protein sequences may accumulate within such resources, without any accompanying annotation to state that the predicted gene or protein sequence might be erroneous. This problem is likely to be most severe in the case of genomes where gene prediction is only weakly supported by expressed sequence information or information obtained by comparative genomics.

The key question is: are there signs that may indicate that the predicted structure of a protein-coding gene might be erroneous? The rationale of the MisPred project (Nagy

Corresponding author: Laszlo Patthy, Biological Research Center, Hungarian Academy of Sciences, Institute of Enzymology, Budapest, Hungary (e-mail: patthy@enzim.hu)

et al. 2007) is that a protein-coding gene is likely to be mispredicted if some of its features (or features of the protein it encodes) conflict with our current knowledge about protein-coding genes and proteins. As will be illustrated below, the MisPred pipeline can detect such conflicts, thereby providing valuable tools for the quality control of gene predictions.

2 Quality control of gene predictions

2.1 Principles of quality control

Each MisPred tool is based upon generally accepted rules about the properties of protein-coding genes and correctly folded, functionally competent protein molecules. Each tool combines reliable bioinformatic methodologies, as well as in-house programs, to analyze protein sequences. Any sequence, which is considered to be in conflict with one of these rules, is deemed to be erroneous (i.e. abnormal or mispredicted). By identifying erroneous protein sequences, the MisPred pipeline serves to inform both the creators of the predictive algorithms as well as experimentalists of the reliability of predictions, to thereby assist in the improvement of the quality of the available datasets. The principles of quality control are illustrated below with five of the MisPred tools. For each of these five tools, the MisPred pipeline contains specific routines, each focusing on a special type of conflict with one of the dogmas.

2.1.1 Violation of some generally valid rules about proteins

The first three Conflicts are based on the concept that since some protein domains occur exclusively in the extracytoplasmic space, some occur only in the cytoplasm and others are found only in the nucleus, the domain composition of proteins may be used to predict their subcellular localization (Mott et al. 2002; Tordai et al. 2005). The subcellular localization of proteins, however, is determined primarily by appropriate sequence signals therefore the presence (or absence) of such signals must be in harmony with the domain composition of the protein. Proteins that violate these rules are considered to be erroneous.

For the domain-based prediction of subcellular localization of proteins only those Pfam-A domain families (Finn et al. 2006) have been incorporated into the MisPred pipeline that are exclusively extracellular, cytoplasmic or nuclear, respectively. Pfam-A domains that are known not to be restricted to a particular cellular compartment, such as immunoglobulin domains and fibronectin type III domains (i.e., domains that are “multilocale”), were not utilized in these analyses. Our domain co-occurrence analyses (Tordai et al. 2005) have identified 166 obligatory extracellular, 115 obligatory cytoplasmic and 126 obligatory nuclear Pfam-A domain families as being restricted

to the respective subcellular compartment, the majority of which are also identified as such in the SMART database (Letunic et al. 2004).

2.1.1.1 Conflict between the presence of extracellular Pfam-A domain(s) in a protein and the absence of appropriate sequence signals

This MisPred tool (hereafter referred to as **Conflict 1**) identifies proteins containing extracellular Pfam-A domains which occur exclusively in extracellular proteins or extracytoplasmic parts of type I, type II, and type III single pass or multispinning transmembrane proteins and examines whether the proteins also have secretory signal peptide, signal anchor or transmembrane segments that could target these domains to the extracellular space. Proteins that contain obligatory extracellular domains but lack secretory signal peptide, signal anchor and transmembrane segment(s) are considered erroneous since in the absence of these signals their extracellular domain (usually rich in disulfide-bonds) will not be delivered to the extracytoplasmic space where it is properly folded, stable and functional. Mislocalized extracellular domains are likely to be misfolded in the reductive milieu of the cytoplasm and such proteins are likely to be rapidly degraded by the protein quality control system of the cell.

2.1.1.2 Conflict between the presence of extracellular and cytoplasmic Pfam-A domains in a protein and the absence of transmembrane segments

This MisPred tool (hereafter referred to as **Conflict 2**) is based on the principle that multidomain proteins that contain both obligatory extracellular and obligatory cytoplasmic domains must have at least one transmembrane segment to pass through the cell membrane. The MisPred tool associated with Conflict 2 identifies proteins containing both extracellular and cytoplasmic Pfam-A domains and examines whether they also contain transmembrane helices. If a protein contains both obligatory extracellular and cytoplasmic domains but lacks transmembrane segment(s) it is considered to be erroneous.

2.1.1.3 Co-occurrence of nuclear and extracellular domains in a predicted multidomain protein

This MisPred tool (hereafter referred to as **Conflict 3**) is based upon the rule that protein domains that occur exclusively in the extracellular space and those that occur exclusively in the nucleus do not co-occur in a single multidomain protein (Tordai et al. 2005). The explanation for this rule is that a protein that contains both extracellular and nuclear domains would not be delivered to a compartment where both types of domains would be correctly folded and fully functional. Accordingly, proteins that contain both extracellular and nuclear domains are deemed erroneous.

2.1.1.4 Domain size deviation

This MisPred tool (hereafter referred to as **Conflict 4**) is based on the observation that the number of residues in closely related members of a globular domain family usually fall into a relatively narrow range (Wheelan et al. 2000). This phenomenon is due to the fact that insertion (or deletion) of longer segments into (or from) structural domains may yield proteins that are unable to fold efficiently into a correctly folded, viable and stable protein (Wolf et al. 2007). The MisPred tool for Conflict 4 used Pfam-A domain families that have a well-defined and conserved sequence length range and well-characterized members of the family (in the UniProtKB/Swiss-Prot database) did not deviate from the average domain size by more than 2 SD values. Approximately 85% of all Pfam-A families present in Metazoa turned out to be suitable for this task. Predicted proteins containing domains that deviate by more than 2 SD from the average length for that domain family were deemed erroneous.

2.1.2 Violation of some generally valid rules of protein-coding genes

2.1.2.1 Chimeric proteins parts of which are encoded by exons located on different chromosomes

This MisPred tool (hereafter referred to as **Conflict 5**) is based on the rule that a protein is encoded by exons located on a single chromosome. If a predicted protein sequence is identified as chimeric by this MisPred tool it is deemed erroneous.

3 Results

3.1 Validation of the MisPred pipeline

The Swiss-Prot section of UniProtKB is the gold standard of protein databases. The information available therein is integrated with other databases and each entry is manually annotated and curated by experts in the field. We have used Swiss-Prot as the benchmark with which to validate the concepts behind the MisPred pipeline, based

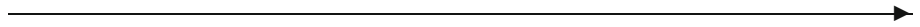


Fig. 1 Error detected by MisPred routine for Conflict 1: the case of the Swiss-Prot entry YL15_CAEEL (Q11101). The hypothetical homeobox protein C02F12.5 predicted for chromosome X contains an extracellular Kunitz_BPTI domain but was found to lack both a signal peptide and transmembrane helices. This protein, that also contains a nuclear Homeobox domain, arose through the *in silico* fusion of a gene related to the homeobox protein HM07_CAEEL (P20270) and the Kunitz_BPTI containing protein CBG14258, Q619J1_CAEER. (A) Alignment of YL15_CAEEL and Q619J1_CAEER shows close homology only in the C-terminal region. (B) Alignment of the YL15_CAEEL_corr1 (the corrected version of the N-terminal constituent of YL15_CAEEL) and HM07_CAEEL. (C) Alignment of YL15_CAEEL_corr2 (the corrected version of the C-terminal constituent of YL15_CAEEL) and Q619J1_CAEER

A	
	1 50
q619j1_caehr:	~~~~~
y115_caee1	MTSKTRMTSNKFAYDFFPWSNDTNSSQQIKNIKPPPKRSNRP TKRRTFTS
	51 100
q619j1_caehr:	~~~~~MFVWSAAVLIFSSVVP TFAQYGC I . . . SE
y115_caee1	EQVTLLELEFAKNEYI CKDRRGELAQ T IELTECQVKTWF QNRRTKKRSSE
	101 150
q619j1_caehr:	<u>LTF GKACPQNKTS TKWFFDAKLSF CYPYQFLG CDEGSNSFES SDI CLES C</u>
y115_caee1	<u>LKFGTACSENKTS TKWYYD SKLLF CYPYKYLGC GEGSNSFESNENCLSEC</u>
	151 200
q619j1_caehr:	<u>KPADQF SCGGNTDADGICF SP SDS GCKKGTDCVMGGNIGF CCNKATQDEW</u>
y115_caee1	<u>KPADQF SCGGNTGPDGVCF AHGDQ GCKKGTVCVMGGMVGF CCDKKIQDEW</u>
	201 250
q619j1_caehr:	<u>NKEHSP TCSKGSVVQFKQWFGMTPLIGRNCANHF CPAGSTCIQGWTAHC</u>
y115_caee1	<u>NKENS PKCLKGQVVQFKQWFGMTPLIGRS CSHNF CPEKSTCVQGWTAHC</u>
	251
q619j1_caehr:	<u>CQ</u>
y115_caee1	<u>CQ</u>
B	
	1 50
y115_caee1_corr1	MTSKTRMTSNKFAYDFFPWSNDTNSSQQIKNIKPPPKRSNRP TKRRTFTS
hm07_caee1	~~~~~MKHEMVFTFLMMVREASTSRIPRRRTFTS
	51 100
y115_caee1_corr1	EQVTLLELEFAKNEYI CKDRRGELAQ T IELTECQVKTWF QNRRTKKRSF I
hm07_caee1	EQLYLLEMYFAQSQYVGC DERERLARILSLDEYQVKTWF QNRRIRMRREA
	101
y115_caee1_corr1	~~~
hm07_caee1	NK
C	
	1 50
q619j1_caehr:	MFVWSAAVLIFSSVVP TFAQYGC I SEL TFGKACPQNKTS TKWFFDAKLSF
y115_caee1_corr2	MLFF TLLIQLF . . LVPVL CQYACSELKFGTACSENKTS TKWYYD SKLLF
	51 100
q619j1_caehr:	CYPYQFLG CDEGSNSFES SDI CLESCKPADQF SCGGNTDADGICF SP SDS
y115_caee1_corr2	CYPYKYLGC GEGSNSFESNENCLSECCKPADQF SCGGNTGPDGVCF AHGDQ
	101 150
q619j1_caehr:	GCKKGTDCVMGGNIGF CCNKATQDEWNKEHSP TCSKGSVVQFKQWFGMTP
y115_caee1_corr2	GCKKGTVCVMGGMVGF CCDKKIQDEWNKENS PKCLKGQVVQFKQWFGMTP
	151 178
q619j1_caehr:	LIGRNCANHF CPAGSTCIQGWTAHCQ
y115_caee1_corr2	LIGRSCSHNF CPEKSTCVQGWTAHCQ

on the expectation that very few, if any, of the Swiss-Prot sequences would be truly erroneous. Examination of human, mouse, rat, chick, zebrafish, worm and fly Swiss-Prot entries has indeed revealed that none of the entries violate the rules underlying Conflicts 2 and 3. The majority of truly erroneous sequences were returned for Conflicts 1 and 4, however, these accounted for only 0.05% (human), 0.04% (mouse), 0.31% (rat), 1.16% (chicken), 0.08% (zebrafish), 0.33% (worm) and 0.08% (fly) of the entries.

An example of a Swiss-Prot entry identified as erroneous by Conflict 1 is YL15_CAEEL (Q11101) from *Caenorhabditis elegans*, which has an obligatory extracellular Kunitz_BPTI domain but no secretory signal peptide or transmembrane segment(s). This protein is also in conflict with another rule: in addition to the obligatory extracellular Kunitz_BPTI domain it also contains a predominantly nuclear Homeobox domain. Analysis of this Swiss-Prot entry revealed that the nuclear and extracellular domains of this protein are encoded by distinct, tandem genes whose exons have been erroneously joined *in silico*. The correct structures of the constituent genes could be predicted using the sequences of ‘correct’ homologs (Fig. 1).

In summary, the fact that the number of Swiss-Prot entries identified by MisPred as erroneous is very low, attests to both the high quality of this database and the reliability of the MisPred approach.

3.2 Errors detected by the MisPred tools in public databases

3.2.1 Analysis of the TrEMBL section of UniProtKB

The TrEMBL section of UniProtKB was included in the MisPred analyses since the entries found in this database are employed by various extrinsic gene prediction methodologies and therefore greatly influence the quality of the resultant datasets.

Analysis of the human protein sequences in the TrEMBL database has revealed that the error rate for Conflict 1 (7.49%), Conflict 4 (4.92%) and Conflict 5 (0.33%) are orders of magnitude higher than those for the Swiss-Prot dataset. The large number of erroneous sequences detectable by Conflict 1 and 4 come primarily from protein fragments translated from non-full length cDNAs: the incomplete proteins lack signal peptides, parts of domains etc. Another major source of error is that some of the transcripts arose through aberrant splicing and parts of domains may be missing from the hypothetical proteins they encode.

An example of a TrEMBL entry identified as erroneous by Conflict 4 is Q5T951_HUMAN which is encoded by an alternative transcript of the gene for human carnitine O-acetyltransferase. Figure 2 shows the three dimensional structure of full length human carnitine O-acetyltransferase (CACP_HUMAN, P43155), the region missing from Q5T951_HUMAN is highlighted in yellow. Since Q5T951_HUMAN violates the integrity of Pfam-A domain Carn_acyltransf (PF00755) it is predicted to be a non-viable protein.

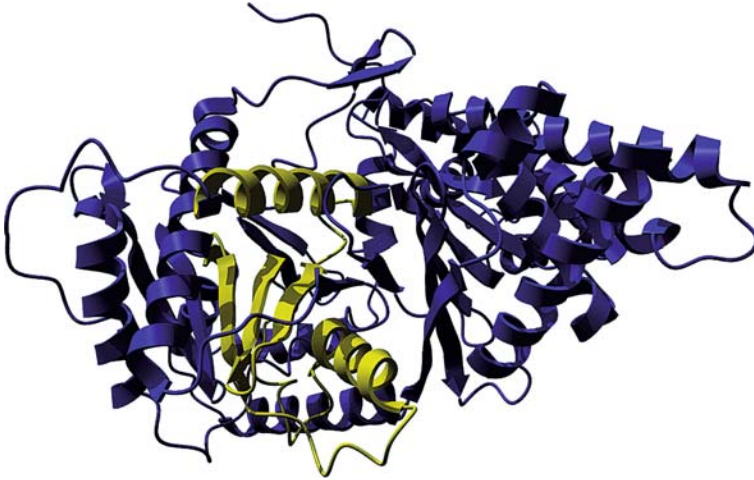


Fig. 2 Error detected by MisPred routine for Conflict 4: the case of TrEMBL entry Q5T951_HUMAN. The figure shows the three dimensional structure (1NM8.pdb) of human carnitine O-acetyltransferase (CACP_HUMAN, P43155). An alternative transcript of human carnitine O-acetyltransferase gene encodes a protein (Q5T951_HUMAN) that lacks several key structural elements (highlighted in yellow) of this domain

The relatively high rate of erroneous human proteins detected by the tool for Conflict 5 reflects the abundance of chimeric proteins generated by chromosomal translocation in cancer cell lines. However, a surprisingly large proportion (37%) of chimeric human TrEMBL entries are not annotated as such and are derived from cDNAs cloned from apparently normal tissues.

No erroneous sequences were returned for either Conflicts 2 or 3. Since the TrEMBL entries are translated from experimentally determined cDNAs, it appears that neither the omission of internal transmembrane helices separating extracellular and cytoplasmic domains (Conflict 2), nor the forbidden co-occurrence of extracellular and nuclear domains (Conflict 3) affects the quality of this database.

3.2.2 Analysis of sequences predicted by the EnSEMBL and GNOMON gene prediction pipelines

MisPred analyses of sequences predicted by the two pipelines have revealed that in the case of both pipelines the majority of erroneous entries are returned for Conflicts 1 and 4 (Table 1).

The relatively high number of erroneous proteins detectable with the tool for Conflict 1 is due to the fact that detection of exons encoding signal peptides is one of the most difficult tasks in gene prediction. In vertebrates, secretory signal peptides are frequently encoded by distinct, short, poorly conserved exons that may be easily missed

Table 1 Results of MisPred analyses of human, opossum, chicken and zebrafish protein sequences predicted by the EnsEMBL- and NCBI/GNOMON-pipelines

	Percent of protein sequences identified as erroneous by MisPred				
	Conflict 1	Conflict 2	Conflict 3	Conflict 4	Conflict 5
EnsEMBL					
<i>Homo sapiens</i>	0.57%	0.00%	0.002%	1.76%	0.00%
<i>Monodelphis domestica</i>	2.02%	0.05%	0.00%	1.17%	0.00%
<i>Gallus gallus</i>	1.71%	0.004%	0.004%	1.58%	0.00%
<i>Danio rerio</i>	3.39%	0.002%	0.00%	1.64%	0.01%
NCBI/GNOMON					
<i>Homo sapiens</i>	0.92%	0.00%	0.00%	2.52%	0.00%
<i>Monodelphis domestica</i>	1.26%	0.01%	0.01%	0.55%	0.00%
<i>Gallus gallus</i>	1.66%	0.01%	0.01%	2.50%	0.00%
<i>Danio rerio</i>	2.22%	0.02%	0.02%	1.52%	0.004%

by the gene finding programs, particularly since few cDNAs and ESTs may be available that cover the 5' parts of vertebrate protein-coding genes. This type of error was particularly evident in chordate species more distantly related to human, mouse and rat. For example, in the case of the EnsEMBL predictions for frog, pufferfish and zebrafish ~40% of proteins containing extracellular domains were identified by Conflict 1 as erroneous. The *Caenorhabditis elegans* and *Drosophila melanogaster* entries returned significantly lower rates of erroneous sequences (~10%) than the non-mammalian chordates analyzed. This is probably due to the fact these invertebrates have compact (intron-poor) genomes and these features significantly increase the accuracy of gene-prediction.

A significant proportion (>1%) of sequences were returned for Conflict 4 for all vertebrates analyzed, suggesting that erroneous omission or insertion of exons (causing deviation of domain size) is a major source of error in gene prediction. This problem appeared to be less serious in the case of *Caenorhabditis elegans* and *Drosophila melanogaster* (error rate is ~0.3%), probably due to the fact these invertebrates have intron-poor genomes and this significantly increases the reliability of gene-prediction. Conflict 2 identified very few erroneous sequences (i.e. proteins containing extracellular and cytoplasmic domains but lacking transmembrane helices) among those predicted by the EnsEMBL and GNOMON pipelines. This is probably due to the fact transmembrane helices (unlike secretory signal peptides) are found in the middle or at the 3' parts of genes – regions which are better represented in cDNA and EST libraries. Analysis of the few erroneous sequences returned for Conflict 2 by both methodologies revealed that they usually resulted from *in silico* fusion of tandem genes encoding extracellular and cytoplasmic proteins. The same type of error, i.e. the *in silico* fusion of

distinct, tandem genes encoding extracellular and nuclear proteins, respectively, also accounts for the relatively few sequences returned for Conflict 3 by both prediction pipelines. It is noteworthy in this respect that such mispredicted sequences were principally encountered in the case of *Fugu rubripes* proteins, which may be due to the fact that intergenic distance is shorter in the compact genome of the pufferfish.

The MisPred tool for Conflict 5 identified no human EnsEMBL or GNOMON-predicted entries as being chimeric. This result is not unexpected in view of the high quality of contig assembly and chromosomal assignment in the case of the human genome. The *Danio rerio* sequences analyzed, however, returned a few mispredicted entries for Conflict 5 by both prediction pipelines, which may be attributed to inaccurate contig assembly and/or chromosomal assignment in the case of the zebrafish genome.

Although the proportion of erroneous human sequences generated by the NCBI/GNOMON pipeline is slightly greater for both Conflict 1 and Conflict 4 than in the cases of EnsEMBL-predicted sequences (see Table 1), this does not necessarily reflect a difference in the reliability of the two methodologies. Note for example, that in the case of opossum the error rate seems to be lower for GNOMON-predicted sequences. A major difference between the two datasets is that the EnsEMBL database is a comprehensive source of both known genes, well characterized experimentally, as well as predicted genes, whereas NCBI's GNOMON-predicted sequences (distinguished by unique identifiers) are devoid of well-characterized genes. Accordingly, the differences we observed may be a consequence of the difference in gene populations covered by the two datasets. In order to directly compare the performance of the two pipelines MisPred analyses were performed on only those protein-coding genes for which both EnsEMBL and GNOMON have made at least one prediction. Analysis of these datasets revealed

Table 2 Results of MisPred analyses of human, opossum, chicken and zebrafish protein sequences predicted by the EnsEMBL- and NCBI/GNOMON-pipelines in the case of genes for which both pipelines have at least one prediction

	Percent of protein sequences identified as erroneous by MisPred				
	Conflict 1	Conflict 2	Conflict 3	Conflict 4	Conflict 5
EnsEMBL					
<i>Homo sapiens</i>	0.83%	0.00%	0.00%	1.73%	0.00%
<i>Monodelphis domestica</i>	1.30%	0.00%	0.00%	1.13%	0.00%
<i>Gallus gallus</i>	1.84%	0.00%	0.00%	2.59%	0.00%
<i>Danio rerio</i>	3.08%	0.01%	0.00%	1.91%	0.00%
NCBI/GNOMON					
<i>Homo sapiens</i>	1.06%	0.00%	0.00%	1.36%	0.00%
<i>Monodelphis domestica</i>	1.15%	0.02%	0.01%	0.40%	0.00%
<i>Gallus gallus</i>	1.57%	0.00%	0.00%	3.71%	0.00%
<i>Danio rerio</i>	1.77%	0.02%	0.01%	3.37%	0.00%

that the differences were significantly diminished, suggesting that there are no major differences in the reliability of the two pipelines (Table 2).

4 Alternative interpretations of the results of MisPred analyses

4.1 MisPred has a low false positive rate

The MisPred pipeline is dependent upon the reliability of bioinformatic programs incorporated into each of the MisPred tools (programs used for the detection of Pfam-A domains, secretory signal peptides, transmembrane helices, chromosomal localization, etc.). However, each tool remains a predictive methodology, with specific parameters and pre-defined sets of heuristics for sequences analysis. This means that sequences may be incorrectly identified as erroneous if, for example, the bioinformatic tools fail to recognize secretory signal peptide sequences or transmembrane helices (Conflicts 1 and 2) or full-length Pfam-A domains (Conflicts 1, 2, 3 and 4). Analyses of the benchmark Swiss-Prot proteins, however, have revealed that the false positive rate generated by the MisPred pipeline is lower than 0.1%.

4.2 MisPred detects errors in gene prediction

The conclusion that a predicted protein is erroneous (i.e. the gene is mispredicted) is reinforced if it is shown that alternative transcripts of the same gene or orthologs/paralogs of the gene encode protein(s) that do not violate the given rule. It is important to point out that the same information that is used to reinforce the conclusion that the protein/gene is mispredicted is also suitable for the correction of that error (see Fig. 1). In this way, the MisPred pipeline provides not only tools for the identification of possible errors but (the associated FixPred pipeline) also aids in the correction of these errors. The final validation of the conclusion that a protein/gene is mispredicted is when we actually correct it (e.g. by targeted search for exons that correct the error).

Although MisPred may help correct major errors in gene prediction, it must be emphasized that the tools for Conflicts 1, 2, 3 and 4 tend to underestimate the number of mispredictions. For example, a limitation of Conflict 4 is that only a fraction of predicted proteins may comprise members of well-characterized Pfam-A domain families suitable for the detection of domain size deviation and only major deviations from normal domain size can be used to detect erroneous proteins. In the case of Conflicts 1, 2 and 3 a serious limitation is that only a relatively small fraction of Pfam-A domain families (~11%) can be used as unambiguous markers of extracellular/sub-cellular localization.

Since in the case of human genes it is estimated that at the transcript level about 55% of the genes may be mispredicted (Guigó et al. 2006), the fact that MisPred detects about

2–3% indicates how far away we still are from identifying all transcript-level errors. Additional tools are being developed to close the gap between the number of mis-predicted proteins identified by the current tools and those that remain unidentified.

4.3 MisPred detects “errors” of biological processes

The MisPred tools identified erroneous proteins not only in databases containing sequences predicted by gene prediction pipelines but also among proteins translated from experimentally validated cDNAs and ESTs (Tress et al. 2007). The majority of the erroneous proteins detected by the MisPred tool for Conflict 1 (e.g. lacking signal peptides) in such experimental databases (e.g. TrEMBL) are likely to be mislocalized and rapidly degraded by the quality control system of the cell, therefore such abnormal proteins are unlikely to fulfill a meaningful biological function. Similarly, the erroneous proteins detected by MisPred tool for Conflict 4 (e.g. proteins with truncated domains, see Fig. 2) are likely to be misfolded and rapidly degraded. The transcripts encoding such non-viable versions of ‘normal’ proteins usually arise by aberrant splicing of the primary transcript. Recent studies have indeed shown that although some 40–80% of human multiexon genes produce splice variants, there is little evidence that these isoforms have a role as functional proteins (Tress et al. 2007).

The majority of chimeric transcripts identified by Conflict 5 are formed from chimeric genes resulting from chromosomal translocation principally in tumor cell lines. An alternative explanation for the origin of interchromosomal transcripts is that chimeric proteins might be formed through transchromosomal transcription, i.e. if genes located on different chromosomes are expressed in the same “transcription factory” they may give rise to chimeric transcripts (Unneberg and Claverie 2007).

4.4 MisPred discovers exceptions to generally valid rules

Analysis of all of the datasets has revealed that there exist genuine exceptions to some of the rules upon which the MisPred tools are based. For example, some of the proteins identified as erroneous by Conflict 1 turned out to be false positives: they are secreted to the extracellular space via non-classical means through leaderless secretion (Bendtsen et al. 2004).

Similarly, analyses of the sequences returned for Conflicts 1, 2, and 3 have identified cases where members of a domain family previously considered to be restricted to a particular cellular compartment, turned out to be multilocale.

The observation that many of the chimeric proteins identified in the TrEMBL database by the tool for Conflict 5 originated from apparently normal tissues raises the possibility that formation of chimeric proteins may be more common in normal cells than previously thought, suggesting that transchromosomal transcription may be quite significant (Unneberg and Claverie 2007).

5 Conclusions

Since the MisPred pipeline is able to detect erroneous protein sequences it is useful in the quality control of the prediction of protein-coding genes. First, by identifying mis-predicted entries in public databases, it may inform users about possible errors. Second, by pinpointing the actual errors in predictions, it may guide the correction of these errors.

References

- Bendtsen J, Jensen L, Blom N, Von Heijne G, Brunak S (2004) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Design Selection* 17: 349–356
- Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34: D247–D251
- Gnomon description (2003) <http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.html>
- Guigó R, Flicek P, Abril J, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic V, Birney E, Castelo R, Eyraes E, Ucla C, Gingeras T, Harrow J, Hubbard T, Lewis S, Reese M (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol* 7(Suppl 1): S2.1–S3.1
- Hubbard T, Aken B, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer S, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Overduin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez X, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E (2007) Ensembl 2007. *Nucleic Acids Res* 35: D610–D617
- Letunic I, Copley R, Schmidt S, Ciccarelli F, Doerks T, Schultz J, Ponting C, Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res* 32: D142–D144
- Mott R, Schultz J, Bork P, Ponting C (2002) Predicting protein cellular localization using a domain projection method. *Genome Res* 12: 1168–1740
- Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L (2007) MisPred Database for mispredicted and abnormal proteins. <http://mispred.enzim.hu/index.html>
- Tordai H, Nagy A, Farkas K, Banyai L, Patthy L (2005) Modules, multidomain proteins and organismic complexity. *FEBS J* 272: 5064–5078
- Tress M, Martelli P, Frankish A, Reeves G, Wesselink J, Yeats C, Olason P, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski R, López G, Sadowski M, Watson J, Fariselli P, Rossi I, Nagy A, Kai W, Starling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramírez F, Schlicker A, Denoeud F, Jones P, Kerrien S, Orchard S, Antonarakis S, Reymond A, Birney E, Brunak S, Casadio R, Guigo R, Harrow J, Hermjakob H, Jones D, Lengauer T, Orengo C, Patthy L, Thornton J, Tramontano A, Valencia A (2007) The implications of alternative splicing in the ENCODE protein complement. *P Natl Acad Sci USA* 104: 5495–5500
- Unneberg P, Claverie J (2007) Tentative mapping of transcription-induced interchromosomal interaction using chimeric EST and mRNA data. *PLoS ONE* 2: e254
- Wheelan S, Marchler-Bauer A, Bryant S (2000) Domain size distributions can predict domain boundaries. *Bioinformatics* 16: 613–618
- Wolf Y, Madej T, Babenko V, Shoemaker B, Panchenko AR (2007) Long-term trends in evolution of indels in protein sequences. *BMC Evol Biol* 7: 19

SECTION 2

Gene regulation and expression

CHAPTER 2.1

Evaluating the prediction of cis-acting regulatory elements in genome sequences

O. Sand, J.-V. Turatsinze and J. van Helden

Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, Boulevard du Triomphe, Bruxelles, Belgium

1 Introduction

Transcriptional regulation plays an essential role in all steps of morphogenesis, by controlling the specific subsets of genes that will be expressed in different cell types, and at different times during embryonic development. The control of gene expression is also crucial to maintain the basic cellular functions (e.g. cell divisions) and the response of the organism to its environment (e.g. metabolic regulation). The spatio-temporal control of gene expression is ensured by interactions between transcription factors and specific loci, called cis-acting regulatory elements.

Since a decade, a large number of computer tools have been developed to predict cis-regulatory elements in genome sequences (see Table 1 for a partial list of existing tools). A first type of approach, called *pattern matching*, consists in predicting the putative binding sites for a given transcription factor, based on some prior knowledge about its specificity for DNA binding. In a second type of approach, called *pattern discovery*, one starts from a set of co-regulated genes to infer motifs that are likely to reflect the binding specificity of some as yet unknown transcription factors.

Pattern matching and pattern discovery tools rely on various biological and statistical assumptions, which determine the choices of a series of parameters. The proper tuning of these parameters has a determinant impact on the quality of the predictions. It is thus necessary to establish criteria for estimating the reliability of the predicted cis-acting elements. The assessment of pattern detection methods (pattern discovery + pattern matching) involves several important choices regarding the testing sets, the testing protocol, and the evaluation statistics. Table 2 summarizes the main parameters for establishing the testing sets.

Corresponding author: Olivier Sand, Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université Libre de Bruxelles, Campus Plaine, CP 263, Bld du Triomphe, 1050 Bruxelles, Belgium (e-mail: oly@bigre.ulb.ac.be)

Table 1 Pattern discovery tools commonly used to detect cis-regulatory elements (partial list)

Program	URL	References	Algorithm	Motif format
<i>AlignACE</i>	http://atlas.med.harvard.edu/cgi-bin/alignace.pl	(Roth et al. 1998)	Gibbs	PSSM
<i>ANN-Spec</i>	http://www.cbs.dtu.dk/~workman/ann-spec/	(Workman and Stormo 2000)	Neural networks + Gibbs	
<i>BioProspector</i>	http://seqmotifs.stanford.edu/	(Liu et al. 2001)	Gibbs	PSSM
<i>Coresearch</i>		(Wolfertetter et al. 1996)	Word counts	PSSM
<i>Consensus</i>	http://bioweb.pasteur.fr/seqanal/interfaces/consensus.html	(Hertz et al. 1990)	Greedy	PSSM
<i>dyad-analysis</i>	http://rsat.scmdb.ulb.ac.be/rsat/	(van Helden et al. 2000b)	Word statistics	Spaced pairs
<i>gibbs</i>	http://bayesweb.wadsworth.org/gibbs/gibbs.html	(Neuwald et al. 1995; Neuwald et al. 1997)	Gibbs	PSSM
<i>GLAM</i>	http://zlab.bu.edu/glam/	(Frith et al. 2004)	Gibbs + simulated annealing	
<i>Hermes</i>			Word statistics	Words
<i>Improbizer</i>	http://www.soe.uccs.edu/~kent/improbizer/improbizer.html		EM	
<i>MEME</i>	http://meme.sdsc.edu/meme/website/meme.html	(Bailey and Elkan 1994)	EM	PSSM
<i>MITRA</i>	http://www.w1.cs.columbia.edu/compbio/mitra/	(Eskin and Pevzner 2002)	Composite patterns	
<i>MotifSampler</i>	http://www.esat.kuleuven.ac.be/~thijs/Work/MotifSampler.html	(Thijs et al. 2001)	Gibbs	PSSM
<i>NestedMICA</i>	http://www.sanger.ac.uk/Software/analysis/nmica/	(Down and Hubbard 2005)		PSSM
<i>oligo-analysis</i>	http://rsat.scmdb.ulb.ac.be/rsat/	(van Helden et al. 1998)	Word statistics	Words
<i>PhyloGibbs</i>	http://www.imsc.res.in/~rsidd/phylogibbs/		Gibbs sampler with pre-selection for conserved blocks (using Dialign)	PSSM
<i>QuickScore</i>	http://algo.inria.fr/dolley/QuickScore/	(Regnier and Denise 2004)	Word statistics	Words
<i>SeSiMCMC</i>	http://favorov.imb.ac.ru/SeSiMCMC/	(Favorov et al. 2005)	Gibbs	
<i>SPAT</i>	http://stat.genopole.cnrs.fr/spatt/	(Nuel 2005)		
<i>Trawler</i>	http://ani.embl.de/trawler/	(Ettwiller et al. 2007)	Word statistics	
<i>Weeder</i>	http://159.149.109.16/Tool/ind.php	(Pavesi et al. 2001, 2004)	Consensus-based	
<i>YMF</i>	http://bio.cs.washington.edu/software.html	(Sinha and Tompa 2003)	Word statistics	Words

Table 2 Typology of data sets used to evaluate cis-regulatory element predictions

Usage	Context	Sites	Advantages	Limitations
Positive control: evaluation of sensitivity, specificity, accuracy.	Artificial sequences (e.g. generated with a Markov model)	Artificial sites (e.g. generated from a PSSM)	Control on all the parameters (number of sites, motif variations, sequence composition, sequence length). Useful to check theoretical models	Performances might differ between artificial sets and real conditions
	Artificial sequences	Implanted biological sites	All the “positive” sites (implanted) are known	Performances mainly reflect the fit between random model of sequence generator and that of the motif detector
	Biological sequences	Biological sites in their context	All the true sites are available for the predictor, even if they are not annotated yet	Answer can be obtained from databases. Programs can be over-fitted because parameters were estimated with the same DB. Some real sites can be absent from the annotation (“ <i>false false positives</i> ” FFP)
	Biological sequences	Implanted biological sites	All the “positive” sites (implanted) are supposedly known	The number of implanted sites might differ from natural conditions. Annotation-based: under-estimation (many sites are not annotated)
Negative control: estimation of the rates of false positives.	Artificial sequences	None	Control on the sequence composition (background model)	Performances mainly reflect the fit between random models of predictor and of sequence generator, respectively
	Random selection of biological sequences	Not specific to the considered factor	Indicates the rate of false positive in real conditions	For pattern matching, one may occasionally include some regulated gene in the random selection. For pattern discovery, this should not make a problem, since regulatory signals will be “diluted” among promoters regulated differently

In this chapter, we will present methods for evaluating the quality of the predictions of cis-acting regulatory elements. We will start from some concrete study cases, which will serve us to introduce basic concepts. We will then generalize the concepts and present some protocols for a large-scale evaluation of cis-regulatory element prediction. In the conclusion, we will propose a few general good practices for the assessment of predictive tools.

2 Transcription factor binding sites and motifs

The binding specificity of a transcription factor relies on its preferential affinity for short DNA sequences, the transcription factor binding sites. We distinguish the concept of *transcription factor binding site (TFBS)* from that of *motif*.

A *binding site* is a location where a transcription factor binds on DNA sequence. It corresponds to a precise fragment of DNA, which can be described by its position (start,

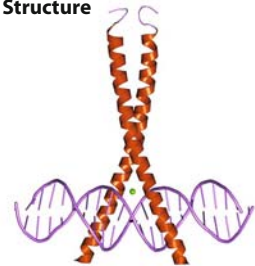
A Aligned sites

#	Site ID	Site sequence	Gene	Start	End
1	R01508	T G A C G T C A C G	RNU4C	-247	-217
2	R15485	T G A C G T C A A G	MITF	-153	-134
3	R04356	T G A C G T C A A G	CCNA	-74	-57
4	R16066	T G A C G T C A C C	CRF	-226	-211
5	R00592	T G A C G T C A T G	TSHA	-142	-117
6	R14792	T G A T G T C A C T	TCR-Vbeta 8.1	-69	-50
7	R12540	T A A C G T C A C A	ARA160	-7324	-7298
8	R04029	T G A C A T C A C G	TPA	-232	-219
9	R16774	G T A C G T C A C G	Egr1	-76	-57
10	R02906	G T A C G T C A C G	BETA2AR	-79	-46
11	R09519	T G A C G T C C A T	DBH	-186	-171
12	R14826	G C A C G T C A A G	BDNF		
13	R16161	T G A C G A C A C	PSEN1		
14	R14781	T C A C G T A A C T	LOR	-120	-90
15	R17209	T G A C G C T A C G	SLC25A3	-138	-116
16	R08102	T G A G C T C A C T	TCP228	-208	-178
17	R04963	T G A C G T C T G A	QM	-189	-166
18	R12331	A A A C G T C A T C	pCIP	-252	-228
19	R14791	A A A T G T C A C A	TCR-Vbeta 8.1	-32	-8

B Consensus

T G A C G T C A C N

E Structure



C Count matrix

		position	1	2	3	4	5	6	7	8	9	10
Residue	A	2	3	19	0	1	1	1	17	5	3	
	C	0	2	0	16	1	1	17	1	11	3	
	G	3	12	0	1	17	0	0	0	1	9	
	T	14	2	0	2	0	17	1	1	2	4	

D Sequence logo



Fig. 1 From binding sites to binding motif. **A:** collection of annotated *binding sites* for the human transcription factor CREB. **B:** degenerate consensus representing the specificity of the transcription factor. **C:** position-specific scoring matrix (PSSM) describing the *binding motif* derived from the aligned annotated binding sites. **D:** sequence logo. **E:** Structure of the interface between the CREB protein and DNA

end), and its sequence. A binding motif is an abstract representation of the binding specificity of a transcription factor. It does not correspond to any particular genomic position, but rather to a collection of sites.

This is exemplified in Fig. 1A, which shows a list of binding sites for the cAMP-response element binding protein (CREB). The sites have been aligned to highlight the conservation of some residues at different positions. This position-specific conservation is represented by a *motif*, which summarizes, in some abstract way, the specificity of the interactions between the CREB transcription factor and DNA interactions. Binding motifs can be described in different ways (Fig. 1B–D). The simplest representation of a binding motif is the consensus (Fig. 1B), which summarizes the residue conservation on the basis of somewhat arbitrary rules (Cavener 1987)

A more expressive way to summarize the residue conservation is the *position-specific scoring matrix* (PSSM, Fig. 1C), where each row represents a residue, each column a position in the alignment of the binding sites, and numbers indicate the counts of residues at each position. The advantage of this representation is that it takes into account all the variations observed at each position of the aligned binding sites.

A *sequence logo* can be derived from the count matrix to provide a visual and intuitive representation of the position-specific residue specificity (Fig. 1D). This representation has been proposed by Schneider and Stephens (1990), as an application of Shannon’s information theory to DNA sequences (Schneider et al. 1986).

3 Scanning a sequence with a position-specific scoring matrix

In this section, we explain how PSSM can be used to detect putative binding sites in a given DNA sequence. Various scoring schemes have been implemented, and we will use here the terminology and statistics proposed by Hertz et al. (1990) and Hertz and Stormo (1999).

In order to detect putative binding sites in a DNA sequence, each segment of this sequence is compared with the PSSM, and assigned a score (called *weight score*), which indicates the likelihood for this segment to be an instance of the motif.

In short, the weight score is the log-likelihood between two probabilities.

$$W_S = \ln \left(\frac{P(S|M)}{P(S|B)} \right)$$

In this formula, $P(S|M)$ is the probability for the sequence segment (S) to be an instance of the motif (M), and $P(S|B)$ the probability of the same sequence segment to appear by chance, according to the background model (B).

Chapter 2.1: Evaluating the prediction of cis-acting regulatory elements in genome sequences

A

Sequence fragment	C	C	A	G	G	A	T	C	T	T		
Prior frequencies										Max info per column	1.39	
Residue	Prior										pseudo-count	1
A	0.249											
C	0.250											
G	0.251											
T	0.250											
sum	1.000											

B

Probability of a sequence segment under the Background model										
residue r	C	C	A	G	G	A	T	C	T	T
P(r B)	0.250	0.250	0.249	0.251	0.251	0.249	0.250	0.250	0.250	0.250
P(S B)	9.53E-07	=PROD[P(r B)]								

$$P(S|B) = \prod_{j=1}^w p_{r_j}$$

C

Counts											
position	1	2	3	4	5	6	7	8	9	10	Sum
A	2	3	19	0	1	1	1	17	5	3	52
C	0	2	0	16	1	1	17	1	11	3	52
G	3	12	0	1	17	0	0	0	1	9	43
T	14	2	0	2	0	17	1	1	2	4	43

D

Frequencies											
position	1	2	3	4	5	6	7	8	9	10	Sum
A	0.11	0.16	1.00	0.00	0.05	0.05	0.05	0.89	0.26	0.16	2.74
C	0.00	0.11	0.00	0.84	0.05	0.05	0.89	0.05	0.58	0.16	2.74
G	0.16	0.63	0.00	0.05	0.89	0.00	0.00	0.00	0.05	0.47	2.26
T	0.74	0.11	0.00	0.11	0.00	0.89	0.05	0.05	0.11	0.21	2.26

$$f_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^4 n_{i,j}}$$

E

Frequencies, corrected											
position	1	2	3	4	5	6	7	8	9	10	Sum
A	0.11	0.16	0.96	0.01	0.06	0.06	0.06	0.86	0.26	0.16	0.27
C	0.01	0.11	0.01	0.81	0.06	0.06	0.86	0.06	0.56	0.16	0.27
G	0.16	0.61	0.01	0.06	0.86	0.01	0.01	0.01	0.06	0.46	0.23
T	0.71	0.11	0.01	0.11	0.01	0.86	0.06	0.06	0.11	0.21	0.23

$$f'_{i,j} = \frac{n_{i,j} + k/4}{\sum_{i=1}^4 n_{i,j} + k}$$

F

Probability of a sequence segment under the matrix model										
residue r	C	C	A	G	G	A	T	C	T	T
P(r M)	0.01	0.11	0.96	0.06	0.86	0.06	0.06	0.06	0.11	0.21
P(S M)	4.26E-10	=PROD[P(r M)]								

$$P(S|M) = \prod_{j=1}^w f'_{r_j}$$

G Weight of a segment, computed from thesegment probabilities
 Weight -7.714 = ln(P(S|M)/P(S|B))

$$W_s = \ln\left(\frac{P(S|M)}{P(S|B)}\right)$$

H

Weight matrix											
position	1	2	3	4	5	6	7	8	9	10	Sum
A	-0.80	-0.43	1.35	-2.99	-1.39	-1.38	-1.38	1.24	0.05	-0.43	-6.15
C	-3.00	-0.80	-3.00	1.18	-1.39	-1.39	1.24	-1.39	0.81	-0.43	-8.16
G	-0.43	0.89	-3.00	-1.39	1.24	-3.00	-3.00	-3.00	-1.39	0.61	-12.47
T	1.05	-0.80	-2.99	-0.80	-2.99	1.24	-1.39	-1.39	-0.80	-0.16	-9.03
sum	-3.18	-1.13	-7.64	-4.00	-4.53	-4.53	-4.53	-4.53	-1.33	-0.41	-35.80

$$w_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

I Weight of a sequence segment, computed from the weight matrix
 residue r C C A G G A T C T T
 W(r) -3.00 -0.80 1.35 -1.39 1.24 -1.38 -1.39 -1.39 -0.80 -0.16
 Weight -7.714 = SUM[W(r)]

$$W_s = \ln\left(\frac{P(S|M)}{P(S|B)}\right) = \sum_{j=1}^w w_{r_j}$$

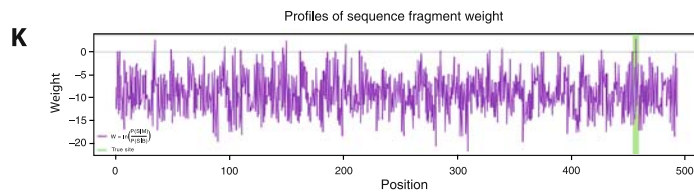
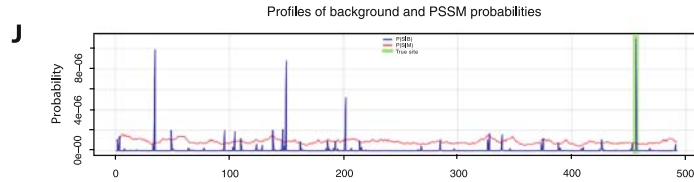


Figure 2 summarizes the steps leading from the count matrix to the weight score. Let us assume that we want to identify putative binding sites in the promoter sequence of Cholecystokinin (Cck), which is a target gene of the transcription factor CREB. This sequence is reproduced below.

```
>RNO|25298|Cck
CCAGGATCTTAAAATTCTGTAAGACTAGAATCCAGGAGGCCAACTGTGATTGAGTTCTGAAAAAT
CTCCCAGAGAACATGCCAGAATTACATTTGCTGACACCTAGTCTGTGAGGGTCCCCCGGTTTCCT
AGACAAACTCCTGCTTCTCTCCGGGAGTAGGGGTGGCACCCCTCCCTGAAGAGGACTCAGCAGAGG
GTACCCCGCTGGGAGGGGCTATCCTCATTCACTGGGCCGTTTCCCTTCTCCCGGGGGGCCACT
TCGATCGGTGGTCTCTCCAGTGGCTGCCTCTGAGCACGTGTCTGCCGGACTGC
```

We will assign a score to each position of this sequence. Since the CREB matrix contains 10 columns, we take a sequence segment of 10 nucleotides, starting from the first position: $S = \text{CCAGGATCTT}$., and compute its score. We will then score the segment of 10 bp starting at the second position (CAGGATCTTA), and so on until the last 10 bp segment of the sequence (GATGACTGGC). Let us see the detail of this computation for the first segment.

3.1 Background probability

The first step is to estimate $P(S|B)$, which is the *background probability* of the sequence $S = \text{CCAGGATCTT}$, i.e. its probability to be generated by chance according to some background model B . Figure 2A shows an example of a simple background model, where each residue has a constant probability. This is called a Bernoulli model, and it assumes that in a genome sequence, residues generally follow each other in an independent way.

Note that Bernoulli models are an over-simplification of biological sequences, because the assumption of independence generally does not hold. For instance, it is well known that genome sequences contain a higher frequency of poly-A or poly-T fragments than what would be expected by the product of residue probabilities. Another clear case of dependency is the avoidance of the CpG dinucleotides in mammalian genomes: the frequency of CG dinucleotide is much lower than the product of frequencies of C and G. In addition, mammalian promoters contain some

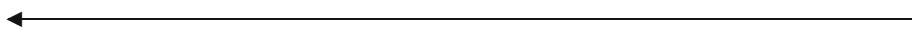


Fig. 2 Utilisation of position-specific scoring matrices to predict binding sites. Example of sequence scanning result with the CREB PSSM. (A) background model. (B) background probability of a sequence segment. (C) Count matrix. (D) frequency matrix. (E) Frequency matrix corrected with a pseudo-count. (F) probability for a sequence segment to be an instance of a motif: $P(S|M)$. (G) Weight of the sequence segment, computed as the log-ratio between motif and background probabilities. (H) Weight matrix. (I) weight score of a sequence segment, directly computed from the weight matrix. (J) probability profiles of a sequence. (K) weight profile of a sequence

regions, called CpG island, where CpG dinucleotides are more frequent than in the rest of the genome. The background model should thus vary according to the genomic region. More elaborate background models can be conceived with Markov chains, but these are out of scope for this introductory chapter. A didactic description of Markov models and their applications to biological sequences can be found in Robin et al. (2005).

Coming back to our example, and if we temporarily accept the Bernoulli simplification (Fig. 2A), the probability of a sequence fragment is simply the product of the prior probabilities of its residues. This leads us to estimate that the probability of the sequence $S = \text{CCAGGATCTT}$ given the background B is $P(S|B) = 9.39\text{E-}7$ (Fig. 2B).

3.2 Probability of a sequence segment given the motif

We now need to estimate the second element of the weight score, the probability of the sequence fragment *given the motif*: $P(S|M)$. For this, we first convert the count matrix (Fig. 2C) into a *frequency matrix* (Fig. 2D) where each cell indicates the relative frequency of a residue at a given position of the aligned binding sites (Fig. 1A).

These frequencies can be used to estimate the probability to observe a given residue at a given position of a binding site (i.e. an instance of the motif). However, if we do so, we will have a problem with the null values. For example, the 4th position of the alignment does not contain any occurrence of the residue A. There is thus a 0 value in the corresponding cell of the PSSM frequency matrix (1st row, 4th column). We should however keep in mind that our PSSM was built on the basis of a relatively small number of annotated binding sites ($n = 19$). The fact that none of the 19 sites annotated in TRANSFAC contains an A at this position might be casual, and we could imagine that, in the future, some experiment will reveal new CREB binding sites with an A at the fourth position. In order to leave some probability for events unobserved so far, we can use a pseudo-count (k) that will be “shared” between all the residues of each column of the matrix in order to obtain *corrected frequencies* (f'_{ij}).

$$f'_{ij} = \frac{n_{ij} + k/4}{\sum_{i=1}^A n_{ij} + k}$$

There is no golden rule for deciding about the most appropriate value for the pseudo-count. In this example, we will arbitrarily set it to $k = 1$. The corrected frequencies (Fig. 2E) can now be used to estimate the probability of the sequence fragment given the motif. For this, we align each letter of our sequence TGTAATAATA with one column of the matrix (Fig. 2F), and we take its probability in the corresponding row. The probability of the sequence is the product of these position-specific residue probabilities, which gives $P(S|M) = 3.62\text{E-}9$ for the sequence CCAGGATCTT.

The *score weight* (W_S) can now be computed as the natural logarithm of the ratio between the two probabilities (Fig. 2G):

$$W_S = \ln\left(\frac{P(S|M)}{P(S|B)}\right) = \ln\left(\frac{9.39E-7}{3.82E-9}\right) = -5.558$$

Since we modeled the sequence according to a Bernoulli model, an equivalent result would have been obtained by pre-computing a *weight matrix* (Fig. 2H), where each cell already contains the log-ratio of the estimated residue probabilities.

$$w_{i,j} = \ln\left(\frac{f'_{ij}}{p_i}\right)$$

The weight matrix nicely indicates the impact of each residue on the final score, as highlighted by the shading of the negative values in the weight matrix: negative scores (shaded cells) indicate residues that are not favorable to the binding. Under the Bernoulli assumption, the weight score of a sequence segment is simply the sum of scores of its residues at the corresponding positions of the weight matrix. It is easy to demonstrate that the two ways to compute the weight score (Fig. 2G and I) give identical results. Consistently, this is what we observe for our example.

$$W_S = \sum_{j=1}^w w_{rj} = -7.714$$

The negative value we obtained suggests that the 10 bp segment at the first position of the Cck promoter has a rather low affinity for the transcription factor CREB, and is thus not a good binding site for it.

3.3 Scanning profiles

We can now apply the same procedure to the 10 bp segments starting at each successive position of the sequence.

```
CCAGGATCTT
 CAGGATCTTA
  AGGATCTTAA
   ...
```

Figure 2J shows the resulting profiles of probabilities for the promoter sequence of Cck. The red profile indicates the background probability $P(S|B)$, and shows erratic variations along the sequence. The motif probability $P(S|M)$ (blue curve) is generally much lower than the background probability, but we observe some very acute peaks. The rightmost of these peaks (highlighted with green background) corresponds to a previously characterized CREB binding site (Haun and Dixon 1990). The other peaks might either be effective binding sites that have not been characterized yet, or false predictions of our scanning program.

4 Evaluating pattern matching results

4.1 Evaluation statistics

If we consider high scoring positions as predicted TF binding sites, how can we evaluate the correctness of such predictions?

The classical procedure relies on a testing set, where the binding sites have been characterized by “wet lab” experiments (DNase footprinting, gel shift, ...). These annotated sites are compared with those predicted (Fig. 3A). This comparison can be done either at the level of the nucleotides (Fig. 3B), or at the level of the sites (Fig. 3C).

The comparison *at the level of the nucleotides* (Fig. 3B) is conceptually simple: each nucleotide of the testing sequences is considered as an individual case, which will take one among four possible statuses (Fig. 3D).

1. *True Positive (TP)* if it belongs to both an annotated site and a predicted site;
2. *True Negative (TN)* if is included neither in a predicted nor in An annotated site;

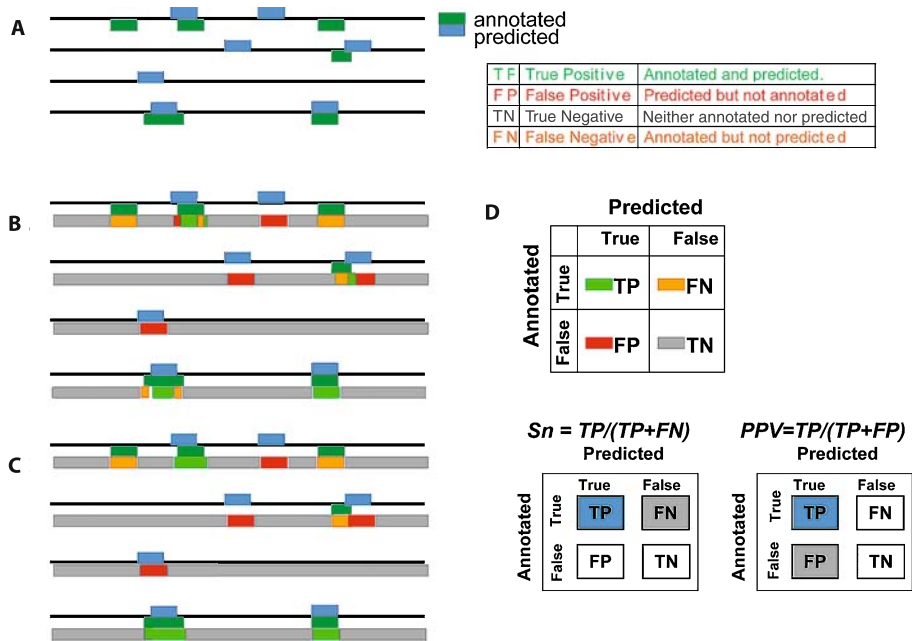


Fig. 3 Schematic illustration of the comparison between annotated and predicted sites. (A) Annotated (dark green) and predicted (blue) binding sites. (B) comparison at the level of nucleotides. (C) comparison at the level of sites. (D) Computation of the sensitivity (Sn) and positive predictive value (PPV) from a contingency table

3. *False Positive (FP)* if it belongs to a predicted site, but no annotated site;
4. *False Negative (FN)* if it belongs to an annotated site, but is not included in any predicted site.

The prediction *at the level of the sites* (Fig. 3C) is based on the same four categories, but generalized to the level of the whole site.

1. A predicted site is considered as TP if it has a sufficient overlap with an annotated site, and as FP otherwise.
2. An annotated site is considered as FN if it is not matched with a sufficient overlap by any predicted site.

The concept of TN seems less intuitive at the level of site, since a true negative would be a non-annotated and non-predicted site. But how would we define the boundaries of a site that is neither annotated, nor predicted? Should we consider the whole region between two annotated sites as a “negative site”? In practice, this is not very important, because, as we will see below, the main statistics derived to estimate the quality of the predictions rely only on TP, FP, and FN, and we have good reasons to carefully avoid using the number of TN.

The four categories defined above can serve as basis for deriving various evaluation statistics (Fig. 3D). The *Sensitivity (Sn)* is the fraction of annotations that is covered by the predictions.

$$Sn = \frac{\text{predicted annotated}}{\text{annotated}} = \frac{TP}{TP + FN}$$

The *Positive Predictive Value (PPV)* is the fraction of predictions that is also found in the annotations.

$$PPV = \frac{\text{predicted annotated}}{\text{predicted}} = \frac{TP}{TP + FP}$$

The Sn and PPV provide us with complementary information about the quality of the predictions. There is generally a tradeoff between Sn and PPV, which can be determined by turning up or down some thresholds on the predictive score, as illustrated on Fig. 4 for the CREB matrix. The dataset used for this evaluation comprises 23 promoter sequences of 500 nucleotides upstream of the transcriptional start site, and containing 25 annotated CREB binding sites. When the lower threshold is set to a trivially low value ($W_s \geq 25$), all the positions of the sequence are predicted as binding sites (Fig. 4A). The sensitivity is thus 100%, but the predictive value of the predictions is almost null (leftmost side of the curves on Fig. 4B). When the score increases, sensitivity shows a step-wise decrease (blue curve), reflecting the progressive “loss” of annotated site (those having a weight score lower than the threshold). In

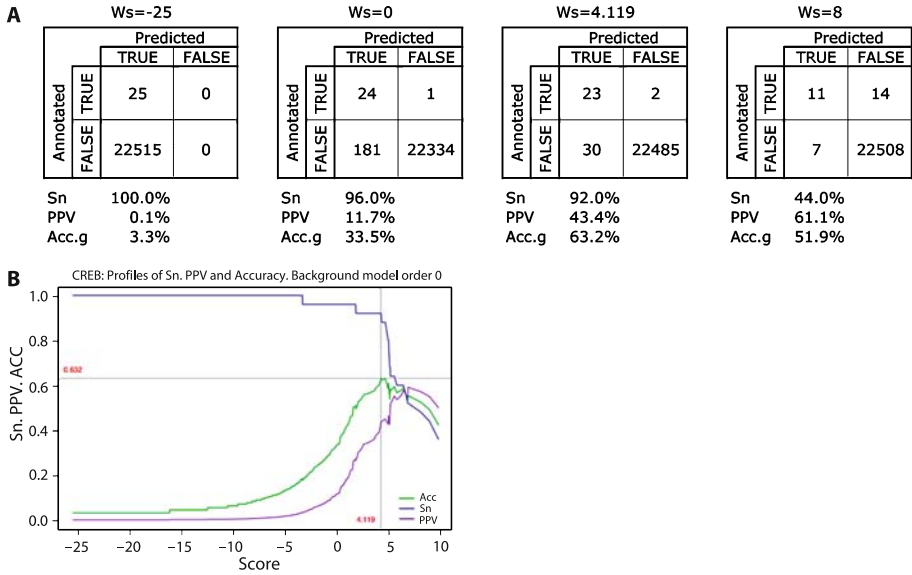


Fig. 4 Impact of the score threshold on the correctness of pattern matching. (A) Examples of Sn, PPV, Acc_g for selected values of weight scores. (B) Sn, PPV and accuracy profiles (internal validation). (C) The same profiles with a Leave-One-Out validation

parallel, the positive predictive value (purple curve) increases with the score, because we discard more and more non-annotated sites from our predictions (the number of FP progressively decreases). The curve is however not monotonous, because the FP is occasionally affected by the loss of an annotated site.

4.2 Accuracy profiles

The tradeoff between sensitivity and PPV is usually measured by computing an accuracy, defined as their arithmetic mean: $Acc_a = (Sn + PPV)/2$. This statistics can however be misleading in some extreme cases. For example, if we set the score to -25 , we will reach a maximal sensitivity ($Sn = 1$) since all possible positions are predicted as sites (this is of course a trivial choice that should never be done in practice). Of course, this sensitivity is at the cost of the PPV, since the fraction of correct predictions is almost null. However, we obtain an accuracy $Acc_a > 0.5$, since it is the arithmetic average between a value of 1 and an almost null value. This accuracy of 0.5 will give us the artificial impression that our matrix is doing a half good job. We will not be able to distinguish this trivial case from a situation where we would predict half of the sites with a reasonably restricted number of predicted sites.

A simple and efficient way to avoid this trap is to compute the accuracy as the geometric mean Acc_g .

$$\text{Acc}_g = \sqrt{\text{Sn} \cdot \text{PPV}}$$

If we select conditions giving either a very low Sn or a very low PPV, the geometric accuracy will be very low. A high geometric mean will be obtained only if we have a good score for *both* Sn and PPV.

Figure 4A shows the values of Sn, PPV and Acc_g derived from the contingency tables for four threshold values. Figure 4B shows the profiles of Sn, the PPV and Acc_g as a function of the weight score W_S . The accuracy curve shows that the optimal tradeoff between Sn and PPV gives an accuracy of 63.2%. This optimal accuracy is obtained by predicting as binding sites all the positions having a score $W_S > 4.119$ with the CREB matrix.

4.3 Avoiding circularity in the evaluation

An important point which is often overlooked, is that the evaluation should rely on set of binding sites that were not used for building the position-specific scoring matrix. However, in the previous section, we tested the Sn, PPV and *Accuracy* on the same binding sites that had been used to build the PSSM. This procedure is called *internal validation*, and it leads to an obvious bias, since each site used to build a matrix contributes in a positive way to its own scoring.

In theory, we would thus need two sets of annotated sites: a *training set*, used to build the PSSM (as in Fig. 1), and a *testing set* for measuring the accuracy of the predictions. Unfortunately, this is not always feasible in practice, because the collections of annotated binding sites are generally too sparse (there are many factors for which we have less than 10 binding sites).

A solution to this problem is to apply a Leave-One-Out (LOO) procedure, which consists in discarding one site (the *left out* element), and building a position-specific scoring matrix with the $n - 1$ remaining sites. This PSSM is used to score the left out site. The procedure is then iterated for each binding site of the collection. We thus obtain n scores, each being computed with a slightly different PSSM, built from $n - 1$ binding sites.

The LOO evaluation is more stringent than the internal evaluation, but it is necessary to obtain an unbiased estimation of the predictive capability of a matrix.

The LOO validation of PSSM has been applied to a collection of PSSM describing transcription factors from drosophila, and this analysis clearly showed that the internal validation clearly over-estimates the predictive qualities of PSSMs (Aerts et al. 2007).

4.4 Why the statistics involving TN should be avoided

The Sn, PPV and Acc_g statistics only use three of the four values of the contingency table: TP, TN and FP. Other statistics have been defined that rely on the number of true

negatives (TN), but these should carefully be avoided when analyzing pattern matching results.

Before explaining the reason for this, we ought to introduce a semantic remark: the statistics that we call here PPV is alternatively called “specificity” (Sp) in some articles. However, the word “*specificity*” is ambiguous, because it has another widely accepted definition.

$$Sp = \frac{\text{neither annotated nor predicted}}{\text{not annotated}} = \frac{TN}{FP + TN}$$

According to the latter definition, the specificity measures the capability of a method to “reject” the non-annotated features. However, in typical conditions, this score is strongly affected by the overwhelming predominance of negative elements. Indeed, cis-acting elements typically cover a very small fraction of the sequences, so that a program that predicts a reasonably small number of sites will always have a very high Sp score, even if its predictions are wrong.

For the sake of illustration, let us consider a simple numeric example: a sequence of 10,000 nucleotides is annotated with 8 binding sites, each of length 10. Some pattern matching program predicts 20 wrong sites (also of length 10), and does not detect any of the annotated sites. This result is obviously very poor, but the specificity will fail to indicate how bad the situation is. We can compute the specificity for this example in the following way.

- All the predictions are false, so that the number of false positives (in terms of nucleotides) is $FP = 20 \cdot 10 = 200$.
- The true negative nucleotides are those that are neither predicted, nor annotated: $TN = 10,000 - 20 \cdot 10 - 8 \cdot 10 = 9720$.
- The FP and TN values can now be entered in the formula:

$$Sp = \frac{TN}{FP + TN} = \frac{9720}{200 + 9720} = 0.98$$

The specificity is 98%, despite the fact that our predictions are completely incorrect! This simple example shows that the specificity (as defined here) is generally misleading, and should never be used to assess the correctness of predicted genomic features. More generally, all the scores that involve the TN should be avoided for the evaluation of pattern matching results.

4.5 Difficulties for the evaluation of pattern matching

In summary, the evaluation of pattern matching relies on a set of sequences where the binding sites for a given transcription factor are annotated. The correctness of the predictions is estimated with three statistics: Sn, PPV, and Acc_g.

Although this protocol seems fairly simple to apply, a recurrent problem is the incompleteness of the annotations. Furthermore, although several databases contain annotations of transcription factor binding sites, none of them can claim to contain fully annotated sequences. Databases depend on a human annotation effort, and there is always a lag between the publication of binding sites and their encoding in the database. But more importantly, the experimental characterization of binding sites requires heavy experiments, and many actual binding sites are likely to have escaped experimental detection so far.

As a direct consequence of the incompleteness of annotations, the predictions that we label as “false positives” might include some active binding sites, which have still not been characterized experimentally. The estimated rate of false positives is thus likely to be an over-estimate (and, accordingly, the PPV is under-estimated).

A more general problem is that the interactions between a transcription factor and DNA are not all-or-none. For thermodynamical reasons, it is energetically more favorable for a transcription factor to be bound to DNA than to float in the nucleoplasm. Transcription factors generally have “some” affinity for any piece of DNA, and their specificity results from their higher affinity for some particular successions of nucleotides. The concept of “binding site” is a convenient simplification to denote some positions that are bound by the factor with a sufficiently high affinity to be detected by methods such as gel shift assays or DNase footprinting. It would however be more correct to consider DNA as a “landscape” displaying zones of more or less intense affinity for the factor. Some recent methods explicitly formulate models that take into account this affinity landscape, as described by Manke et al. in the next chapter.

5 Discovering motifs in promoter sequences

In the second part of this chapter, we will discuss the evaluation of pattern discovery results. A typical situation is to analyze a set of genes showing similar expression profiles in some microarray experiment, as can be retrieved from the ArrayExpress microarray database (see the chapter by Brazma et al. in the same volume). We suppose that the similarity between these expression profiles is due to some transcription factor that regulates them in a coordinated fashion. However, we ignore the identity of this hypothetical transcription factor, and we have no idea about its binding motif. Thus, we cannot use any predefined motif to apply pattern matching, as we did above. Instead, we will try to discover motifs (patterns), without any other information than the promoter sequences of the co-regulated genes.

Since 1997, this problem of pattern discovery has attracted a particular attention from bioinformaticians and biologists, due to the advent of the microarray technologies (DeRisi et al. 1997). Repositories of microarray data such as

ArrayExpress (Brazma et al. 2003) and Gene Omnibus (Barrett et al. 2007) contain hundreds of expression profiles for several model organisms, and the quantity of available data increases every day.

Many approaches have been developed for discovering cis-regulatory motifs in promoters of co-regulated genes (Hertz et al. 1990; Lawrence et al. 1993; Bailey and Elkan 1994; Neuwald et al. 1995; Brazma et al. 1998a, b; Roth et al. 1998; van Helden et al. 1998; Sinha and Tompa 2000; van Helden et al. 2000b; Liu et al. 2001; Thijs et al. 2001). In this chapter, we will focus on a single method, *oligo-analysis* (van Helden et al. 1998). A first reason is that this method was developed by one of us, and we are thus in a better position to evaluate it than other methods. Another reason is that the algorithm is fairly simple: the program *oligo-analysis* counts the occurrences of each possible oligomer in the promoters of the co-regulated genes, and applies a significance test to evaluate its level of over-representation, relative to some background model.

We will present one example of pattern discovery, and show how to evaluate the results, in terms of sensitivity and predictive value. We will then apply the evaluation to whole collections of regulons annotated for human and yeast, respectively. We will present a method for choosing optimal parameters on the basis of some global evaluation statistics. The concepts presented below can be extended to perform similar evaluations for other pattern discovery algorithm.

5.1 Example of pattern discovery result

Figure 5 shows a typical example of pattern discovery result in promoters of 12 human genes regulated by the transcription factor HNF-4. Promoter sequences were retrieved over 1.5 kb upstream from transcription start sites, in the repeat-masked version of the genome. Each set of promoters were further purged with the program REPuter (Kurtz et al. 2001), in order to mask redundant fragments. Such redundancy can occur when the data set contains recently duplicated genes, or pairs of neighbour genes transcribed in divergent directions (and thus having their promoter in the intergenic region).

The program *oligo-analysis* (van Helden et al. 1998) was used to detect over-represented hexanucleotides. Each hexanucleotide was regrouped with its reverse complement, in order to reflect the strand-insensitive activity of human cis-regulatory

Fig. 5 Examples of pattern discovery result. (A) Significant patterns detected with *oligo-analysis* in promoter sequences of 12 human genes regulated by the transcription factor HNF-4. (B) Assembly of these patterns using the program *pattern-assembly*. (C) Comparison between discovered patterns (columns) and HNF-4 binding sites annotated in TRANSFAC (rows). Perfect matches (6 residues) are highlighted in bold, with a green background. Single-mismatches (5 matching residues) are displayed with cyan background. Only the perfect matches are considered as valid for the evaluation

elements. The number of occurrences of each hexanucleotide was counted, and compared to the expected number of occurrences, estimated on the basis of a background model built from the whole set of human promoters.

The primary result of *oligo-analysis* is a set of 20 hexanucleotides that passed the binomial significance test with an E-value ≤ 1 (Fig. 5A). These 20 hexanucleotides can be assembled to form larger motifs (Fig. 5B), suggesting that they reveal different fragments of the binding motif recognized by some transcription factor.

5.2 Evaluation statistics

How can we evaluate a pattern discovery result consisting in a set of partly overlapping oligonucleotides? A simple way to treat this question is to compare each significant oligonucleotide with each annotated binding site. The result of such a comparison is shown in Fig. 5C, where each row corresponds to one HNF-4 binding site annotated in the TRANSFAC database (Wingender 2004; Wingender et al. 1996), and each column represents one hexanucleotides detected by *oligo-analysis*. The values indicate the number of matching residues between a site and a discovered oligomer (the comparison was made on both strands, and the hexanucleotides were slid along the sites to test all possible alignments). For the evaluation, we will only take into consideration the perfect matches, i.e. the correspondences over 6 base pairs (highlighted in bold in Fig. 5C).

We define the following types of correspondences.

- An annotated site is considered as a *True Positive Site* (TPS) if it is matched by at least one discovered motif, and as *False Negative Site* (FNS) otherwise.
- A discovered oligonucleotide is considered as a *True Positive Motif* (TPM) if it matches at least one annotated site, and as *False Positive Motif* (FPM) otherwise. For the same reasons as discussed above, it makes not much sense to think about true negative sites (this would be the whole genome except the annotated site) or true negative motifs (all the hexanucleotides that were not reported as significant), because these TN numbers are so high, compared to the TP, FN and FP, that all the derived statistics would become un-informative.

From these correspondence values, we derive the following statistics.

The site sensitivity (S_n) is the fraction of sites matched by at least one discovered motif.

$$S_n = \frac{\text{discovered sites}}{\text{annotated sites}} = \frac{\text{TPS}}{\text{TPS} + \text{FNS}}$$

In our study case (Fig. 5C), 12 of the 14 annotated sites are matched by at least one hexanucleotides, the sensitivity is thus $S_n = 12/14 = 0.86$.

The motif predictive positive value (PPV) is the fraction of motifs that match at least one site.

$$\text{PPV} = \frac{\text{discovered motifs}}{\text{correct motifs}} = \frac{\text{TPM}}{\text{TPM} + \text{FPM}}$$

For the HNF-4 example, we detected 20 significant oligonucleotides, among which 11 match at least one annotated site. We thus have a $\text{PPV} = 11/20 = 0.55$.

The geometric accuracy is the geometric mean between S_n and PPV, as defined above.

$$\text{Acc}_g = \sqrt{S_n \cdot \text{PPV}}$$

With the HNF-4 example, we obtain $\text{Acc}_g = \sqrt{0.86 \cdot 0.55} = 0.69$. The HNF-4 example thus gives a very good result: almost all the annotated sites are covered by the discovered motifs, and more than half of the significant hexanucleotides correspond to at least one site. This example is obviously not representative for all the human regulons. In order to gain an idea about the general performances of the program, we will now apply the same analysis for all the annotated regulons, in yeast and in human.

5.3 Correctness of predicted motifs for a collection of annotated regulons

In order to assess the general performances of the algorithm, we selected two model organisms: yeast (*Saccharomyces cerevisiae*) and human (*Homo sapiens*). The yeast *Saccharomyces cerevisiae* is currently the pet organism of many bioinformaticians, because its genome is relatively compact (12 Mb), its regulatory regions are restricted to a relatively short region (~ 500 bp) upstream of the genes, and there are several large-scale data sets available (genome, transcriptome, proteome, interactome, ...). On the contrary, *Homo sapiens* is probably the most challenging organism for the study of regulation, because of the huge size of its genome (3 Gb), the large distances between regulatory elements and their target genes (sometimes several Mb), and the fact that those signals are drawn in an ocean of non-coding sequences. In addition, 40% of the genome is covered by repetitive elements, which causes specific problems for the definition of background models.

We applied *oligo-analysis* to discover over-represented motifs in promoters of co-regulated genes obtained from the TRANSFAC database (Wingender 2004; Wingender et al. 1996). The yeast regulons obtained from TRANSFAC and complemented with annotations were taken for one of our previous publication (Simonis et al. 2004).

For each regulon, we detected over-represented hexanucleotides, compared discovered motifs with annotated sites, and computed S_n , PPV and Acc_g as described in the previous section. In addition, for the yeast promoters, we also tested another algorithm, called *dyad-analysis*, which detects spaced pairs of trinucleotides (van Helden et al.

Table 3 Evaluation of the correctness of motifs discovered with *oligo-analysis* in all the yeast regulons having at least 5 annotated target genes

Regulon	Genes	Sites	Max. Sig.	Matched sites	Significant patterns	Matching patterns	PPV	Sn	Acc _g	Rank
GLN3	31	1	24.32	1	11	1	0.09	1.00	0.30	1
GCN4	40	18	22.47	11	8	6	0.75	0.61	0.68	2
DAL80	19	2	20.79	2	15	4	0.27	1.00	0.52	3
BAS1	17	2	15.95	2	7	3	0.43	1.00	0.65	4
MSN2	56	1	14.7	1	28	8	0.29	1.00	0.53	5
PHO2	21	5	14.43	1	6	1	0.17	0.20	0.18	6
PDR1	16	9	14.24	9	17	11	0.65	1.00	0.80	7
MSN4	58	3	13.33	1	24	8	0.33	0.33	0.33	8
CBF1	16	1	9.56	1	6	1	0.17	1.00	0.41	9
PDR3	10	8	9.2	7	10	9	0.90	0.88	0.89	10
INO2	19	3	8.39	2	7	3	0.43	0.67	0.53	11
INO4	19	1	8.39	0	7	0	0.00	0.00	0.00	12
MET4	10	1	8.17	1	8	2	0.25	1.00	0.50	13
SWI6	10	3	8.09	3	10	2	0.20	1.00	0.45	14
HSF1	21	4	7.27	2	6	2	0.33	0.50	0.41	15
MIG1	26	15	6.26	12	22	7	0.32	0.80	0.50	16
DAL81	10	2	5.98	0	4	0	0.00	0.00	0.00	17
DAL82	6	2	5.61	0	2	0	0.00	0.00	0.00	18
RPN4	11	2	4.92	1	9	4	0.44	0.50	0.47	19
UPC2	9	1	4.24	1	8	1	0.13	1.00	0.35	20
LEU3	8	5	4.2	4	4	3	0.75	0.80	0.77	21
STE12	13	5	3.88	5	3	2	0.67	1.00	0.82	22
SWI4	8	3	3.76	3	9	4	0.44	1.00	0.67	23
ADR1	11	4	3.75	3	7	2	0.29	0.75	0.46	24
RCS1	9	8	2.95	8	3	3	1.00	1.00	1.00	25
REB1	19	10	2.7	7	4	3	0.75	0.70	0.72	26
ACE2	5	2	2.6	2	6	4	0.67	1.00	0.82	27
SWI5	8	4	2.57	4	8	4	0.50	1.00	0.71	28
NDT80	11	1	2.53	1	5	2	0.40	1.00	0.63	29
PHO4	8	6	2.52	6	6	5	0.83	1.00	0.91	30
UME6	26	4	2.47	2	5	5	1.00	0.50	0.71	31
MET31	5	2	2.41	2	4	3	0.75	1.00	0.87	32
MOT3	15	2	2.09	0	8	0	0.00	0.00	0.00	33
RAP1	32	20	1.92	7	3	2	0.67	0.35	0.48	34
CAT8	10	18	1.66	5	4	2	0.50	0.28	0.37	35
ZAP1	52	8	1.51	5	5	5	1.00	0.63	0.79	36
RFX1	5	9	1.42	0	6	0	0.00	0.00	0.00	37
YAP1	31	2	1.35	0	3	0	0.00	0.00	0.00	38
HAC1	7	6	1.11	2	6	2	0.33	0.33	0.33	39
XBP1	5	13	1.02	2	3	1	0.33	0.15	0.23	40
ROX1	15	25	0.89	20	2	2	1.00	0.80	0.89	41
MAC1	9	5	0.82	0	1	0	0.00	0.00	0.00	42

(continued)

Table 3 (Continued)

Regulon	Genes	Sites	Max. Sig.	Matched sites	Significant patterns	Matching patterns	PPV	Sn	Acc _g	Rank
GCR1	18	11	0.76	0	2	0	0.00	0.00	0.00	43
HAP3	15	2	0.65	1	4	1	0.25	0.50	0.35	44
HAP2	14	2	0.45	1	2	1	0.50	0.50	0.50	45
HAP4	14	1	0.28	0	3	0	0.00	0.00	0.00	46
PIP2	19	1	0.17	0	1	0	0.00	0.00	0.00	47
GAL4	9	16	0.12	0	1	0	0.00	0.00	0.00	48
ABF1	37	25	0	0	0	0	0.00	0.00	0.00	49
HAP1	7	6	0	0	0	0	0.00	0.00	0.00	50
MCM1	14	24	0	0	0	0	0.00	0.00	0.00	51
NRG1	5	2	0	0	0	0	0.00	0.00	0.00	52
SKO1	11	3	0	0	0	0	0.00	0.00	0.00	53
Average	16.98	6.40	5.26	2.79	6.28	2.43	0.35	0.52	0.41	
Median	14.00	4.00	2.60	1.00	5.00	2.00	0.32	0.50	0.45	

2000b). The results are summarized in Table 3 (*oligo-analysis* in yeast promoters), Table 4 (*dyad-analysis* in yeast promoters) and Table 5 (*oligo-analysis* in human promoters), respectively. For the yeast, *oligo-analysis* gives generally good performances, with a relatively good sensitivity (mean over all yeast regulons = 52%), a lower PPV (mean = 35%), and a geometric accuracy of 41%.

The analysis of failures further reveals interesting cases:

1. For the MOT3 regulon, the annotated binding sites are CAGGCA and AGGCAA. The most significant motif, CAGGAAA, matches these two sites with a single mismatch. In addition, a second motif is detected, AGGCACG, which matches the first site but misses its first residue. The predictions are thus likely to be correct, but to reveal a different variant (CAGGAAA) or a shifted fragment (AGGCACG) of the annotated motif. This highlights the difficulty to evaluate regulons for which only two sites are available.
2. For the regulons DAL81 and DAL82, the program fails to report the annotated motifs, but detects, with a strong significance, the GATA-box (GATAAG), which is bound by the other factors involved in nitrogen regulation (DAL80, GLN3). These factors interact with DAL81 and DAL82 for the regulation of its target genes. The detected motifs are thus not properly speaking “false positives”, even though they are considered as such for the evaluation.
3. For the regulons GAL4 and ABF1, the motifs are spaced dyads. These motifs are missed by the program *oligo-analysis*, but detected with a high significance by *dyad-analysis* (Table 4) However, some other motifs, which were detected with a weak significance by *oligo-analysis*, are lost by *dyad-analysis*, so that the mean performances over all the yeast regulons are almost identical (Sn = 55%, PPV = 38%, Acc_g = 43%).

Table 4 Evaluation of the correctness of motifs discovered with *dyad-analysis* (van Helden et al. 2000b) in all the yeast regulons having at least 5 annotated target genes

# Fam	Members	Sites	Sig. Max.	TPsites	nb_pat	TP_pat	PPV	Sn	Acc _g	Rank
GLN3	31	1	22.74	1	12	1	0.08	1.00	0.29	1
GCN4	40	18	21.12	8	9	6	0.67	0.44	0.54	2
DAL80	19	2	19.26	2	12	3	0.25	1.00	0.50	3
PDR1	16	9	17.24	9	24	20	0.83	1.00	0.91	4
ZAP1	52	8	15.28	8	14	12	0.86	1.00	0.93	5
BAS1	17	2	14.62	2	5	2	0.40	1.00	0.63	6
MSN2	56	1	13.27	1	67	7	0.10	1.00	0.32	7
PHO2	21	5	13.1	1	5	1	0.20	0.20	0.20	8
PDR3	10	8	12.44	8	22	19	0.86	1.00	0.93	9
MSN4	58	3	11.91	1	61	7	0.11	0.33	0.20	10
CBF1	16	1	8.25	1	7	1	0.14	1.00	0.38	11
CAT8	10	18	7.18	16	11	7	0.64	0.89	0.75	12
MET4	10	1	6.85	1	6	3	0.50	1.00	0.71	13
GAL4	9	16	6.84	8	7	5	0.71	0.50	0.60	14
INO2	19	3	6.83	3	7	6	0.86	1.00	0.93	15
INO4	19	1	6.83	1	7	3	0.43	1.00	0.65	16
SWI6	10	3	6.3	3	7	1	0.14	1.00	0.38	17
HSF1	21	4	5.94	3	8	6	0.75	0.75	0.75	18
PIP2	19	1	5.79	1	5	3	0.60	1.00	0.77	19
LEU3	8	5	5.36	5	12	11	0.92	1.00	0.96	20
MIG1	26	15	4.85	14	33	10	0.30	0.93	0.53	21
DAL81	10	2	4.56	0	3	0	0.00	0.00	0.00	22
DAL82	6	2	4.23	2	6	2	0.33	1.00	0.58	23
MOT3	15	2	3.63	0	8	0	0.00	0.00	0.00	24
RPN4	11	2	3.58	1	8	4	0.50	0.50	0.50	25
SWI4	8	3	3.14	3	6	4	0.67	1.00	0.82	26
ABF1	37	25	3.1	18	2	2	1.00	0.72	0.85	27
UPC2	9	1	2.91	1	6	1	0.17	1.00	0.41	28
STE12	13	5	2.56	5	3	2	0.67	1.00	0.82	29
ADR1	11	4	2.44	2	4	2	0.50	0.50	0.50	30
HAP1	7	6	2.1	1	4	2	0.50	0.17	0.29	31
GCR1	18	11	1.83	0	3	0	0.00	0.00	0.00	32
RCS1	9	8	1.63	7	4	3	0.75	0.88	0.81	33
SWI5	8	4	1.45	4	6	5	0.83	1.00	0.91	34
REB1	19	10	1.35	5	1	1	1.00	0.50	0.71	35
ACE2	5	2	1.33	2	7	5	0.71	1.00	0.85	36
PHO4	8	6	1.19	0	2	0	0.00	0.00	0.00	37
NDT80	11	1	1.14	1	3	1	0.33	1.00	0.58	38
UME6	26	4	1.14	2	7	4	0.57	0.50	0.53	39
MET31	5	2	1.09	1	1	1	1.00	0.50	0.71	40
YAP1	31	2	1.07	0	2	0	0.00	0.00	0.00	41
XBP1	5	13	0.96	0	2	0	0.00	0.00	0.00	42
RFX1	5	9	0.53	0	4	0	0.00	0.00	0.00	43

(continued)

Table 4 (Continued)

# Fam	Members	Sites	Sig. Max.	TPsites	nb_pat	TP_pat	PPV	Sn	Acc _g	Rank
HAC1	7	6	0.49	0	1	0	0.00	0.00	0.00	44
MAC1	9	5	0.48	0	1	0	0.00	0.00	0.00	45
RAP1	32	20	0.48	0	2	0	0.00	0.00	0.00	46
MCM1	14	24	0.29	0	1	0	0.00	0.00	0.00	47
ROX1	15	25	0.27	0	3	0	0.00	0.00	0.00	48
HAP2	14	2	0	0	0	0	0.00	0.00	0.00	49
HAP3	15	2	0	0	0	0	0.00	0.00	0.00	50
HAP4	14	1	0	0	0	0	0.00	0.00	0.00	51
NRG1	5	2	0	0	0	0	0.00	0.00	0.00	52
SKO1	11	3	0	0	0	0	0.00	0.00	0.00	53
Average	16.98	6.40	5.30	2.87	8.32	3.26	0.38	0.55	0.43	
Median	14.00	4.00	3.10	1.00	5.00	2.00	0.33	0.50	0.50	

Not surprisingly, the results obtained with human regulons are much worse (Table 5): the mean performances are quite poor ($Sn = 15\%$, $PPV = 24\%$, $Acc_g = 18\%$). For about half of the regulons, *oligo-analysis* fails to detect the correct motif (this explains why the median values are almost null, in contrast with the mean values). These poor performances are not specific of this program, but to come from the intrinsic difficulty of extracting motifs from promoters in vertebrate. The main problem comes from the fact that our analysis is restricted to proximal promoters, whereas many vertebrate factors regulate their target genes at distance, *via* binding sites located further away upstream, or within introns, or even downstream of the target genes. Multi-genome approaches where the analysis is restricted to conserved genomic fragments, are likely to improve the predictions, but will not be discussed in the scope of this chapter. To our knowledge, such approaches have been tested on a restricted number of study cases, but a systematic evaluation is still missing.

5.4 Distributions of motif scores in positive and negative testing sets

Pattern discovery programs return one or several scores associated with each predicted motif. A good scoring scheme should in principle help us to discriminate relevant from spurious motifs. In this section, we propose a protocol to assess the capability of a program to distinguish between relevant and spurious motifs, by comparing score distributions obtained with a positive and a negative control set, respectively. For this part of the evaluation, we do not attempt to evaluate whether the discovered motifs match or not the annotated ones (as has been treated in the previous section). An advantage of this protocol is thus that we deliberately avoid all the problems related to incomplete or inaccurate annotations.

Table 5 Evaluation of the correctness of motifs discovered with *oligo-analysis* in all the human regulons having at least 5 annotated target genes

Regulon	Genes	Sites	Max. Sig.	Matched sites	Significant patterns	Matching patterns	PPV	Sn	Acc _g	Rank
T00671_p53	28	35	8.08	10	24	14	0.58	0.29	0.41	1
T00764_SRF	7	13	5.06	3	46	3	0.07	0.23	0.12	2
T00759_Sp1	74	177	4.34	152	35	34	0.97	0.86	0.91	3
T01609_HIF-1	12	18	4.26	16	36	12	0.33	0.89	0.54	4
T00915_YY1	6	11	3.71	2	4	2	0.50	0.18	0.30	5
T00149_COUP-TF1	7	12	3.08	5	10	4	0.40	0.42	0.41	6
T00250_Elk-1	5	10	2.81	0	22	0	0.00	0.00	0.00	7
T00423_IRF-1	10	21	2.67	1	5	1	0.20	0.05	0.10	8
T02758_HNF-4	12	14	2.62	12	20	11	0.55	0.86	0.69	9
T00590_NF-kappaB	17	25	2.55	2	14	5	0.36	0.08	0.17	10
T00874_USF1	12	15	2.43	1	9	1	0.11	0.07	0.09	11
T00140_c-Myc	10	17	2.39	0	2	0	0.00	0.00	0.00	12
T00035_AP-2alphaA	17	27	2.15	7	10	7	0.70	0.26	0.43	13
T00045_COUP-TF2	6	8	1.82	3	18	3	0.17	0.38	0.25	14
T01542_E2F-1	6	12	1.68	2	6	4	0.67	0.17	0.33	15
T00593_NF-kappaB1	11	18	1.61	0	6	0	0.00	0.00	0.00	16
T00261_ER-alpha	13	21	1.39	3	2	2	1.00	0.14	0.38	17
T04096_Smad3	5	7	1.39	1	7	1	0.14	0.14	0.14	18
T00167_ATF-2	16	18	1.39	0	7	0	0.00	0.00	0.00	19
T06124_NRSF	5	6	1.36	0	3	0	0.00	0.00	0.00	20
T01950_HNF-1B	7	9	1.3	0	3	0	0.00	0.00	0.00	21
T00029_AP-1	35	53	1.27	25	3	2	0.67	0.47	0.56	22
T02068_PU.1	5	8	1.26	4	3	2	0.67	0.50	0.58	23
T00163_CREB	16	26	1.13	1	6	1	0.17	0.04	0.08	24
T01948_NF-AT1	5	11	1.02	3	6	2	0.33	0.27	0.30	25
T00221_E2F	8	17	1	9	3	2	0.67	0.53	0.59	26
T00308_GATA-2	5	5	0.98	0	2	0	0.00	0.00	0.00	27
T02338_Sp3	6	10	0.96	1	5	1	0.20	0.10	0.14	28
T01945_NF-AT2	6	7	0.94	0	1	0	0.00	0.00	0.00	29
T00112_c-Ets-1	8	16	0.88	0	2	0	0.00	0.00	0.00	30
T01580_STAT6	5	7	0.87	0	4	0	0.00	0.00	0.00	31
T01951_HNF-1C	6	7	0.86	0	2	0	0.00	0.00	0.00	32
T00721_RAR-beta	5	6	0.8	1	2	1	0.50	0.17	0.29	33
T03828_HNF-4alpha	7	11	0.73	3	2	1	0.50	0.27	0.37	34
T01616_RBP-Jkappa	5	5	0.7	0	3	0	0.00	0.00	0.00	35
T00133_c-Jun	24	32	0.59	0	4	0	0.00	0.00	0.00	36
T00539_NF-1	5	8	0.52	0	2	0	0.00	0.00	0.00	37
T00306_GATA-1	6	50	0.48	0	2	0	0.00	0.00	0.00	38
T00641_POU2F1	9	23	0.38	0	1	0	0.00	0.00	0.00	39
T01345_RXR-alpha	9	10	0.37	1	1	1	1.00	0.10	0.32	40
T00241_Egr-1	5	9	0.29	0	1	0	0.00	0.00	0.00	41
T00878_USF2	5	5	0.26	0	3	0	0.00	0.00	0.00	42

(continued)

Table 5 (Continued)

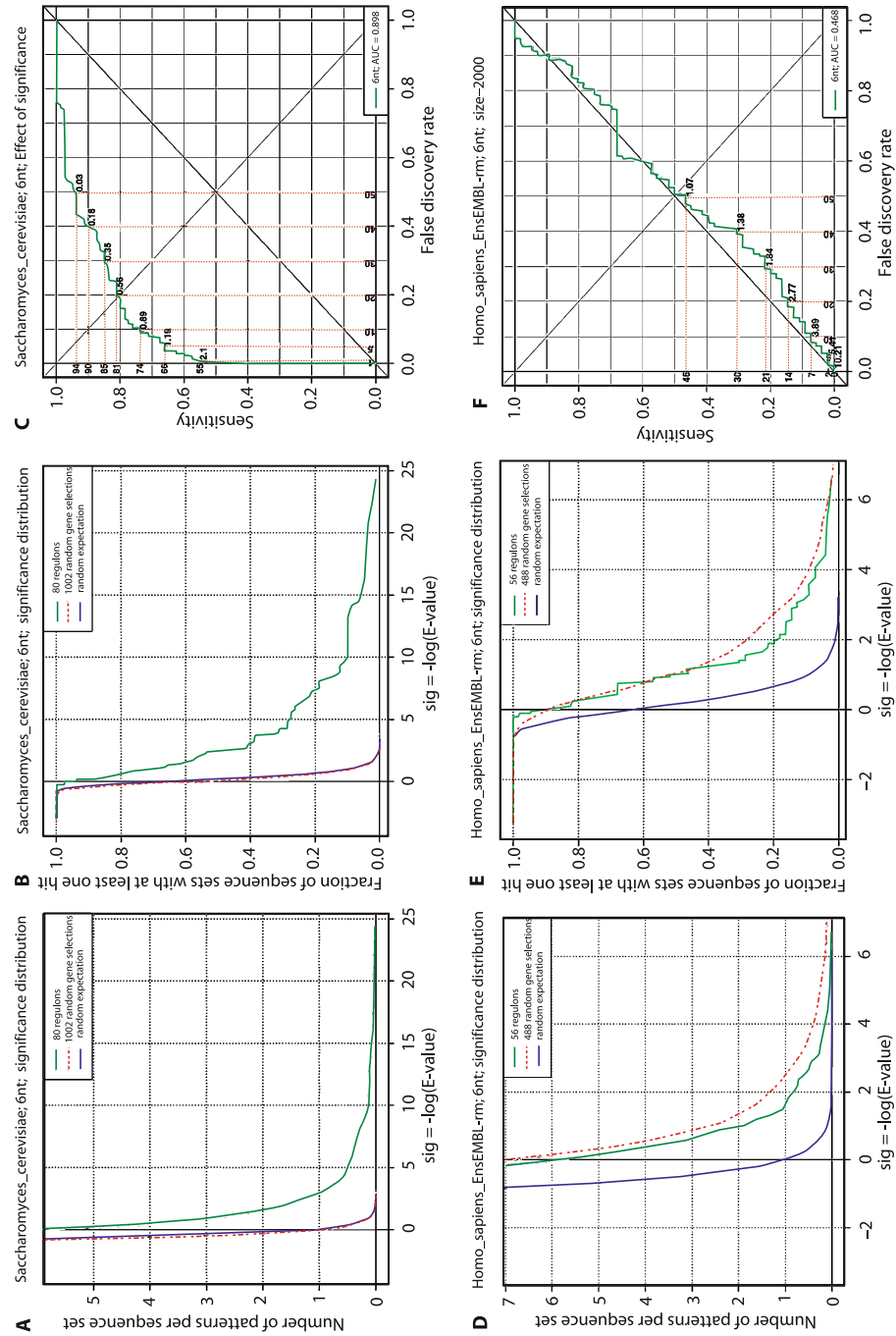
Regulon	Genes	Sites	Max. Sig.	Matched sites	Significant patterns	Matching patterns	PPV	Sn	Acc _g	Rank
T00368_HNF-1A	15	20	0.22	3	4	2	0.50	0.15	0.27	43
T00040_AR	5	15	0.21	0	1	0	0.00	0.00	0.00	44
T00594_RelA	12	16	0.19	4	2	1	0.50	0.25	0.35	45
T00113_c-Ets-2	7	9	0	0	0	0	0.00	0.00	0.00	46
T00123_c-Fos	10	13	0	0	0	0	0.00	0.00	0.00	47
T01462_Fra-1	5	6	0	0	0	0	0.00	0.00	0.00	48
T01553_MITF	7	11	0	0	0	0	0.00	0.00	0.00	49
T01977_JunB	5	6	0	0	0	0	0.00	0.00	0.00	50
T01978_JunD	6	7	0	0	0	0	0.00	0.00	0.00	51
T04759_STAT1	7	8	0	0	0	0	0.00	0.00	0.00	52
Mean	10.6	17.9	1.4	5.3	6.8	2.3	0.24	0.15	0.18	
Median	7.0	11.5	1.0	0.5	3.0	0.5	0.03	0.02	0.04	

One difficulty of the score-based assessment is that existing programs return various scores to qualify the discovered motifs. Matrix-based pattern discovery programs return one or several among the following scores : MAP (*gibbs*, *AlignACE*), Log-likelihood (*MEME*, *MotifSampler*), Information Content (*consensus*, *MotifSampler*), Consensus Score (*MotifSampler*), *P*-value (*consensus*), *E*-value (*consensus*, *MEME*). String-based pattern discovery tools also assign various scores to oligomers (words, dyads, or degenerate motifs): expected/observed ratio, *z*-score (*YMF*, *oligo-analysis*, *dyad-analysis*, *Trawler*), binomial significance (*oligo-analysis*, *dyad-analysis*, *SPATT*), compound Poisson significance (*Hermes*). We will show that the analysis of score distributions permits us to select the most discriminating score, in order to obtain the best of each program.

The score-based assessment relies on two complementary datasets:

1. The positive set is made of a collection of annotated regulons. We selected all regulons having at least 5 genes in our collection (80 regulons for yeast, 50 regulons for the human).
2. The negative set consists in sets of genes randomly selected in the genome of interest (yeast or human, respectively). These genes are probably regulated individually, but since they are regrouped on a random basis, we do not expect to find over-represented motifs in the sets of promoters analyzed for this negative test.

The distribution of scores obtained in the negative and positive sets are displayed on Fig. 6, for yeast (A–C) and human (D–F), respectively. Figure 6A shows the number of patterns returned per gene set (regulon or random selection), as a function of the significance score (abscissa). This significance score is a minus-log transform of the *E*-value: $\text{sig} = -\log_{10}(\text{E-value})$. The random selections (dotted line) perfectly follow



the theoretical curve estimated from the binomial distribution. This means that the E -value returned by *oligo-analysis* provides a reliable estimation of the rate of false positives. In the positive control (green curve), the program returns a significantly larger number of patterns than expected by chance. The most significant patterns indeed generally correspond to annotated motifs (Table 3).

Figure 6B gives a complementary information on the same result set: instead of plotting the number of hexanucleotides returned per sequence set, we show the number of sequence sets (regulons or random selections) for which at least one hexanucleotide is returned (ordinate), as a function of the significance score (abscissa). Here as well, the negative set (random gene selections) perfectly follows the theoretical curve (blue), whereas motifs are found in a much higher fraction of the positive set (regulons).

In human promoters, the distribution curves show a very bad behaviour: the number of motifs is as high in random gene selections as in regulons (Fig. 6D, E). This reflects a problem with the background model: the E -value estimated from the binomial statistics (blue curve) strongly under-estimates the empirical rate of false positives (dotted lines). This explains the poor PPV obtained for human regulons in the analysis of motif correctness (Table 5).

5.5 The Receiver Operating Characteristics (ROC) curve

Another very expressive way to display the results is to directly compare the two distributions (regulons and random selections) on a ROC curve (Fig. 6C, F). ROC (Receiver Operating Characteristics) curves were defined in the field of signal detection theory, as a plot where the X axis shows the false positive rate, and the Y axis the sensitivity. The *false positive rate* measures the fraction of negative elements that are erroneously considered as positive: $FPR = FP / (FP + TN)$. In our case, this is the fraction of random gene selections for which at least one pattern is returned above a given significance score. The *sensitivity* (S_n , also called *True Positive Rate*) is defined as above, as the fraction of positive elements that are correctly detected: $S_n = TP / (TP + FN)$. In our case, this is the fraction of regulons for which at least one pattern is returned above a given significance score. Note that the sensitivity reported here differs from that defined for the correctness analysis in Table 3, because we are not testing whether the motifs do or not correspond to the annotations. In this part of the analysis, sensitivity and FPR are defined in terms of the number of motifs reported in two data sets: genes that are co-regulated (regulons), and genes that are supposedly not (random selections). Our only



Fig. 6 Score distribution of discovered motifs in positive and negative control sets. Pattern discovery in yeast promoters (A–C) or human promoters (D–F). (A, D) number of over-represented oligonucleotides per sequence set. Green curves : number of motifs per regulon; Blue curves: expected rate of false positives; Red curve: empirical rate of false positive, estimated by measuring the number of motifs in random selections of genes (B, E) top-ranking motif for each sequence set (C, F) ROC-like curves

purpose is to assess whether a given score (the significance score for the time being) allows us to discriminate between these two data sets.

A ROC curve represents the evolution of Sn and FPR when the score varies: when the significance score increases, the total number of motif decreases which affects both sensitivity and FPR. An ideal predictor would reach 100% of sensitivity for a FPR of 0%. A random predictor would return as many motifs in the negative as in the positive set, and would thus follow the diagonal. Real-life predictors are usually located in the upper left triangle, between the random and the perfect predictors.

Let us analyze the ROC curve obtained with the detection of over-represented hexanucleotides in yeast promoters (Fig. 6C). We observe that, with a significance of 0.03, motifs are detected in 93% of the regulons, but also in 50% of the random gene selections. When the threshold on significance increases, both sensitivity and FPR decrease, but the FPR decreases faster. For example, if we accept a rate of 10% of false positive, the sensitivity is 74%, and with a restrictive threshold ($\text{sig} \geq 2.1$), we can reduce the FPR to 1%, but still detect motifs in 55% of the regulons.

A classical way to estimate the global performance of a ROC curve is to measure the *Area Under the Curve* (AUC). A perfect predictor has an area of 1, whereas a random predictor has an area of 0.5 (the lower right triangle of the plot). With the yeast promoters (Fig. 6D), we obtain an AUC of 0.9, which is pretty good. This is far from being the case for human promoters: the ROC curve follows the diagonal and the AUC is 0.47, suggesting that, for human sequences, *oligo-analysis* performs as badly as a random predictor.

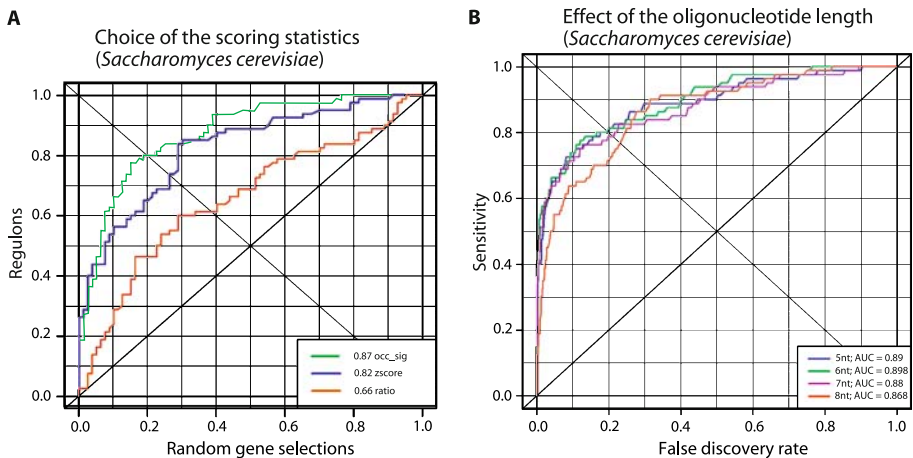


Fig. 7 Utilization of ROC curves to estimate optimal parameters for pattern discovery. (A) comparison between different scores returned by *oligo-analysis*, with yeast promoters: binomial significance (*occ_sig*), z-score, expected/observed ratio. (B) Performances of *oligo-analysis* on yeast promoters, with different oligonucleotide sizes (from 5 to 8)

5.6 Using ROC curves to find optimal parameters

It is very convenient to draw several ROC curves on the same plot, in order to compare the performances of a program under different parameter conditions, or to compare several programs (Fig. 7).

For instance, on Fig. 7A, we compare the ROC curves obtained with 3 alternative scores returned by *oligo-analysis*. The binomial significance (*occ_sig*) clearly outperforms the z-score, and the observed/expected ratio gives very poor results. We also compared the performances of *oligo-analysis* with different oligonucleotide sizes (Fig. 7B). Apparently, the global performances (with 80 regulons and 1002 random gene selections) are similar for pentanucleotides (5 nt), hexanucleotides (6 nt) and heptanucleotides (7 nt). Octanucleotides (8 nt) give slightly weaker results, as estimated from the AUC.

It is important to realize that the AUC reflects the behaviour of the curve over the whole range of FPR. In practice, we generally do not even want to consider a program that would return 50% of FPR or more. Thus, beyond the simple optimization of the AUC, we should also focus on the left side of the curve, which generally corresponds to the conditions wished for our predictions (low FPR values). In the range from 0 to 20% FPR, Fig. 7B clearly shows that octanucleotides give a lower sensitivity than shorter oligonucleotides.

6 Methodological issues for evaluating pattern discovery

In 2005, Tompa et al. (2005) organized a comparative assessment of 12 pattern discovery methods, based on sequence sets from 4 model organisms (yeast, drosophila, mouse and human). This assessment was organized as a community experiment, where each developer was invited to test his/her own program on some test sets. Developers also participated to the discussions about the evaluation statistics. Despite the huge effort put in this community-based experiment, the results of this first evaluation were rather deceiving: all the programs, without exception, had an average accuracy below 15%. This poor result contrasted with our experience, since we usually obtained higher ratings in our published and unpublished tests, at least for the yeast *Saccharomyces cerevisiae*. After this first evaluation, the organizers and some participants had the opportunity to discuss further about the strengths and weaknesses of this first assessment, and we could identify several reasons why the results were apparently so poor.

1. Most datasets only contained very few genes (less than 5 in most cases), although pattern discovery requires a sufficient number of genes in order to distinguish the signal (cis-regulatory sites) from the noise (the surrounding sequences).

2. A good part of the test sequences were artificially built by implanting binding sites extracted from a transcription factor database in some foreign sequences. The foreign sequences were either selected at random in the whole collection of promoters, or generated randomly following a Markov model. The advantage of this approach is that the positions of the correct sites are perfectly known. However, it is well known that, even in the best available databases, annotated sites only represent a fraction of those present in promoters. Thus, sequences with implanted sites contain much less instances of the regulatory motifs than native promoter sequences, so that the signal-to-noise ratio was weaker than in real conditions.
3. The results had to be submitted in the form of a set of predicted sites, which were then compared to the positions of the annotated/implanted sites. The results were thus analyzed with the same statistics as defined above for pattern matching (at the level of nucleotides and at the level of sites). However, this implied that the assessment only considered the final result of two consecutive processes: pattern discovery and pattern matching. We would like to consider those two steps separately, and to evaluate on the one hand the motifs returned by some motif discovery results, and on the other hand the binding sites predicted by scanning sequences with those motifs.

7 Good practices for evaluating predictive tools

In this section, we proposed an alternative protocol that addresses the problems mentioned above and permits to assess the results of any pattern discovery algorithms.

1. We use as testing set a complete collection of all the regulons stored in the TRANSFAC database (Wingender 2004; Wingender et al. 1996).
2. We restrict the analysis to those factors for which the database contains at least 5 target genes, in order to have a reasonable signal-to-noise ratio.
3. All sequences used in this protocol are promoter sequences retrieved from the genome, and the TFBS are in their native location.
4. The evaluation is based on the motifs returned by the pattern discovery algorithms, without requiring a further step of pattern matching. There is thus no need to scan the sequences in order to predict the binding sites. Of course, this protocol might be combined with a subsequent assessment of the sites predicted by scanning the sequences with the discovered motifs, as described in the first part of the chapter.

These methodological choices are probably not perfect, rather, we propose them as a tradeoff between the ideal situation and the constraints imposed by the available data. There are certainly alternative ways to perform such evaluations, and other statistics can be used for the same purpose. We would however like to insist on some aspects that go

beyond the choice of a precise statistics, and belong to what could be called “*good practices for evaluation*”.

7.1 Use comprehensive data sets

In too many cases, the performances of a new published method are illustrated on the basis of a few selected examples. Such selections can be justified by didactic purposes, but should never be considered as an evaluation. A quantitative evaluation should rely on an exhaustive data set, which was selected before the evaluation. The regulons should never be selected *a posteriori*. This means that we also have to report negative results and the cases for which our program fails, because they are an essential part of the evaluation. Besides, a detailed analysis of the resisting cases is often the key to an improvement of the methods.

7.2 Think about your negative control

Think about your negative control. An essential quality of a predictive method is its capability to return a negative answer when there is nothing to be predicted. Some programs have been conceived to optimize sensitivity, but this is generally at the cost of specificity. A simple negative test is to generate random sequences and submit them to the program for pattern detection (matching or discovery). A well-tuned program should report no motif (pattern discovery) or no site (for pattern matching). Such a simple test is however too optimistic, because generating random sequences relies on some theoretical background model (Bernoulli, Markov), which might be too simple to reflect the complexity of biological sequences (especially for vertebrate genomic regions). A more stringent test is to select random fragments in the genome of interest, and to submit those biological sequences to the program. Of course, these sequences will contain instances of some transcription factor binding motifs. However, since they were selected at random, they should not contain any enrichment for a particular transcription factor. The number of sites or motifs reported should thus be much lower than in promoters of co-regulated genes.

7.3 Ensure neutrality

In this chapter, we only evaluated two methods developed by our group (*matrix-scan* and *oligo-analysis*). An obvious question is “*how do these methods compare with those developed by other people*”? A comparison with other programs is often requested by referees when we submit a paper describing a new method. Some journals consider it as an editorial requirement for accepting a methods paper. However, it seems obvious to us that we are not in the best position to answer this question, for two reasons: (1) our judgment may be flawed by our motivation (we generally want to show that our

program works better than its competitors); (2) even if we adopt an attitude as neutral as possible, we generally know much better how to fine-tune the parameters for our own program than for those developed by third-parties.

We envisage two types of evaluations that can ensure neutrality: *user-based* and *community-based* assessment. In a community-based assessment, a series of testing sets are published, and each developer is allowed to test his/her methods, and send the results to the evaluation committee. This method ensures that each tool is fine-tuned in the best way, with the expertise of its own developer. On the contrary, in a *user-based* assessment, the assessor has not contributed to develop any of the programs tested. Of course, this kind of assessment does not guarantee the best utilization of each program, since a naïve user is likely to be more familiar with some programs than with other ones. Such user-based evaluations integrate several inter-dependent aspects such as the performances of the program itself, its ease of use, the complexity of its parameters, and the quality of the documentation. Somehow, the user-based assessment is more realistic, since it estimates the level of performances that other naïve users can hope to obtain from the programs.

8 What has not been covered in this chapter

In the scope of this book chapter, it is not possible to entirely cover all the aspects of the evaluation of cis-regulatory element prediction. We restricted the presentation to the evaluation of two approaches: a pattern matching method relying on position-specific scoring matrices, and a pattern discovery method based on the detection of over-represented oligomers.

We did not attempt to cover matrix-based methods for pattern discovery, such as consensus (Hertz et al. 1990), MEME (Bailey and Elkan 1994) or the Gibbs sampler (Lawrence et al. 1993; Neuwald et al. 1995; Roth et al. 1998; Thijs et al. 2001). For such programs, the analysis of score distributions can be performed exactly in the same way as we did for *oligo-analysis*. Such an evaluation is already very informative regarding the discriminative power of the various scores returned by these motif discovery programs: information content, consensus score, log-likelihood, MAD, . . . (results not shown).

Another important aspect that is not covered here is the application of comparative genomics to detect conserved elements in the promoter sequences (phylogenetic footprints). Our group recently published a systematic evaluation of footprint discovery in Bacteria (Janky and van Helden 2008), using the same concepts as described here.

We did not either treat the evaluation of programs that predict cis-regulatory modules (CRM), i.e. genomic regions enriched in binding sites for one transcription factor (homotypic models) or several transcription factors (heterotypic models). An evaluation of CRM predictions was recently performed in the fruit fly *Drosophila melanogaster*, using either a single genome, or a combination of 12 genomes of the same genus (Aerts et al. 2007).

9 Materials

To analyze the data and draw the figures of this chapter, we used various programs of the Regulatory Sequence Analysis Tools (RSAT), a specialized software suite for the detection of cis-acting elements in genome sequences (van Helden 2003; van Helden et al. 2000a). These tools can be accessed via a web interface (<http://rsat.bigre.ulb.a.be/rsat/>), or used as Web services (the latter requires some programming skills).

The RSAT project started in 1997 in the Centro de Investigacion de la Fijacion de Nitrogeno (Cuernavaca, Mexico), and has been pursued since 1998 at the Université Libre de Bruxelles (Belgium). Since 2004, the BioSapiens project contributed to its funding, in particular for the assessment of pattern discovery in various model organisms, and for the development of Web services.

Abbreviations

Acc _g	Geometric accuracy
AUC	Area Under the Curve (under the ROC curve)
FN	False Negative
FP	False Positive
FPR	False Positive Rate
PPV	Positive Predictive Value
PSSM	Position-specific scoring matrix
ROC	Receiver Operating Characteristics
Sn	Sensitivity
TF	Transcription factor
TFBS	Transcription factor binding site
TN	True Negative
TP	True Positive

Acknowledgements

This project was funded by the BioSapiens Network of Excellence funded under the sixth Framework program of the European Communities (LSHG-CT-2003-503265). JVT is supported by a doctoral grant from the Fonds pour la Recherche dans l'Industrie et l'Agriculture (F.R.I.A.) allowed by the Belgian Fonds National de la Recherche Scientifique.

JvH acknowledges Martin Tompa for having impuled a collaborative spirit in the first community-based assessment of pattern predictions, and for several days of thought-stimulating discussions at the 2nd Barbados Workshop on Genomics and Gene Regulation.

References

- Aerts S, van Helden J, Sand O, Hassan BA (2007) Fine-Tuning Enhancer Models to Predict Transcriptional Targets across Multiple Genomes. *PLoS ONE* 2: e1115
- Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2: 28–36
- Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007) NCBI GEO: mining tens of millions of expression profiles-database and tools update. *Nucleic Acids Res* 35: D760–D765
- Brazma A, Jonassen I, Eidhammer I, Gilbert D (1998a) Approaches to the automatic discovery of patterns in biosequences. *J Comput Biol* 5: 279–305
- Brazma A, Jonassen I, Vilo J, Ukkonen E (1998b) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res* 8: 1202–1215
- Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA (2003) ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 31: 68–71
- Cavener DR (1987) Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res* 15: 1353–1361
- DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680–686
- Down TA, Hubbard TJ (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res* 33: 1445–1453
- Eskin E, Pevzner PA (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics* 18 (Suppl 1): S354–S363
- Ettwiller L, Paten B, Ramialison M, Birney E, Wittbrodt J (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat Methods* 4: 563–565
- Favorov AV, Gelfand MS, Gerasimova AV, Ravcheev DA, Mironov AA, Makeev VJ (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics* 21: 2240–2245
- Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z (2004) Detection of functional DNA motifs *via* statistical over-representation. *Nucleic Acids Res* 32: 1372–1381
- Haun RS, Dixon JE (1990) A transcriptional enhancer essential for the expression of the rat cholecystokinin gene contains a sequence identical to the –296 element of the human *c-fos* gene. *J Biol Chem* 265: 15455–15463
- Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15: 563–577
- Hertz GZ, Hartzell GW 3rd, Stormo GD (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci* 6: 81–92
- Janky R, van Helden J (2008) Evaluation of phylogenetic footprint discovery for the prediction of bacterial cis-regulatory elements. *BMC Bioinformatics* (in press)
- Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* 29: 4633–4642
- Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262: 208–214
- Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*: 127–138
- Neuwald AF, Liu JS, Lawrence CE (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* 4: 1618–1632
- Neuwald AF, Liu JS, Lipman DJ, Lawrence CE (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Res* 25: 1665–1677

- Nuel G (2005) S-SPatt: simple statistics for patterns on Markov chains. *Bioinformatics* 21: 3051–3052
- Pavesi G, Mauri G, Pesole G (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* 17(Suppl 1): S207–S214
- Pavesi G, Mereghetti P, Mauri G, Pesole G (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* 32: W199–W203
- Regnier M, Denise A (2004) Rare events and Conditional Events on random strings. *DMTCS* 2: 191–214
- Robin S, Rodolphe F, Schbath S (2005) *DNA, words and models – statistics of exceptional words*. Cambridge University Press
- Roth FP, Hughes JD, Estep PW, Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* 16: 939–945
- Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18: 6097–6100
- Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol* 188: 415–431
- Simonis N, Wodak SJ, Cohen GN, van Helden J (2004) Combining pattern discovery and discriminant analysis to predict gene co-regulation. *Bioinformatics* 20: 2370–2379
- Sinha S, Tompa M (2000) A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol* 8: 344–354
- Sinha S, Tompa M (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 31: 3586–3588
- Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17: 1113–1122
- Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 23: 137–144
- van Helden J (2003) Regulatory sequence analysis tools. *Nucleic Acids Res* 31: 3593–3596
- van Helden J, Andre B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 281: 827–842
- van Helden J, Andre B, Collado-Vides J (2000a) A web site for the computational analysis of yeast regulatory sequences. *Yeast* 16: 177–187
- van Helden J, Rios AF, Collado-Vides J (2000b) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* 28: 1808–1818
- Wingender E (2004) TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol* 4: 55–61
- Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 24: 238–241
- Wolfertstetter F, Frech K, Herrmann G, Werner T (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput Appl Biosci* 12: 71–80
- Workman CT, Stormo GD (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput* 2000: 467–478

CHAPTER 2.2

A biophysical approach to large-scale protein-DNA binding data

T. Manke, H. Roeder and M. Vingron

Max Planck Institute for Molecular Biology, Berlin, Germany

A key mechanism of gene regulation involves the binding of transcription factors to the promoter regions of their respective target genes. This mechanism has long been studied for individual promoters and specific transcription factors using a number of experimental techniques, such as DNase footprinting (Galas and Schmitz 1978), gel-shift assays (Fried and Crothers 1981) and SELEX (Tuerk and Gold 1990).

Recent experimental advances have added a genome-wide perspective to this research. Molecular biologists can now determine the relative binding strength of a transcription factor to thousands of different promoter regions with a method called chromatin immunoprecipitation followed by microarray analysis (ChIP-chip {X7E “ChIP-chip”}). In such experiments transcription factors are extracted from the cell, along with various DNA-fragments to which they are bound. The nucleotide sequence of these fragments can then be determined by hybridization of the DNA to a specially designed microarray (Lee et al. 2002). Here the key idea is to quantify the different amounts of bound fragments corresponding to different sites in the genome through the intensity of the hybridization signal. In principle, this provides a quantitative measure for differences in binding strength, but one still needs to normalize against sequence-specific background signals. Therefore two channels of intensities are considered: the red channel (R) for the signal intensity, and the green channel (G) for a factor-independent background signal. The binding ratio (R/G) provides a relative measure of binding strength.

The emergence of such data has highlighted the fact that transcription factors can bind to DNA with a range of different affinities, and it has triggered also a more quantitative approach to theoretical binding models. Prior to large-scale binding data, computational models had focused almost exclusively on the prediction of new binding sites from a limited number of observed sites. In other words, the models aimed to generalize sparse observations. In contrast, binding data is now abundant and the

Corresponding author: Thomas Manke, Max Planck Institute for Molecular Biology, Ihnestr. 73, 14195 Berlin, Germany (e-mail: manke@molgen.mpg.de)

challenge is to rationalize these observations in terms of some underlying model. Here we review recent efforts to characterize protein-DNA interactions starting from a biophysical framework.

1 Binding site predictions

The results from individual binding experiments or *de-novo* motif discovery can be represented as alignments of real or putative binding sites, as illustrated in Fig. 1 for the serum response factor (SRF). Such an alignment can also be represented by a **motif matrix** {XE “*motif matrix*”} which contains the observed nucleotide counts or frequencies at each position of the alignment. For ease of visualization, the information content at each position is often used to represent the nucleotide preferences of a transcription factor.

Such representations basically model nucleotide occurrences at different positions within the binding site as independent from each other, an assumption which is hard to falsify with limited data (Stormo 1990). One may therefore generate random instances of “binding sites” by picking nucleotides according to their position-specific distribution in the matrix. In the example above one would choose the letter G at the first position with high probability ($39/46 \approx 0.85$). Many years of detailed binding site studies in various species and for different factors have produced hundreds of such

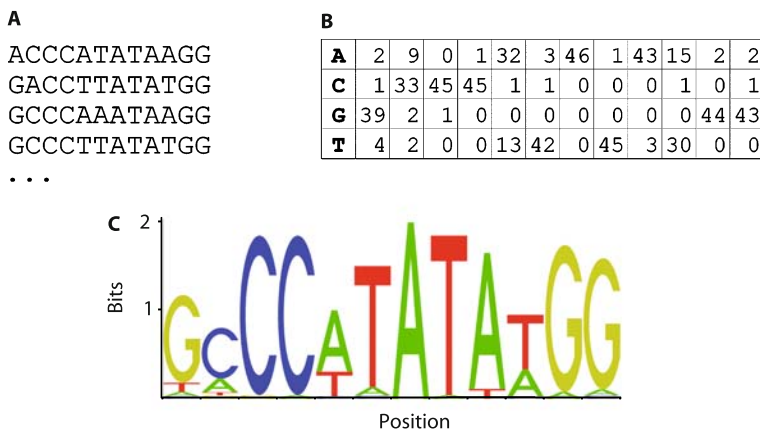


Fig. 1 Binding site preferences of SRF. A) shows part of an alignment of 46 sequences which were found to bind SRF. The motif matrix in B) counts the observed nucleotides at each of the 12 positions in the alignment. From this matrix one can determine the frequency distribution of nucleotides at each position, which is conveniently characterized by its information content. The sequence logo in C) uses this information content to determine the height at each position (≤ 2 bit) and scales each nucleotide according to their relative frequency

matrix models, which are stored in databases such as TRANSFAC (Wingender et al. 1996) or JASPAR (Sandelin et al. 2004). In a similar spirit, one can also define a probabilistic model for the sequence background which does not contain any sites. For example, one can define a random background sequence model, in which all nucleotides occur with some specified frequency and independent of their positions. If both the binding site model (provided by the motif nucleotide frequency matrix) and the background model are specified, one can assign a likelihood ratio to any arbitrary sequence site. This likelihood score quantifies whether the site is more likely to be generated by the sequence model, or by the background. This approach has frequently been reviewed in more detail (Stormo 2000; Bulyk 2005; D’haeseleer 2006). By choosing an appropriate score threshold {XE “score threshold for hit based methods”}, one defines putative binding sites which exceed the score threshold. To distinguish such predictions from biological sites, they are also referred to as *motif hits*. Importantly, any score threshold also determines an expected false positive rate (motif hits which are generated by the background model), as well as an expected rate of false negatives (sites which are generated from the motif matrix, but which score below the threshold). While the hit-based approach is statistically sound it has certain disadvantages which one would like to overcome.

First, the traditional hit-based methods discard all quantitative information and do not discriminate between strong and weak binding signals after the threshold has been put in place, see Fig. 2. This makes any follow-up analysis sensitive to the threshold. For example, such analyses often aim to quantify a degree of correlation between experimental results and computational predictions, or the correlation of binding patterns for different

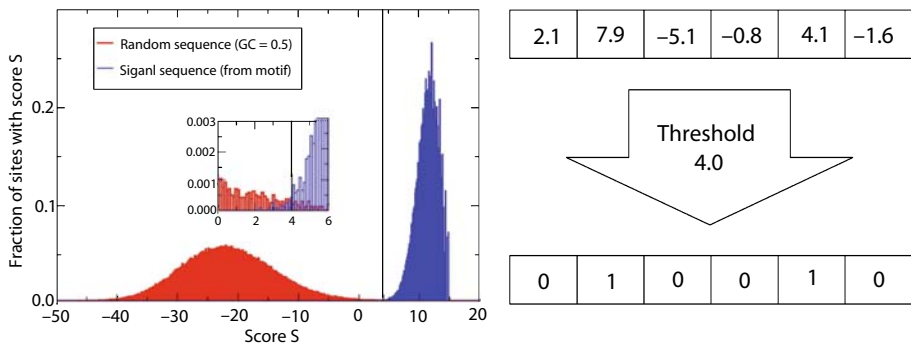


Fig. 2 Given two score distributions of background (red) and signal (blue) one can choose a particular threshold (black line) such as to fix the false positive and false negative rates. In general the two distributions overlap, as illustrated by the inset. Therefore a threshold reflects some compromise between specificity and sensitivity, with the help of which one can classify any sequence as containing the motif or not. This gives rise to *binding profiles*, as exemplified by the left figure for seven hypothetical promoter sequences. In this approach one typically disregards quantitative differences among different binding sites, i.e. the difference between the scores 4.1 and 7.9 in the example

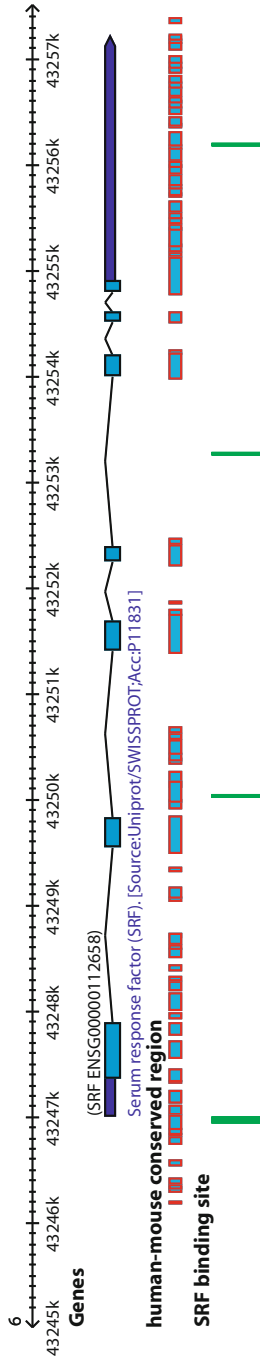


Fig. 3 Phylogenetic Footprinting of the human SRF promoter region. Evolutionary conserved sequence regions between human, mouse and chicken are highlighted as light-blue tracks and can be interpreted as regions which have evolved under tighter selective constraints, such as exons. A number of statistically significant SRF motifs are shown by the green ticks of the lowest track. The left-most motif is conserved in human and mouse and corresponds to the well-known functional binding site, which underlies the auto-regulatory capacity of SRF

transcription factors. If the binding profiles merely encodes (in a Boolean manner) whether a promoter region contains a motif hit or not, such correlation measures are not very robust against changes in the threshold. Second, it is not immediately clear how hit-based predictions can help to model quantitative binding data and how binding site predictions could be improved systematically. Third, rather than obtaining individual binding sites, current binding data frequently assigns a binding strength to longer sequence regions, and the hit-based approaches must be extended to obtain “regional scores”. Fortunately, there is a simple biophysical interpretation of binding scores, which permits calculation of a regional binding affinity. Fourth, the binary assignment of “motif hits” cements the conceptual prejudice that predicted sites correspond to functional links. In fact, the same approach is often applied to experimental data, where interactions that are stronger than some threshold are interpreted as regulatory interactions. Finally, the hit-based approach does not usually help to identify potential regulators for a given sequence region. Given the large number of known motifs, there is still a large number of statistically significant motif occurrences in any stretch of sequence. This problem can be alleviated somewhat, if one focuses on specific transcription factors of interest or invokes functional data, such as evolutionary sequence conservation or gene expression. The approach of *phylogenetic footprinting* {XE “*phylogenetic footprinting*”} is illustrated in Fig. 3 for the SRF promoter region, which shows several conserved motif occurrences for SRF, which is a known transcription factors with auto-regulatory capacity (Schratt et al. 2002; Dieterich et al. 2003).

While phylogenetic footprinting may reduce the number of false positive predictions, the restriction of search space may also miss functionally important sites (false negatives). Moreover, even conserved sequence regions tend to harbor an overabundance of possible motif occurrences for known transcription factors, and one would still like to rank them according to some rationale; their relative binding strength, for example. In the following we will focus on a new methodology to predict and compare more accurately the strength of transcription factors to sequence regions.

2 Affinity model {XE “affinity model, TRAP”}

As was pointed out before, with the emergence of large-scale binding data, the experimental situation has shifted from sparse knowledge of individual binding sites to genome-wide measurements of binding strength. Here we will review a simple biophysical model, with the help of which one can quantify the binding affinity of a transcription factor to individual sequence sites and longer sequence regions (Roeder et al. 2007). First consider many copies of some DNA site, S_l , which extends from sequence position l to $l + W - 1$. The equilibrium constant for protein-DNA complex formation at this site is given by the ratio of concentrations of the complex $[T \cdot S_l]$, with respect to the concentrations of the free transcription factor $[T]$, and the unoccupied

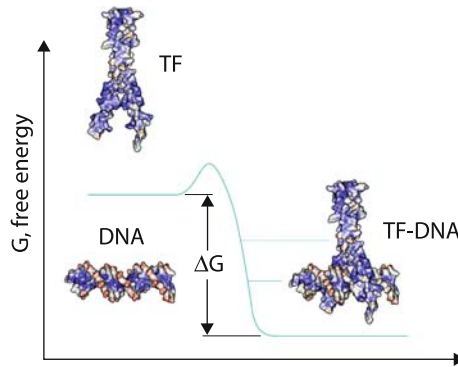


Fig. 4 This figure illustrates the energetics of protein-DNA interactions for one hypothetical site. Generally the free energy of the complex is lower than that of the unbound state, indicating that most TF will bind DNA. The difference in free energy determines the fraction of bound sequences according to Boltzmann equilibrium distribution

sequence site $[S_l]$. According to the Boltzmann formalism, this ratio is determined by the difference in free energy, $\beta\Delta G$ (Zumdahl 1998)

$$K_{\text{eq}} = \frac{[T \cdot S_l]}{[T][T_l]} = e^{-\beta\Delta G_l} \quad (1)$$

This is illustrated in Fig. 4, where the energy difference between the unbound educts and the complex is plotted in thermal units of $\beta = kT$. The correlation being the lower the binding energy of the bound state, the higher the concentration of complex at equilibrium.

The energy difference depends on the nucleotide composition of the site, and in the following we measure all energy differences with respect to the lowest possible energy, i.e. the consensus site to which we assign $E_0 = 0$.

$$a_l = \frac{[T \cdot S_l]}{[S_l][T \cdot S_l]} = \frac{R_0 e^{-\beta E_l}}{1 + R_0 e^{-\beta E_l}} \quad (2)$$

We call this fraction the local affinity, which can be assigned to any sequence site S_l . Here $R_0 = [T]e^{-\beta\Delta G_0}$ is a positive, sequence-independent parameter, and $E_l \geq 0$ is a site-dependent *mismatch energy*. Following the classical model of Berg and von Hippel (1987), the mismatch energy for many transcription factors can be calculated as independent contributions from each basepair within a sequence site of width W .

$$\beta E_l = \sum_{w=1}^W \epsilon_w(M, \lambda) \quad (3)$$

Here the individual contributions, ϵ_w depend on a known motif matrix, M , and a dimensionless scale-parameter (Roeder et al. 2007). This formalism is similar to the

score calculation, but the physical framework has the advantage that the energy model also makes prediction on the site-specific affinity as in Eq. (2). Moreover, this approach also allows to determine the affinity to longer DNA sequence regions of length L , by summing up all contributions from both strand, a_l , and anti-strand sequences, $a_{\bar{l}}$

$$A(R_0, \lambda) = \sum_{l=1}^L a_l + a_{\bar{l}} \quad (4)$$

For a more detailed exposition we refer the reader to our original work (Roeder et al. 2007), where we have also included a correction term for palindromic motifs. The above model depends on two parameters, R_0 and λ , which can be tuned for a given set of binding data. In our work we utilized the large-scale ChIP-chip experiments from the laboratory of Richard Young (Harbison et al. 2004). This group provides binding data for the relative binding affinities of more than 200 yeast transcription factors to intergenic sequences under various cellular conditions. We use the experimental ratio, R/G , of intensities for bound vs. unbound fragments and compared it to the prediction for the affinity A . In this case the analyzed sequence regions correspond to the intergenic regions in yeast ($L \approx 1000$ bp). Figure 5 provides an example for such comparison between our prediction and the experimental binding ratios for transcription factor Leu3 and some 6000 intergenic regions, which were measured in rich media condition. In this case we achieve a highly significant correlation coefficient of 0.31, when the two parameters, R_0 and λ are optimized.

Remarkably, most transcription factors gave a similar value of $\lambda \approx 0.7$, and a closer inspection of the optimal R_0 revealed that this value depends strongly on the width W of

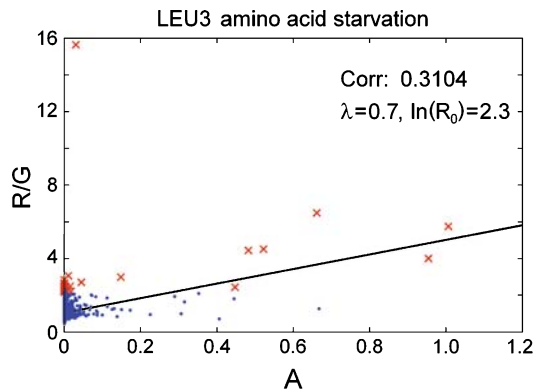


Fig. 5 The correlation of predicted affinity, A , and experimental binding ratio (R/G) for the factor Leu3. The Pearson correlation coefficient is ≈ 0.31 for the optimal choice of parameters. The sequences considered to be bound the TF according to ChIP-chip data are indicated by red crosses

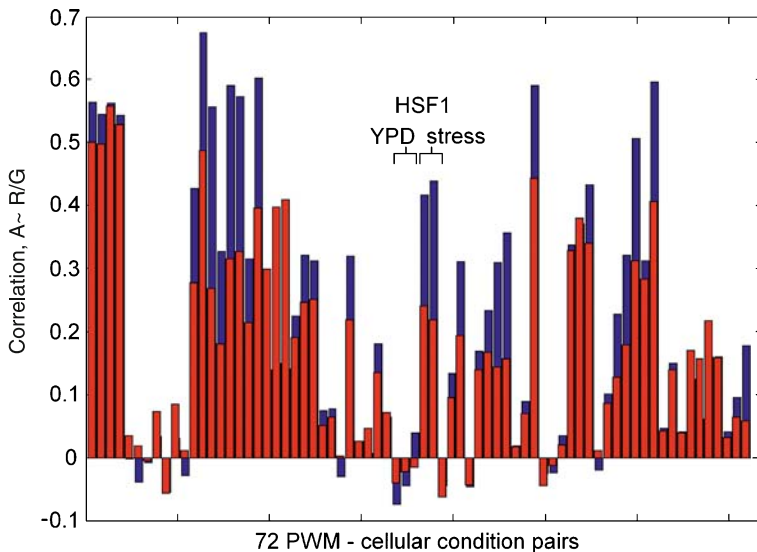


Fig. 6 Illustrates that, for the majority of cases, the affinity model (blue bars) can predict the experimental binding ratio better than the traditional method (red bars). Notice that in several cases, the binding ratios for a given factor were obtained in different cellular conditions, while the affinity predictions only rely on sequence information. They do not take into account any condition specific parameter (such as changes in the binding model, chromatin modifications, or competition with other factors). Therefore we cannot expect to observe correlations for all condition. This is exemplified by HSF1, which was measured in rich media (YPD) and under stress conditions

the motif, which in turn is a good predictor of R_0 . This defines a general model, which we termed TRAP{ XE “TRAP” } for transcription factor affinity prediction. Unlike hit-based sequence annotation, TRAP assigns a certain affinity of a given transcription factor to any sequence fragment. It avoids an arbitrary distinction between bound and unbound sequences, but it still permits a comparison and ranking of different sequences according to their predicted binding affinity to a factor of interest. Figure 6 demonstrates that for a number of transcription factors, TRAP can account for a large fraction of binding data, but not necessarily for all tested environments. For example, the heat-shock factor shows good correlation (~ 0.5) in heat-shock condition, but poor correlation (< 0.1) in rich media. This is an example where the binding ability of the factor is known to change in response to post-translational modification, and where the known motif is an appropriate model only for a specific cellular condition. Notice that the correlations obtained with the affinity-based model are almost always better than those obtained with a hit-based method. For the latter, one may simply count the numbers of hits in a given sequence region. As mentioned before, such a measure does not properly take into consideration differences in binding strength above or below the threshold.

Therefore a threshold is not necessary to model large-scale binding data, and it may even hinder a quantitative comparison between experimental data and model prediction. In contrast, the physical approach provides a quantitative framework for comparison and systematic improvement. It shifts the focus from the prediction of binding sites to the prediction of affinities, which may be weak or strong.

3 Affinity statistics {XE “affinity statistics”}

While the affinity-based model provides a ranking of sequence regions according to their different affinities for a given transcription factor, one cannot always directly compare affinities from different factors for a given sequence. This is because different transcription factors can have very different affinity distributions, as illustrated in Fig. 7. In general, transcription factors can have different binding specificities, a fact which is reflected in a shifted affinity distribution. Moreover, the affinity distributions are neither normal nor easily parameterized. This is a remnant of the discrete motif matrices, which entail discrete binding energies and therefore discrete affinities.

Therefore the theoretical challenge is to provide a proper normalization of the distribution, such that the binding affinities of different factors can be directly compared with each other. The cumulative distribution function provides such normalization, as it assigns a p -value ($0 \leq p \leq 1$) to any affinity. Using an iid background model and simplifying assumptions about the independence of nucleotide positions, one can actually derive this distribution exactly and for any matrix. This is illustrated in Fig. 8 for

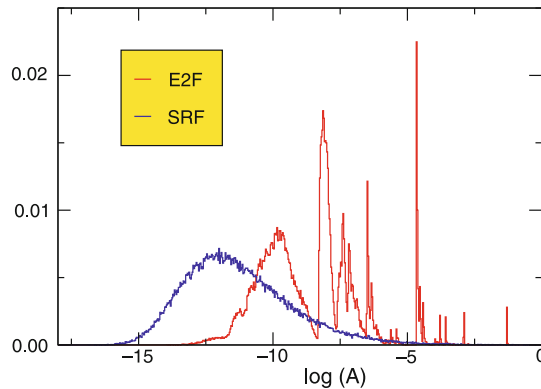


Fig. 7 Different transcription factors have different affinity distribution. In this example the affinities of two well-known transcription factors, E2F and SRF, are compared for a background of random sequences with length $L = 1000$ bp. For the background we assumed a constant GC-content (50%), but the results are very similar for human promoter regions. The latter have a similar average GC-content, but vary considerably around this average

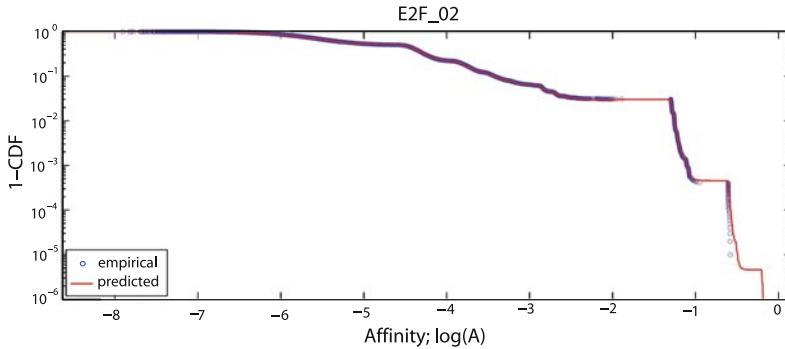


Fig. 8 The empirical distribution of affinities in a random sequence background with GC-content of 50% (blue curve) can be accurately predicted from theoretical considerations (red curve) for any given transcription factor. This illustration is for E2F and a region of length $L = 1000$ bp

the transcription factor E2F and a sequence region of $L = 1000$ bp with GC-content = 50%. One notices a perfect agreement between the predictions and an actual simulation with 100000 measurements. Again it is apparent that the tails cannot generally be described by any standard parameterization. In particular, the log-normal approximation seems to fail completely in the tails. Unfortunately this is the region of high affinities which are of most interest. One should recall here again, that the goal is not to set some significance threshold, but rather to normalize an observed affinity, and to give a statistical meaning to the statement that one factor binds stronger than another.

The precise determination of the cumulative distribution is theoretically rewarding, but not very useful for practical applications as it would have to be repeated for all possible region sizes. Also the background model is overly simplistic and does not capture the heterogeneity observed in most real promoter regions. Therefore it is highly desirable to obtain an efficient parameterization which captures the cumulative distribution at least at some level of accuracy. For the purpose of ranking different transcription factors, it will be quite sufficient to calculate an approximate p -value using a small number of parameters, rather than working with the precise p -values of Fig. 8.

For the following application we used the General Extreme Value distribution {XE “General Extreme Value distribution”}

$$P(x|a, b, c) = \exp\left(-\left[1 + a\frac{x-c}{b}\right]^{-\frac{1}{a}}\right) \quad (5)$$

which provides an estimates for the probability that a certain affinity exceed the value x . The three parameters (a, b, c), have been determined for each motif matrix by fitting the empirical affinity distribution to the parameterized form of Eq. (5). Here we also took the biologically motivated background set of < 34,000 human promoter sequences.

4 Applications

Now we show how the statistical analysis can be applied in a realistic setting. Consider again the promoter region of SRF, which we take to be a 2000 bp region centered at the transcription start site. The biologically relevant question is to decide which transcription factors are most likely to regulate the activity of SRF. The transcription factors with the highest predicted affinity include many unspecific factors which have high affinities throughout the human genome, but which are not specific to the SRF promoter. Therefore one needs the statistical approach to determine those factors, which have a high affinity *specifically* for the SRF promoter, but not the background. In Fig. 9 we show the top-ranking matrices, after the affinity has been converted into a p -value, as described above. The top-ranking matrices include many representatives of the SRF transcription factor itself, which indicates that our statistical approach to binding strength is able to rank known regulators top.

Notice that in this setting we assumed maximal ignorance and included all known transcription factor matrices and the whole sequence region. In particular we did not reduce the search space to conserved sequence regions, and we did not remove uninformative or non-vertebrate matrices. Clearly, the motif matrices are also not independent of each other, but this redundancy could be resolved in a post-processing step of this analysis. While the regulatory mechanisms acting on the SRF-promoter are likely to involve additional sequence elements and other transcription factors, it is encouraging that a key player is correctly identified by the ranking method. The ranking scheme provides a robust approach to quantify the binding strength and to discriminate transcription factors from each other. It should be stressed though, that within the

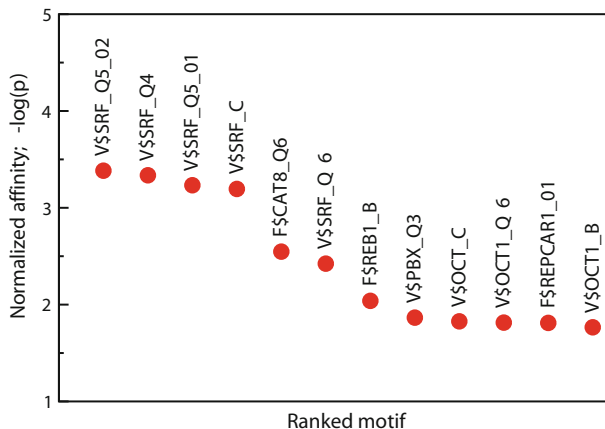


Fig. 9 Top-ranking TRANSFAC matrices according to their normalized affinity with respect to the 2000 bp promoter region of SRF

biophysical framework one does not aim to detect transcription factors with low affinity, which may be important for certain aspects of regulatory control.

5 Summary

Here we reviewed the traditional approach to transcription factor DNA binding and contrasted it with a new method that retains the quantitative information about the affinity of a transcription factor binding to DNA regions. The underlying biophysical model is still simple, but it can be easily tuned, given large-scale binding data and matrix motifs. In contrast to the traditional approach, the affinity model does not introduce thresholds, and does not predict “hits” of transcription factors, but rather their relative binding strength to a given sequence region.

More importantly the model provides a quantitative benchmark, which can be systematically improved. This illustrates that the Berg–von Hippel model (Berg and von Hippel 1987), which was developed in the context of bacterial transcription factors, still has its applicability and successes in eukaryotes.

Moreover, given a certain sequence region and a list of motif matrices, one can now meaningfully rank the corresponding transcription factors according to their specific binding strength to the selected sequence region. This is possible because a simple background model suffices to capture most of the affinity distribution in various sequence backgrounds.

While the physical model and its statistical interpretation have certain successes, there is clearly room for improvement. On the model side one should account for more complicated regulatory mechanisms, such as co-operation among transcription factors and competition with nucleosomes. Further improvements of the statistical model will likely come from a better description of the tails of the distribution, for which certain limit theorems ensure a universal behaviour, which may be parameterised more accurately.

References

- Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 193(4): 723–750
- Bulyk ML (2005) Discovering DNA regulatory elements with bacteria. *Nat Biotechnol* 23(8): 942–944
- D’Haeseleer P (2006) How does DNA sequence motif discovery work? *Nat Biotechnol* 24(8): 959–961
- Dieterich C, Wang H, Rateitschak K, Luz H, Vingron M (2003) CORG: a database for comparative regulatory genomics. *Nucleic Acids Res* 31(1): 55–57
- Fried M, Crothers DM (1981) Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res* 9(23): 6505–6525
- Galas DJ, Schmitz A (1978) DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* 5(9): 3157–3170

- Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431(7004): 99–104
- Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298(5594): 799–804
- Roider HG, Kanhere A, Manke T, Vingron M (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 23(2): 134–141
- Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32(Database issue): D91–D94
- Schratt G, Philippar U, Berger J, Schwarz H, Heidenreich O, Nordheim A (2002) Serum response factor is crucial for actin cytoskeletal organization and focal adhesion assembly in embryonic stem cells. *J Cell Biol* 156(4): 737–750
- Stormo GD (1990) Consensus patterns in DNA. *Methods Enzymol* 183: 211–221
- Stormo GD (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16(1): 16–23
- Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249(4968): 505–510
- Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 1996, 24(1): 238–241
- Zumdahl SS (1998) *Chemical principles*, 3rd edn. Houghton Mifflin Company: Boston, ISBN 0-395-83995-5

CHAPTER 2.3

From gene expression profiling to gene regulation

R. Coulson¹, T. Manke², K. Palin³, H. Roeder², O. Sand⁴, J. van Helden⁴,
E. Ukkonen³, M. Vingron² and A. Brazma¹

¹Microarray Group, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

²Max Planck Institute for Molecular Genetics Ihnestrasse 73, Berlin, Germany

³Department of Computer Science, University of Helsinki, Helsinki, Finland

⁴Laboratoire de Bioinformatique des Génomes et des Réseaux, Université Libre de Bruxelles, Bruxelles, Belgium

1 Introduction

Transcription factors, by binding to particular DNA sequences termed transcription factor-binding sites, play an important role in regulating gene expression in both prokaryotic and eukaryotic organisms. These binding sites lie within promoters (which are located just upstream of a gene and promote transcription of that gene) and enhancers (short DNA elements enhancing transcription levels of genes in a gene cluster, and which need not be particularly close to the genes they act on, or even located on the same chromosome). Binding of transcription factors in these genomic regulatory regions can influence gene transcription rates either positively or negatively. The binding may also be dependant on the interaction with co-activators and co-repressors, in addition to context (e.g. particular histone modifications in the vicinity of the regulatory element). Identifying all transcription factors and their respective binding sites would be an important step towards a more thorough understanding of gene regulation. Regular expression type patterns, as well as nucleotide distribution matrices, have both been used for describing transcription factor-binding sites, e.g. (Bucher 1990; Ghosh 1990; Chen et al. 1995; Wingender et al. 1996). Here we will discuss some of the computational approaches that are used in binding site identification.

The basis of this approach is an assumption that a set of genes tightly co-expressed under a range of conditions are likely to be co-regulated, i.e. the same transcription factor may be binding to the promoters of these genes. This means that if we can find (i) a set of tightly co expressed genes and (ii) the promoter (or enhancer) regions for

Corresponding author: Alvis Brazma, Microarray Group, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK (e-mail: brazma@ebi.ac.uk)

these genes, the statistical analysis of the specific sequence features of these regions can reveal the binding sites. The former has been facilitated by the availability of genome-wide expression data (notably microarrays), and the latter can be achieved by looking for sequences resembling known binding sites of known transcription factors in these regions, or by looking for statistically overrepresented sequence elements in the set of co-expressed promoters (or enhancers).

Algorithms have been proposed for inferring descriptions of binding sites from sets of relatively small number of sequences (about 20) in which all or almost all of the sequences are known to contain the site for the respective transcription factor e.g. (Stormo and Hartzell 1989; Wolfertstetter et al. 1996). With the ability to extract sets of co-expressed genes from transcriptomics experiments, it has been possible to identify novel putative binding sites for yeast transcription factors (Brazma et al. 1998b; van Helden et al. 1998). However, this approach has turned out to be much more elusive for higher eukaryotes; a reason for this is yeast promoter regions are usually located in the direct vicinity of the gene (Mellor 1993) they are regulating (up to 600 bp upstream from the translation start site was used in the papers above), contrasting with mammals where the regulatory regions can be located many thousands of base-pairs either upstream or downstream of their target gene(s). Comparative genomics offers some help – it is assumed that functional genomic regions, including regulatory regions, are conserved between evolutionary related species. Nevertheless identifying novel binding sites by sequence analysis in a similar manner that has been successful in Fungi, has turned out to be difficult in animals. Instead, researchers have concentrated on identifying the presence of transcription factor binding sites utilising a priori knowledge (e.g. position weight matrices).

Recently this avenue of research has gained new impetus with the availability of many different genome-wide gene expression sets in the public domain, most importantly in public databases such as ArrayExpress (Parkinson et al. 2007) and Gene Expression Omnibus (Wheeler et al. 2005). In this chapter we will describe a combination of approaches that takes advantage of the availability of these datasets. Initially, we will discuss how to find sets of tightly co expressed genes (Sect. 2) – if such a set includes a transcription factor, it is possible that this transcription factor is one of the regulators of the set. We will discuss how to use a comparative genomics approach to narrow down the putative regulatory regions (Sect. 3), map known transcription factor binding sites onto these regions (Sect. 4) and combine information about several binding sites to predict them more reliably (Sect. 5). Finally, the feasibility of predicting the presence of novel binding sites by statistical analysis of promoter regions of co-expressed genes will be assessed (Sect. 6).

2 Generating sets of co-expressed genes

The ArrayExpress Warehouse is a database of gene expression profiles, and allows for queries based on a range of gene annotations including gene symbols, Gene Ontology

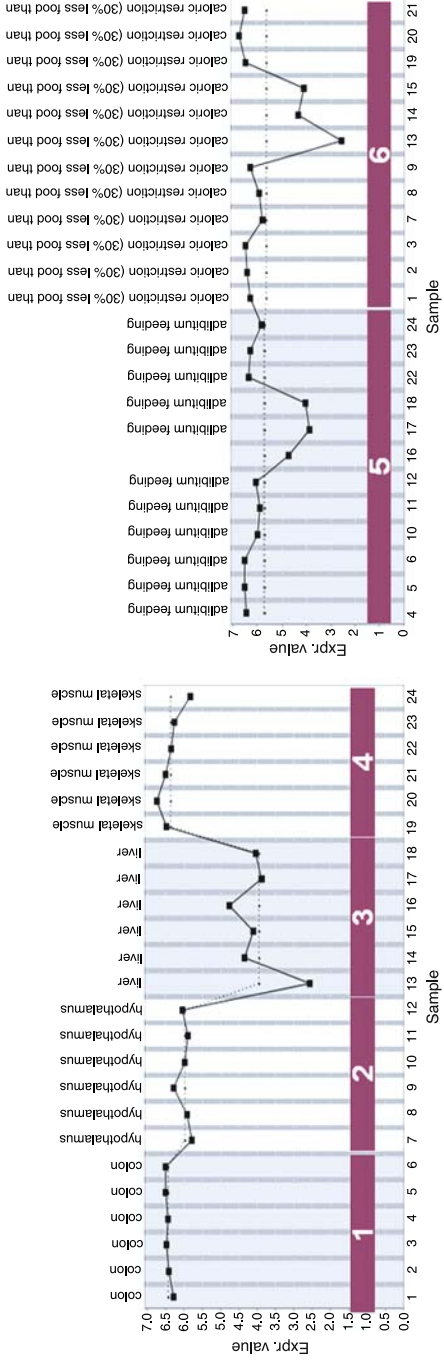


Fig. 1 Differentially expressed transcription factor expression profile

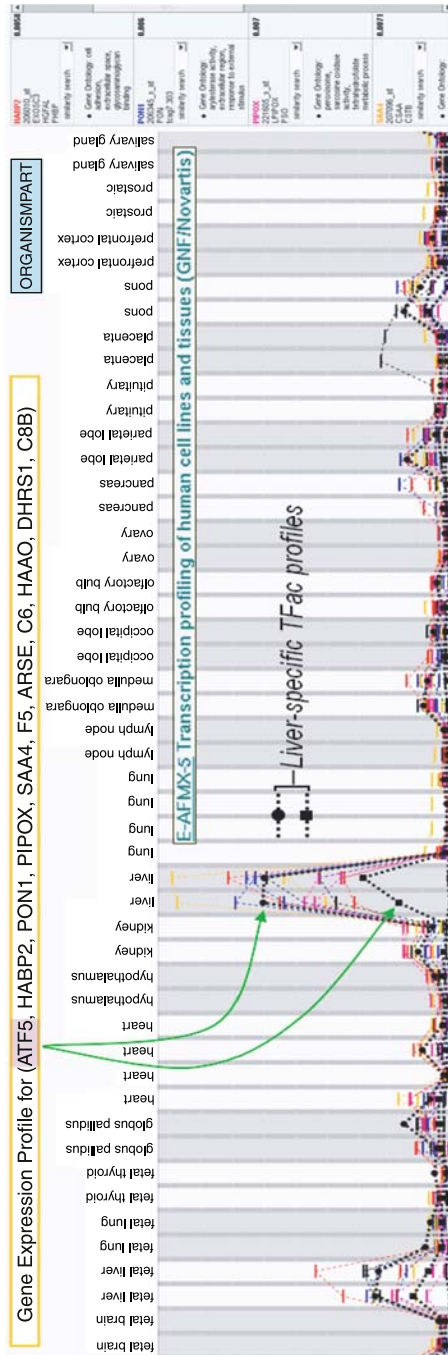


Fig. 2 Co-expressed gene cluster

terms and disease associations. The probesets present on the metazoan Affymetrix microarrays stored in the warehouse have been mapped to ENSEMBL gene entries, and this facilitates the linking of results from protein sequence searches to the expression data contained within the warehouse. An application of this mapping information is the identification of transcription factor expression profiles; human and mouse genes are queried with statistical descriptions of DNA-binding domains (DBDs), and the genes encoding DBD-containing polypeptides (transcription factors) are associated with their corresponding probesets.

The identification of transcription factors differentially expressed between the various experimental factors in the transcriptomics experiments, is initiated by calculating the average expression level in each experimental factor group, and then comparing it with all the other group average expression levels. More specifically, the R software-package “LIMMA” (Linear Models for Microarray Data) – which performs an one-way ANOVA – is employed, and after adjusting the probabilities for multiple testing using the Benjamini–Hochberg False Discovery Rate correction, any DBD-gene with $p < 0.05$ is considered to be differentially expressed between the various experimental factors in the experiments. Hence, if the average expression levels are significantly different between the groups, then the transcription factor is considered to be differentially expressed. In Fig. 1, the average (displayed as a dotted line) is significantly different between groups 1, 2, 3 and 4, but not between groups 5 and 6. Therefore, this transcription factor exhibits differential expression between organism parts i.e. its expression level differs in the colon, hypothalamus, liver and skeletal muscle.

To create sets of co-expressed genes containing transcription factors i.e. a potential regulator of these genes, the genes displaying correlated and significant differential expression for the same experimental factors as the transcription factor are pooled. The co-expression set is then generated by including the differentially regulated genes whose expression profiles are significantly correlated with that of the transcription factor. Figure 2 shows a set of co-expressed genes: the expression profile of a liver-specific transcription factor (ATF5), along with the ten closest expression patterns, is plotted:

The determination of the presence of conserved DNA-motifs in regions both upstream and downstream of the transcription start sites of the genes present in these co-expressed clusters, is discussed below. Furthermore, the co-expressed gene sets can be utilised to ascertain if orthologous transcription factors control the expression of orthologous genes, and if the set members exhibit enriched GO term content.

3 Finding putative regulatory regions using comparative genomics

Chemically active and biologically significant transcription regulatory regions are often assumed to be evolutionarily conserved. This assumption holds for many currently

characterized enhancer elements but there are important counterexamples for this property (Prabhakar et al. 2006). Deletion of large gene deserts including highly conserved DNA segments can yield viable mice and, even more amazingly, deletion of ultraconserved segments of DNA, that work as expression enhancers in transgenic marker assays during development, have almost no noticeable effect on the mouse phenotype (Nobrega et al. 2004; Ahituv et al. 2007). These limiting results are augmented by the recent laboratory results that many of the chemically active DNA locations are not evolutionarily conserved (Birney et al. 2007; Margulies et al. 2007). Keeping these constraints in mind we can proceed to finding conserved non-coding sequences that putatively work as expression enhancers.

The simplest way of finding conserved non-coding sequences is to locally align two evolutionarily related DNA sequences and concentrate on the best conserved elements. This can be done quite efficiently with e.g. the two sequence variant of the BLAST program (Tatusova and Madden 1999). The significance of the local conservation cannot be estimated with the BLAST provided *E*-values that are based on the false null hypothesis of unrelated sequences (Palin and Ukkonen 2008).

Multiple sequence alignment can improve the power of detecting the evolutionarily conserved regions. The multiple sequence comparison benefits from the long total time of independent but parallel evolution among the sequences while the sequences still share a common and a reasonably recent ancestor. This way the biologically significant regions should be well conserved while the insignificant regions have accumulated a large number of mutations such that the sequence conservation should be more pronounced indicator of biological significance as in pair wise alignments.

Currently there are a multitude of whole genome sequences available for a wide range of species. The NCBI is aware of 188 eukaryotic genome sequencing projects in at least assembly stage and 238 projects in more preliminary stage. The comparison of several of these genomes is probably the most efficient method for discovery of evolutionarily conserved sequences. The most interesting genomes are the ones reasonably close to human e.g. the primates or mammals. Larger sequence divergence is likely to lead to loss of sensitivity in the discovery of the regulatory elements that can evolve with reasonably light sequence constraint (Prabhakar et al. 2006). The transcription factor binding sites on the regulatory element can move, allow mutations on the binding DNA and the sites can even be replaced with other similar binding sites nearby.

Evolutionary and sequence pattern extraction through reduced representation (ESPERR) is a recent method for finding the elements conserved for a particular purpose, for example transcriptional enhancer activity, in multiple species (Taylor et al. 2006). The system evaluates the so-called Regulatory Potential Score which discriminates the regulatory regions from neutral sites. ESPERR is a supervised classification system that classifies a multiple DNA alignment to one of two classes. The class can be regulatory/non-regulatory elements or some other DNA sequence annotation.

To avoid handling the discrete alphabet of alignment columns, which is of size 5^k for k species allowing gaps, the ESPERR reduces the representation of the alignment columns to small, in the range of 15–30, set of symbols, each standing for a set of alignment columns. The symbol sets are learned heuristically to maximize the discrimination between the two classes. The initial symbol set is generated by clustering the alignment columns with similar ancestral nucleotide distribution. The initial symbol set is further truncated by joining/extracting initial symbols such that the final classifier improves in accuracy. The accuracy of the classifier is estimated with cross-validation. The search for the best symbol set is stochastic hill climbing with few methods to escape local optima.

The actual classification of the multiple alignments is done according to the log likelihood ratio between two variable order Markov models (VOMMs). Both Markov models use the symbol set obtained above. One of the VOMMs models the multiple alignments of the sequences belonging to the class of interest and the other one models the multiple alignments of the sequences not belonging to the class.

Using discriminative system for the classification has the advantage that one can distinguish between conservation due to different reasons. For example ESPERR is able to classify highly conserved non-coding sequences as developmental enhancers/not enhancers with leave-one-out (LOO) cross-validation accuracy of $\sim 83\%$. The data consisted of transgenic reporter results at embryonic day 9.5 with 108 positive and 138 negative samples (Pennacchio et al. 2006). The classification results are very good considering that all of the training data, including the negative data points, is extremely well conserved over all used species (human, mouse, opossum, chicken, frog, zebrafish, and pufferfish). ESPERR finds the less well-conserved putative regulatory regions, however there is not a good ‘gold standard’ for these less well conserved enhancers.

4 Detecting common transcription factors for co-expressed gene sets

Given a set of co-expressed genes as obtained from ArrayExpress a first question to be addressed is whether there exists a transcription factor(s) regulating the expression of some or all the genes in the set thus causing co-expression. If such a factor exists one would expect to find a common DNA sequence motif corresponding to the binding motif of the factor in the promoter region of some or perhaps all genes in the set.

For a large number of transcription factors, binding motifs derived primarily from small scale experiments such as footprinting have been collected over the last two decades and are now available from public databases such as JASPAR (Sandelin et al. 2004) and TRANSFAC (Matys et al. 2006). As described in more detail in Sect. 2.2 such binding motifs of transcription factors are stored in the form of position-specific frequency matrices, where each column in the matrix shows the preference of the factor for a given base at the corresponding position in the binding site. An example of such a

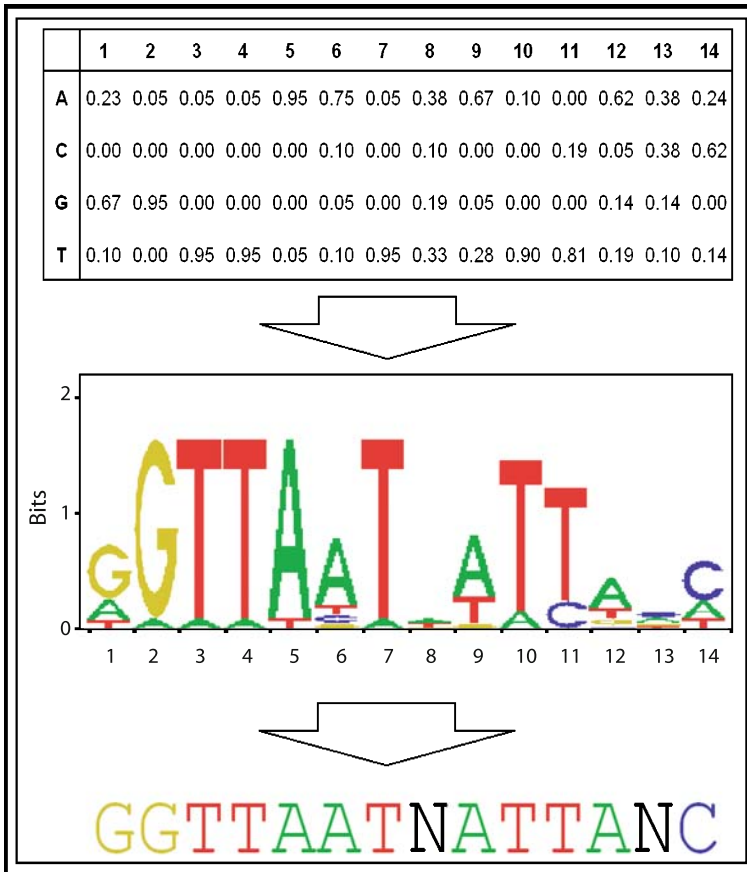


Fig. 3 Frequency matrix of the transcription factor HNF1

frequency matrix is shown in Fig. 3 for the liver-specific transcription factor HNF1. For visual purposes the matrix can be converted into a binding site LOGO (Schneider and Stephens 1990), which indicates the importance of a given base at a given position. Alternatively the binding preference can be reduced to the consensus site for the transcription factor as is shown at the bottom of Fig. 3 for HNF1. The consensus site is usually considered as the site to which the factor binds most strongly.

Given the frequency matrix of a transcription factor one can now scan for potential binding sites of the factor in the DNA sequence corresponding to the promoter region of a gene. Every site in the promoter is thereby compared to the consensus sequence of the factor and a distance measure between the consensus and the given DNA site is computed. Sites exceeding a certain similarity score are considered possible binding sites for the transcription factor. As explained in a previous chapter, various schemes

Table 1 Top target genes for HNF1

Rank	Gene	Expectation score	Belongs to liver set
1	Hgfac	4.12	
2	F13b	1.53	
3	Al182371	1.40	
4	Igfbp1	1.38	
5	Gc	1.02	Yes
6	N/A	0.97	
7	Hc	0.81	
8	Fga	0.78	Yes
9	Mvk	0.75	
10	Mmab	0.52	
11	Afm	0.49	Yes
12	Kif12	0.42	
13	Crp	0.40	Yes
14	Kifc3	0.34	
15	Anpep	0.31	
16	Hoxa9	0.31	
17	Ranbp3l	0.30	
18	Otub1	0.29	
19	Slc26a3	0.28	
20	Fgb	0.28	Yes

exist for computing similarity scores and subsequently applying thresholds to distinguish between binding and non-binding sites. Alternatively similarity scores can be used to calculate the binding probability of a transcription factor to a given site in the DNA sequence. The expected number of transcription factors associated with a promoter is then given by the sum over all individual binding probabilities in the sequence (Roeder et al. 2007). These expectation values can be seen as measure of the affinity between a transcription factor and a promoter and can be used to rank all genes in a genome. As illustrative example Table 1 shows the top ranking 20 genes (out of all 26,000 mouse genes) for the factor HNF1.

Such a ranking allows in the following to investigate whether the transcription factor might play a role in the regulation of a gene set obtained from ArrayExpress. In the case of the factor HNF1 one can see a strong accumulation of genes belonging to the set of liver specific genes among HNF1's top targets. The significance of such an enrichment can be evaluated by applying various statistical test. For instance, the association between HNF1 and the liver specific genes has a $p < 1.0E-15$, which confirms the important role that HNF1 plays in the regulation of gene expression in liver. Following such schemes many important interactions between sets of co-expressed genes and individual transcription factors have been discovered.

Alternative representations for the binding preference of the transcription factor HNF1 (obtained from the JASPAR database). The frequency of a given base at a given

position in experimentally verified binding sites is first stored in the form of a position frequency matrix (indicated on the top). From this matrix a sequence LOGO or the consensus binding site for the transcription factor can be derived. In the latter, positions in the binding site where the transcription factor has no apparent preference for a particular base are indicated by an N, indicating there exists a clear enrichment of liver specific genes among the top ranking targets for this factor.

5 Combining transcription factor information

Over the recent years, several methods have been published that locate clusters of transcription factor binding sites from the non-coding DNA. These so called cis-regulatory modules (CRMs) are assumed to combine the regulatory inputs from the transcription factors binding them and to provide condition- and tissue-specific regulatory output to the basal transcription machinery. The grammar of these cis-regulatory modules is still elusive and the tools for finding new modules vary a lot in their modeling assumptions.

The Enhancer Element Locator (EEL) is a cis-regulatory module prediction tool that makes a lot of biochemically motivated assumptions about the CRM structure and evolution (Hallikas et al. 2006). These assumptions allow efficient analysis of long sequences with large number of transcription factor binding site motifs. EEL finds locally conserved sequences of transcription factor binding sites in two orthologous DNA sequences. The structure of these modules is dictated by the underlying biochemical model and by the mutations occurred after the latest common ancestor of the two compared sequences. A typical module consists of x -binding sites spanning a DNA sequence of length $y-z$.

The reasonably recent common ancestor is a vital assumption for the EEL method because it requires conservation of the exact sequence of binding sites. Because most genes within species have evolved (almost) independently for a long time it is unlikely to find common enhancer elements of two genes by comparing their surrounding DNA regions directly with EEL. More appropriate, and more general, way of finding a putative common regulating transcription factor for a set of genes is to first find CRMs for all genes in the genome and afterwards detect the factors with most overrepresented binding sites in the CRMs of the set of genes.

A simple method for detecting the overrepresented transcription factors is to use apply the Fisher's test on a contingency table. Now assume that we have G genes, of which T have a CRM containing a binding site of the transcription factor of interest. If we are provided a set of C presumably co-regulated genes, x of which have the binding site, we can evaluate the probability of obtaining x , or more, genes with the binding sites, given that we sampled the C genes independently. This situation is depicted in Table 2.

Table 2 Contingency table for TFBS over representation

	With factor	Without factor	
Co-regulated	X	$C - X$	C
Not co-regulated	$T - X$	$G - C - T + X$	$G - C$
	T	$G - T$	

Fisher's test, also known as the hypergeometric test, provides the probability $P(X \geq x)$ which is easily computed with all statistical software packages. Great care should be given to the way of choosing the number of genes G . One has to make sure that on one hand, the genes with/without factor, and on the other hand the co-regulated/not co-regulated genes are really chosen from the set of G genes. It is easy to overlook missing data in one or both of the datasets, which could result in false p -value estimates.

6 “*De novo*” prediction of transcription factor binding motifs

The last problem treated in this chapter will be the “de novo” identification of cis-regulatory motifs from a set of regulatory sequences. Starting from a set of sequences, we want to discover motifs that are found with a higher frequency in a set of promoters of interest than expected by chance. These motifs will be considered as potentially bound by some transcription factor, supposedly responsible for the co-regulation of the genes of interest. Several pattern discovery methods have been developed to tackle this problem (Hertz et al. 1990; Lawrence et al. 1993; Bailey and Elkan 1994; Neuwald et al. 1995; Brazma et al. 1998a; van Helden et al. 1998, 2000; Roth et al. 1998; Hughes et al. 2000; Thijs et al. 2001). In this section, we will present a method based on the detection of over-represented oligonucleotides (van Helden et al. 1998).

Typical applications are the discovery of transcription factor binding motifs from sets of co-expressed genes (expression microarrays), or from sets of genomic fragments where a factor has been shown to bind (chromatin immuno-precipitation experiments). Chromatin immuno-precipitation (ChIP) is a method that permits to fish fragments of DNA bound by a given transcription factor. The “pulled-down” fragments can then be characterized either by sequencing them, or by hybridizing them onto a microarray chip that contains several thousands of genomic fragments. The combination of chromatin immuno-precipitation and microarray hybridization is called ChIP-chip. Odom and co-workers (2004) used the ChIP-chip technology to detect genes whose proximal promoters are bound by three transcription factors (HNF1a, HNF4a and HNF6), in two different cell types (hepatocytes and pancreatic islets).

To illustrate the pattern discovery approach, we report the patterns discovered in 135 promoters bound by HNF6 in hepatocytes (Fig. 4). Promoter sequences were

A

Pattern	Occurrences per position	Expected occurrences per position	Occurrences	Expected occurrences	P-value	E-value	Significance	Rank
AATCAAT ATTGATT	2.94E-04	8.62E-05	27	7.92	8.50E-08	6.90E-04	3.16	1
ATCAATA TATTGAT	2.39E-04	6.67E-05	22	6.13	5.50E-07	4.40E-03	2.35	2
ATCGATC GATCGAT	9.79E-05	1.37E-05	9	1.26	7.00E-06	5.70E-02	1.25	3
GCAATGA TCATTGC	2.39E-04	8.12E-05	22	7.46	1.20E-05	9.60E-02	1.02	4
ATGATTC GAATCAT	2.18E-04	6.95E-05	20	6.39	1.30E-05	1.00E-01	0.99	5
AATCGAT ATCGATT	8.71E-05	1.54E-05	8	1.41	0.00011	9.20E-01	0.04	6

B

Assembly 1		seed: aatcaat	3 words
Direct	Reverse	score	
AATCAAT.	.ATTGATT		3.16
AATCGAT.	.ATCGATT		0.04
.ATCAATA	TATTGAT.		2.35
AATCAATA	TATTGATT	strict	
AATCRATA	TATYGATT	degenerate	

Assembly 2		seed: atcgatc	4 words
Direct	Reverse	score	
GATCGAT.	.ATCGATC		1.25
AATCGAT.	.ATCGATT		0.04
.ATCGATC	GATCGAT.		1.25
.ATCGATT	AATCGAT.		0.04
GATCGATC	GATCGATC	strict	
RATCGATY	RATCGATY	degenerate	

Isolated patterns: 2		
Direct	Reverse	score
GCAATGA	TCATTGC	1.02
ATGATTC	GAATCAT	0.99

C



Fig. 4 Oligonucleotides significantly over-represented in a set of 135 promoters bound by HNF1a. (A) significant oligonucleotides detected with oligo-analysis. (B) motifs resulting from the assembly of the 6 significant oligonucleotides. When alternative residues are found at a given position, we highlight the least significant variants (according to the significance score in the third column). (C) sequence logo of the annotated HNF6 binding motif.

retrieved from -700 to $+200$ from the transcription start site (this is the size of the fragments spotted on Odom's chip), and submitted to the program oligo-analysis (van Helden et al. 1998). The primary result of oligo-analysis (Fig. 4A) is a list of oligomers that show a significant level of over-representation in the sequences analyzed. In the promoters bound by HNF6, the most significant heptamer is the oligonucleotide AATCAAT (regrouped with its reverse complement ATTGATT), which is found in 27 occurrences. The reason for regrouping motifs with their reverse complement is that most transcription factors can recognize their sites irrespective of their orientation (DNA strand). For a random selection of promoters of the same size, we would expect an average of 7.92 occurrences. The P -value ($8.5E-08$) indicates the probability to observe at least 27 occurrences when expecting 7.92. Since the same test is applied to a large number of heptamers, we apply a multi-testing correction, by multiplying the P -value by the number of patterns tested. The resulting E -value ($6.90E-04$) indicates the number of patterns that would be expected with such a level of over-representation by chance, i.e. the expected number of false positives. Since this E -value is very low, the 27 occurrences of AATCAAT are very unlikely to result from chance, and we have thus

good reason to think that they reflect some biological effect (and, of course, our first hypothesis will be that they reflect the enrichment of these promoters in HNF6 binding sites).

The program oligo-analysis compared in the same way the observed and expected number of occurrences for each heptanucleotide found in the input sequences (8090 distinct pairs of heptanucleotides were found with at least one occurrence). Among those, no more than 6 passed the restrictive threshold of $E \leq 1$ (Fig. 4A).

Some of those 6 significant oligomers are strongly mutually overlapping, and can be assembled to form larger motifs, with substitutions at some positions (Fig. 4B). The first assembly is formed by 3 heptanucleotides, that altogether form a partly degenerated 8-mer. The strict consensus (obtained by taking the most significant letter at each position) is AATCAATA, which perfectly fits the motif annotated in TRANSFAC (Fig. 4C). The variant AATCGAT is not part of the annotated motif (which however includes a variant AATCCAT at the same position), and might either be an artifact, or reveal some variant of HNF6 motif that was not present in the 13 known binding sites that were used by TRANSFAC annotators to build the HNF6 matrix (M00639). The second assembly (GATCGATC, Fig. 4A) suggests one further variation at the first position, which might contain either an A or a G.

It should be stated that Fig. 4 shows the typical result of an analysis that worked reasonably well. Unfortunately, this is far from being always the case, especially for human sequences. As mentioned in the introduction, human cis-acting elements are often found at very distant locations from the transcription start site, so that a collection of proximal promoters (as analyzed for this study case) might fail to contain many of the sites actually bound by the factor. In our study cases, we were able to use a collection of relatively short sequences (from -700 to $+200$ around the transcription start site) that corresponds to the fragments hybridized on the ChIP-chip experiment. However, when working with expression microarrays, the only information at hand is the list of up- or down-regulated genes, without any information about the region where the actual transcription factor binds. Another problem is that vertebrate genomes vary a lot in their nucleotide composition, depending on various factors (distance to the TSS, CpG islands, . . .). Consequently, the background models used for estimating the expected frequencies might be insufficient to reflect these heterogeneities, which sometimes results in a large number of false positives. From our experience, *de novo* prediction of regulatory motifs works pretty well in microbial organisms (yeast, bacteria), reasonably well in some insects (*Drosophila*) and plants (*Arabidopsis*), but gives results of variable quality in vertebrates. Systematic evaluations are important to better understand the factors that affect the rate of success of pattern discovery, and to provide clues for selecting optimal parameters and for improving the methods. The evaluation of pattern prediction will be treated separately (see chapter by Sand et al. in this book).

In summary, understanding gene regulation is one of the most challenging targets for genome annotation, since genetic regulation is the key of development, interaction

with the environment, and evolution. The field of regulatory genomics is relatively new, and there is good hope that the methods presented in this chapter will progressively improve in the near future, with the help of high-throughput technologies and the multiplication of sequenced genomes that can be used in comparative genomics.

References

- Ahituv N, Zhu Y, Visel A, et al. (2007) Deletion of ultraconserved elements yields viable mice. *PLoS Biol* 5: 1906–1911
- Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2: 28–36
- Birney E, Stamatoyannopoulos JA, Dutta A, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816
- Brazma A, Jonassen I, Eidhammer I, et al. (1998a) Approaches to the automatic discovery of patterns in biosequences. *J Comput Biol* 5: 279–305
- Brazma A, Jonassen I, Vilo J, et al. (1998b) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res* 8: 1202–1215
- Bucher P (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* 212: 563–578
- Chen QK, Hertz GZ, Stormo GD (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci* 11: 563–566
- Ghosh D (1990) A relational database of transcription factors. *Nucleic Acids Res* 18: 1749–1756
- Hallikas O, Palin K, Sinjushina N, et al. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* 124: 47–59
- Hertz GZ, Hartzell GW 3rd, Stormo GD (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci* 6: 81–92
- Hughes JD, Estep PW, Tavazoie S, et al. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* 296: 1205–1214
- Lawrence CE, Altschul SF, Boguski MS, et al. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262: 208–214
- Margulies EH, Cooper GM, Asimenos G, et al. (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* 17: 760–774
- Matys V, Kel-Margoulis OV, Fricke E, et al. (2006) TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34: D108–D110
- Mellor J (1993) Multiple interactions control the expression of yeast genes. In: Broda PMA, Oliver SG, Sims PFG (eds) *The eukaryotic genome: organization and regulation*. Cambridge University Press, Cambridge, pp 275–320
- Neuwald AF, Liu JS, Lawrence CE (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* 4: 1618–1632
- Nobrega MA, Zhu Y, Plajzer-Frick I, et al. (2004) Megabase deletions of gene deserts result in viable mice. *Nature* 431: 988–993
- Odom DT, Zizlsperger N, Gordon DB, et al. (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science* 303: 1378–1381
- Palin K, Ukkonen E (2008) Statistical significance of above neutral conservation in local alignments. (Submitted)

- Parkinson H, Kapushesky M, Shojatalab M, et al. (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35: D747–D750
- Pennacchio LA, Ahituv N, Moses AM, et al. (2006) *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* 444: 499–502
- Prabhakar S, Poulin F, Shoukry M, et al. (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* 16: 855–863
- Roider HG, Kanhere A, Manke T, et al. (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 23: 134–141
- Roth FP, Hughes JD, Estep PW, et al. (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* 16: 939–945
- Sandelin A, Alkema W, Engstrom P, et al. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32: D91–D94
- Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18: 6097–6100
- Stormo GD, Hartzell GW 3rd (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci USA* 86: 1183–1187
- Tatusova TA, Madden TL (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 174: 247–250
- Taylor J, Tyekucheva S, King DC, et al. (2006) ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res* 16: 1596–1604
- Thijs G, Lescot M, Marchal K, et al. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17: 1113–1122
- van Helden J, Andre B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 281: 827–842
- van Helden J, Rios AF, Collado-Vides J (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* 28: 1808–1818
- Wheeler DL, Barrett T, Benson DA, et al. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 33: D39–D45
- Wingender E, Dietze P, Karas H, et al. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 24: 238–241
- Wolfertstetter F, Frech K, Herrmann G, et al. (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput Appl Biosci* 12: 71–80

SECTION 3

Annotation and genetics

CHAPTER 3

Annotation, genetics and transcriptomics

R. Mott

Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, UK

1 Introduction

This chapter discusses how to combine genome annotations of the type described elsewhere in this book with genetic and functional genomics data to find the genes associated with a phenotype, and in particular with a complex disease. This problem is of fundamental importance; the promise that understanding the molecular basis of common diseases would lead to effective treatments helped motivate and fund the human genome project.

Complex diseases such as cancer, diabetes, cardiovascular disease and depression are defined as conditions with multiple causes, both genetic (due to mutations in the genome) and environmental (everything else). By contrast, a Mendelian disease is caused by mutations in a single gene, with minimal environmental contribution. With a few exceptions such as cystic fibrosis in Caucasians and sickle-cell anaemia in parts of equatorial Africa, most Mendelian diseases are rare and do not impose a major health care burden on society. Most common diseases are complex, the exceptions being caused by infectious agents such as HIV and tuberculosis, and even in these cases there is a genetic contribution to resistance to infection.

In general, most complex diseases have a significant genetic component which we can estimate by examining the co-prevalence of a disease in genetically identical (monozygotic) twins compared to non-identical (dizygotic) twins, who only share 50% of their DNA by descent. Because the average effect due to shared environment should be the same in the two groups, any excess in co-prevalence is likely to be genetic. Thus it is possible to estimate the extent of the genetic contribution to a disease without identifying the causative genes and polymorphisms (Mather and Jinks 1982).

The ultimate aim of gene annotation is to describe the function of every segment of the genome, including protein coding genes as well as micro-RNAs, transcription-factor binding sites and other cryptic functional elements. In addition we want to annotate the

Corresponding author: Richard Mott, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 9JN, UK (e-mail: richard.mott@well.ox.ac.uk)

functional consequence of every polymorphism observed in a population. If we had a perfectly annotated genome then we could predict which genes are relevant to each disease, and there would be no need for further work. However, in fact we have only begun to scratch the surface of the annotation problem, and we will need to be able to integrate data from multiple sources in order to make progress.

Before going further it is important to clarify what is meant by the phrase “gene function”. This turns out to be a surprisingly difficult concept, depending on the context in which the question is being asked. Gene function may be defined at a number of levels. For example, for protein-coding genes, it is important to know in which tissues and at which developmental stages the protein is expressed, and in which splice variants or isoforms. Next, the interactants of the protein are important, as they define the pathways in which the protein functions. Finally we wish to understand the consequences of perturbations to the gene’s DNA sequence; as these may give rise to genetic disease. For some well-studied systems the answers to these questions form a coherent synthesis: For example, we find that for the genes for α and β haemoglobin:

tissue mRNA expression: α and β globins mRNAs are expressed in adult bone marrow, the site of red blood cell production.

interaction: α globin and β globin bind to form haemoglobin.

biological process: haemoglobin is involved in the transport of oxygen.

genetic disease: mutations in the DNA sequences of the globin genes can cause sickle-cell anaemia, thalassaemias and other blood-related diseases.

But in other cases the available data are incomplete or misleading. Even in the well-understood case of the globin genes, non-trivial levels of gene expression are found in other tissues, such as the spleen and in foetal liver, which is the initial organ of red blood cell production. Thus unless one understands the biology of red blood cells these observations could mislead.

Gene function may be defined either in terms of some property of the gene that is common to all individuals, or as the consequences of a genetic variation within the gene which is present in a subpopulation. One might naively assume that the two are necessarily related, but this need not be so. For example α tubulin is a structural protein that is a component of the microtubules found inside cells. Surprisingly, it has recently been shown that mutations to α tubulin cause abnormal neuronal migration in mice and lissencephaly (a lack of normal folds in the brain) in humans (Keays et al. 2007)). The causative mutation prevents tubulin heterodimer formation and *a priori* would not be expected to have this specific effect.

Thus the first, biological, difficulty in annotating genes is that most genes can be pleiotropic with several functions, depending on the context. The second, socio-logical, problem is that direct experimental evidence linking a gene to a particular function is incomplete, and we cannot interpret absence of evidence as evidence of

absence. This is a particular problem when mining the literature for information relating to a gene. A gene that is apparently linked to a disease may well attract further investigation and accumulate more references, possibly spuriously inflating the link. For example there is a vast and mostly inconclusive literature investigating the link between the DRD2 gene and alcoholism (Munafo et al. 2007). At the other extreme a significant fraction of human genes are barely annotated, arising either as computational predictions or as database matches with mRNA sequences. Consequently, when given a list of candidate genes to prioritise for further exploration, there is a natural reluctance to invest expensive resources (such as making and testing a mouse knockout) in a gene about which little is known and which might prove to be a false positive prediction.

The third difficulty is that most quantitative experimental data from high-throughput experiments such as gene expression microarrays (to measure levels of gene expression) or yeast two hybrid studies (to find protein–protein interactions) are indirect and noisy, yet this is the only way to screen large numbers of genes in an unbiased manner. There are now very large databases of high-throughput data available but we are only beginning to understand how to analyse and link it with other resources, and how to resolve apparent contradictions that arise.

We need to make progress in three strategic areas. First, we seek the commonly occurring genetic variants in human populations that are associated with complex disease by looking directly for genotype-phenotype correlations in case-control association studies. Second, we develop appropriate animal models of complex disease; the most common models use mice or rats. Third, in order to integrate these approaches we need accurate comparative functional annotation of the human and other genomes. The remainder of this chapter discusses these issues in greater detail and illustrates the successes and pitfalls that can occur along the way.

2 Genetics and gene function

2.1 Genetic association studies in humans

Whole Genome Association studies (WGA) seek to identify DNA variants that produce phenotypic variation, often in the context of disease. The basic idea is simple: survey a large population of unrelated individuals for disease status (phenotype), and for differences in their DNA (genotypes). Then search for statistically significant correlations between phenotype and genotype (Fig. 1).

Until recently this has proved very difficult, for two reasons. First, overoptimistic assumptions were made about the likely contribution a single genetic variant would make to the likelihood of disease, leading to sample size estimates in the low hundreds that were far too small to detect the very weak signals that actually occur; it is now clear

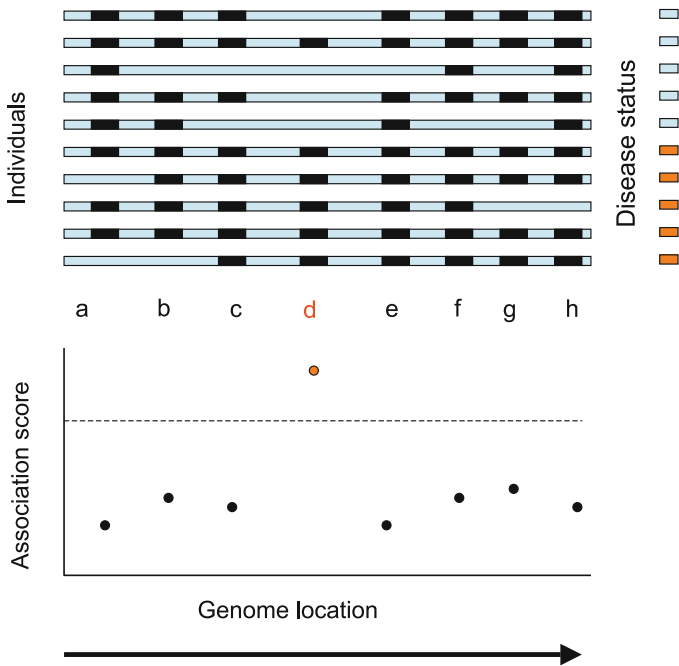


Fig. 1 The principle of genetic association. In the figure, the genomes of 10 unrelated individuals are represented as 10 horizontal bars. Positions of genotyped polymorphisms are labelled a–h, and the genotype of each individual is represented as either light blue or by a black rectangle. The disease status of each individual is represented by the column of blue and orange rectangles on the right. The polymorphism that corresponds most closely to the disease status is taken as being closest to the disease-causing variant. In this example polymorphism d matches the pattern of disease incidence almost perfectly and is therefore the most likely location of a disease gene. The degree of statistical association is represented quantitatively by the graph at the foot of the figure, the choice of units depending on the test of association being used. The horizontal dotted line indicates a threshold for genome-wide statistical significance; the significance of the test of association for each polymorphism is indicated by a black or orange dot at the corresponding position along the genome. In a real whole genome association study involving human subjects, the situation is more complicated because thousands of individuals are genotyped at hundreds of thousands of polymorphisms, and moreover each individual has two copies of each chromosome, which can carry different alleles, but the basic principle is the same

that sample sizes in the thousands are necessary, and in some cases in the tens of thousands. Second, the technology for genotyping around one million single nucleotide polymorphisms (SNPs) at high accuracy and low cost that is necessary to perform an association study has only just recently-become a reality. The result has been that during 2005/2006 there has been a flood of publications of association studies reporting genes associated with many common human diseases (for example see (Nejentsev et al. 2007; Todd et al. 2007; WTCCC 2007; Zeggini et al. 2007)).

It is interesting to survey the newly discovered disease genes and ask which would have been predicted *a priori* based on pre-existing genome annotations: in some cases the genes confirm what had previously been suspected, or at least are not surprising (for example the Insulin gene *INS* in type I diabetes (Todd et al. 2007)), but in many cases, the genes are wholly unexpected (for example the association between the gene *FTO* and obesity (Frayling et al. 2007)), and many previously suspected candidate genes that had been expected to be found did not show association.

One explanation for these negative results is that WGA seeks DNA variants, not genes. A gene may be in a pathway that is vital for a disease, but if it does not contain any functional DNA variants (ie polymorphisms that affect the expression of the gene product in some way, either by altering the amino acid sequence or altering splicing or the level of transcription) then it will not be detected by WGA.

To complicate matters further, the functional DNA variant need not be in or near the gene on which it acts: For example, in the polydactylous mouse mutant Sasquatch, the gene *Shh* is expressed at an ectopic site. Characterization of the mutant led to the identification of an *Shh* enhancer element that lies within intron 5 of a novel gene *Lmbr1* that is also involved in limb development (Clark et al. 2000), but is situated 1 Mb from *Shh* (Lettice et al. 2003). Consider what could happen if an association had been found at this regulatory region: the obvious, but erroneous, conclusion would be that *Lmbr1* was the responsible gene.

The advantage of the genetic approach to identifying disease genes is that it is an unbiased black box that simply finds SNPs that are associated with the phenotype; it does not make any hypotheses about the mechanism linking the DNA to the phenotype. Viewed another way, it is also a weakness because in order to turn the genes into therapies it is necessary to understand these mechanisms. A further caveat is that WGA cannot distinguish between SNPs that cause the phenotypic variation and nearby SNPs that are in tight linkage disequilibrium (LD). This is a particular problem when a tightly linked cluster of genes is identified by this method – it cannot distinguish between them.

LD is a statistical measure of the correlation observed across individuals between pairs of SNPs (Zondervan and Cardon 2004). It is a function of the population size and structure and of the pattern of recombination in the genome, which is distributed non-randomly with recombination being concentrated in hotspots. LD decreases with physical separation between SNPs, and the rate of decrease determines the density of SNPs at which it is necessary to genotype, and the mapping resolution that will be obtained. For example in most Caucasian populations LD decay takes place over about 50 kb on average (although it varies across the genome) which gives single-gene resolution (the average distance between genes being about 100 kb). In some African populations the decay rate is much faster, while in isolated populations (such as Iceland) it is slower (Frazer et al. 2007). In order to detect a functional variant by WGA, it is not necessary to genotype the variant itself, but only a SNP that is in strong LD with the variant; these are called tagging SNPs. Thus there is a trade-off between the number of

tagging SNPs required to cover the genome and hence detect the mapping resolution. At the time of writing, the number of SNPs currently genotyped on a single array is about 500,000, but should reach 1,000,000 during 2008, increasing power to detect associations as a consequence.

3 Use of animal models

Despite the success of human WGA studies, it is clear that only a fraction of the genetic component of complex disease (for example as estimated by twin studies) has been found (Todd et al. 2007). The most likely explanation is that there are many more genes involved, each of very small effect, which are undetectable using the current samples sizes. In order to identify these genes it is therefore helpful to understand the genetic pathways involved in disease rather than taking a simple-minded one-gene-at-a time approach.

The strategy of looking for gene networks rather than genes is sometimes called systems biology. The approach is more tractable in animal models of complex disease rather than directly in humans: it is generally not possible to collect the necessary data from human subjects for ethical and practical reasons.

The mouse and rat are by far the most common models of human disease because they are relatively inexpensive to breed and have a short generation time. In addition, a large number of disease models and other resources are now available. Examples of disease models include specific gene knockouts (e.g. the mouse knockout of the *Hdn* gene exhibits phenotypes similar to those of humans with Huntington's disease (Cha et al. 1998)), inbred strains selected to exhibit a disease-related phenotype (such as the NOD (Non-Obese Diabetic) strain of mice, a model for diabetes (Makino et al. 1980)), or an environmental stimulus that induces the phenotype (such as many behavioural tests (Solberg et al. 2006)).

Knockouts and other transgenic animals are now used routinely to test the phenotypic effects of removing or modifying the expression of a single gene. It is possible to make conditional knockouts in which a gene is only switched off in a particular tissue and developmental stage, which is applicable where an ordinary knockout would be lethal (Ray et al. 2000). There is now an international programme underway to make a knockout of every mouse gene (Austin et al. 2004).

Furthermore, the phenotypic consequences of naturally-occurring variation in the mouse and rat genomes can also be studied to understand the genetic basis of complex disease in a similar way to a WGA in humans. Many inbred strains of mice and rats have been developed in which the genomes of animals within a strain are virtually identical by descent. By making experimental crosses between these inbred strains it is possible to identify quantitative trait loci (QTL) associated with the phenotype. Although the basic principle is the same as in a human WGA, the details and outcomes are different. First,

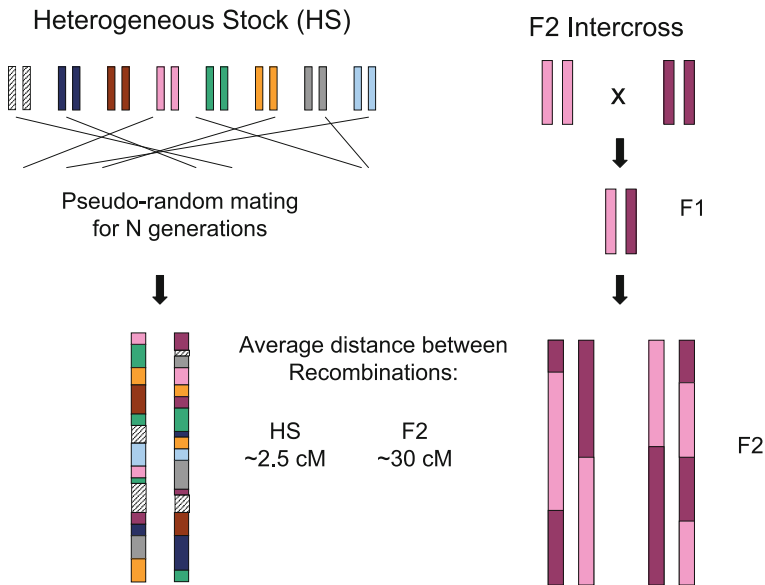


Fig. 2 A schematic illustration of the F2 intercross and heterogeneous stock (HS), two experimental designs used for mapping QTLs in rodents. The genome of each inbred founder is represented by a pair of coloured vertical bars. In the F2 intercross two lines (coloured mauve and purple) are crossed, first to make an F1 generation, in which each pair of chromosomes contains one chromosome descended from each parent, and then again to make an F2 generation, where each chromosome becomes a coarse-grained mosaic of the founder genomes, and where the average distance between recombinants is about 30 centiMorgans (cM). In an HS, eight inbred founder genomes are crossed for many generations until their genomes are much finer-grained mosaics, with a distance between recombinants of about 2.5 cM

most of the inbred strains of laboratory mice are descended from a limited pool of founders, so the extent of linkage disequilibrium found among inbred strains is much greater than in human populations. The practical consequence is that while it is much easier to detect QTLs in crosses between inbred strains using small sample sizes of only a few hundred individuals, the resolution is poor compared to a human WGA: in a typical F2 intercross (Fig. 2) the QTL may encompass over 100 genes. There are several ways of tackling this problem. One is to use gene annotation to narrow the search for candidate genes, although our remarks above indicate that at present annotation is too incomplete for this purpose.

The second approach is to test the effect of knocking out each gene, which is time consuming and expensive, and not feasible for many candidate genes. The third approach is to use special populations of mice whose LD structure is closer to humans, resulting in finer-grained mapping resolution.

One successful example has been the use of a heterogeneous stock (HS). This is a population of mice descended from eight known founder inbred strains that have been

intercrossed for many generations (at least 20) until the genome of each individual is a fine-grained mosaic of the founder chromosomes (Fig. 2). The build-up of historical recombinants gives a comparatively high mapping resolution of about 2 Mb, or about 20 genes. This is still not as precise as in humans, where close to single gene resolution is usually achieved. However, the power to detect associations is much greater, so the yield of QTLs is much better. For example, our laboratory extensively phenotyped a cohort of 2000 HS mice, measuring over 100 phenotypes covering behaviour, asthma and type II diabetes disease models and many physiological measures (Solberg et al. 2006), and identified 843 QTLs each containing 20–30 genes on average (Valdar et al. 2006).

Recombinant Inbred Lines (RIL) are another very important resource. Like an HS, a panel of RILs is descended from a known set of founders which have been crossed for a number of generations. However, they have then been inbred by repeated brother sister matings over about 20 generations until the genome are fixed mosaics of the founders. The advantage of working with RILs is the genomes are fixed so that for example experiments may be replicated under different environmental conditions. Furthermore it is possible to perform a series of experiments on different animals from the same line and treat them as if they were performed on a single individual. This is particularly useful for gene-expression time-course experiments, where the animal must be sacrificed in order to obtain mRNA to assay gene expression. The Collaborative Cross (CC) is an international programme to generate a resource of about 1000 RILs for general use (Churchill et al. 2004). The CC lines are descended from a carefully chosen set of inbred lines which maximises the diversity in the cross.

4 Transcriptomics: gene expression microarrays

Microarrays are used to assay the comparative levels of mRNA expression. By comparing samples collected under different experimental conditions (for example rats fed of high-fat diet compared with normal chow) or different genetic backgrounds, it is possible to identify sets of genes whose expression covaries accordingly. Thus the information delivered is gene-centric and need not relate directly to DNA variation.

In an experiment comparing different conditions, some of the variation in levels of some of the genes identified will be a direct consequence of the difference in experimental conditions, whilst variation in the levels of other genes will be a downstream consequence. The direction of causality usually cannot be inferred from this experiment alone; it delivers a cluster of co-expressed genes. But where these can be superimposed on pathway or genetic mapping data it may be possible to make such inferences.

In an experiment comparing different genetic backgrounds but keeping experimental conditions constant, for example an analysis of gene expression in a panel of recombinant inbred lines it is possible to infer the direction of causality, because we know that variation in DNA causes variation in expression and not *vice-versa*. The

expression levels of each mRNA probe can be thought of as a quantitative trait and mapped accordingly. The result is a genome scan for each gene, indicating expression QTL (eQTL) which are loci harbouring a DNA variant that influences the expression of the gene. In many cases there is an eQTL near to the gene's location; these are called *cis*-eQTL. The other *trans*-eQTLs indicate long-range interactions. In some cases many *trans*-eQTL coincide. These are called "hubs" or "trans-bands" and may correspond to transcription factors; one would then expect all the genes under the control of the same hub to share a common mechanism of transcriptional control, such as a common transcription-factor binding site (Chesler et al. 2005; Li et al. 2006). Thus by identifying clusters of co-expressed genes and then looking for eQTLs for these genes, one can begin to infer the direction of causality.

The main challenge with the analysis of expression data is how to relate it to the identification of the genes responsible for a disease phenotype. The simplest method is to first identify QTLs for the disease, and then look for eQTLs within the QTL interval. Assuming that the right tissue has been assayed for gene expression and that changes in gene expression are indeed responsible for the phenotype, it follows that we should restrict attention to the eQTLs (Wang et al. 2007).

Gene-expression profiling, when combined with genetic mapping data can help to identify candidates (Aherrahrou et al. 2007; Meng et al. 2007). For example, the fatty-acid translocase gene CD36 was identified as a QTL that affects insulin fatty-acid metabolism (Eaves et al. 2002), and complement factor 5 has been suggested as a gene at a QTL that influences susceptibility in a model of asthma (Karp et al. 2000).

Unfortunately, differential gene expression is not always a marker of a QTL. Variants that alter protein structure might not alter expression levels, or there could be compensatory mechanisms that obscure the effect of a QTL on expression. For example a microarray analysis of a model of type I diabetes showed that gene-regulatory systems seemed to be remarkably robust to genetic variation (Eaves et al. 2002).

Second, expression differences might be restricted to certain tissues or developmental stages. For example, expression of 5HT1a receptors (*Htr1a*) in the forebrain is required to modulate anxiety during embryonic and fetal life, but it is not required for the same task in adult animals (Gross et al. 2002). Therefore, expression differences for genes that encode the serotonin receptor that are relevant to anxiety would not be detected in the adult brain, although the phenotype is assessed in the adult animal. Gene-expression analyses will have to be carried out across a large number of tissues at different developmental times to determine whether the gene is differentially expressed.

While in some cases this strategy may work it is important to recognise that the assumptions mean that further verification is needed, usually by knocking out the gene, or by a quantitative complementation test (Yalcin et al. 2004). Furthermore, a technical problem that has only recently been appreciated is that polymorphisms segregating within the microarray probes can cause hybridisation artefacts that appear as differences in mRNA intensity (Walter et al. 2007). This problem is potentially serious, and

consequently any *cis* eQTL identified from a microarray study should be verified by resequencing the probe to determine any polymorphisms.

For some purposes it is possible to use tissue cultures, either human or animal, in place of animals. The advantage of working with live animals is that organ-specific and developmental-specific gene expression studies can be performed. In contrast, most human tissue cultures are blastocysts, which clearly limit the range of applicability to hypotheses that do not depend on the tissue in which the question is answered. However for some systems such as epidermal development, it is possible to make good tissue culture models.

5 Gene annotation

It is now straightforward to extract the gene structure of every annotated gene from a genome browser such as Ensembl (<http://www.ensembl.org>), and to overlay the positions of annotated SNPs and other variants, and to predict any gross phenotypic consequences, such as a missense mutation or a premature stop codon. Other DNA features such as conserved non-coding sequence or predicted transcription factor binding sites are also annotated. The protein domain composition of each translated gene has also been pre-computed and is accessible *via* databases such as Interpro (<http://www.ebi.ac.uk/interpro/>) (Mulder et al. 2007). These features can be thought of as defining the “syntax” of the genome, as they are analogous the task of parsing and spell-checking a natural language. The “semantics” of genome annotation can be thought of as determining the genes’ functions, and at present relatively little can be predicted purely by sequence composition. We can sometimes predict a protein’s subcellular localisation (ie whether it is nuclear, cytoplasmic or signalling) based on sequence characteristics such as domain composition (Mott et al. 2002), and we can predict the 3D fold of the protein if it is similar to a protein whose structure has been solved and in some cases even when it is unique (<http://predictioncenter.org/>) (Moult et al. 2005)). Knowledge of the protein fold is sometimes helpful in defining function, but in general we cannot yet compute the organism from the genome *ab initio*.

The other source of information derives from the scientific literature. The simplest approach is to search abstracts for references to a particular gene, and this process has been largely automated with a several public databases offering quite sophisticated summaries of published gene functions. For example the NCBI Gene database provides a simple mechanism to allow scientists to add to the functional annotation of genes (geneRIFs <http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html>) described in Entrez Gene. OMIM provides a carefully annotated resource but is generally limited to Mendelian disease. IHOP (Information hyperlinked over proteins) provides an automated text-mining database based on Pubmed abstracts. The Mouse Genome Informatics database curates public information about each mouse gene.

It has been clear for some time that the descriptions of gene function must be codified using an ontology, or controlled vocabulary. This simplifies the task of computerised search because the problem of semantic mapping is shifted from the search to the construction of the ontology, which therefore requires considerable manual curation. The best known ontology is GO, the gene ontology, in which function terms arranged hierarchically (<http://www.geneontology.org/>). However, it should be pointed out that even this database is incomplete and at best can only represent the current state of knowledge – it cannot make deductions about gene function.

Knowledge about gene networks, as deduced from pairs of interacting genes, is also integrated into public gene annotations. The problem here is that the quality of interaction data is highly variable: some enzyme pathways are known accurately (eg the KEGG database <http://www.genome.ad.jp/kegg/pathway.html>), whilst data derived from high-throughput assays such as yeast two-hybrid experiments has a high error rate. Gene co-expression data from mRNA microarray experiments also define networks, but cannot readily distinguish between genuine interaction (where the proteins bind) and co-expression (where protein expression is correlated because it is under common control). Nevertheless it seems clear that progress must come from the systems analysis of genes. An example of the approach needed is provided by WebQTL (<http://www.genenetwork.org/>).

In the future we need to develop statistical analyses and databases to handle gene networks and to deal with the noisiness inherent in the data, and to combine this information with other genome annotations.

References

- Aherrahrou Z, Doehring LC, Kaczmarek PM, Liptau H, Ehlers EM, Pomarino A, Wrobel S, Gotz A, Mayer B, Erdmann J, Schunkert H (2007) Ultrafine mapping of *Dyscalc1* to an 80-kb chromosomal segment on chromosome 7 in mice susceptible for dystrophic calcification. *Physiol Genomics* 28: 203–212
- Austin CP, Battey JF, Bradley A, Bucan M, Capecchi M, Collins FS, Dove WF, Duyk G, Dymecki S, Eppig JT, Grieder FB, Heintz N, Hicks G, Insel TR, Joyner A, Koller BH, Lloyd KC, Magnuson T, Moore MW, Nagy A, Pollock JD, Roses AD, Sands AT, Seed B, Skarnes WC, Snoddy J, Soriano P, Stewart DJ, Stewart F, Stillman B, Varmus H, Varticovski L, Verma IM, Vogt TF, von Melchner H, Witkowski J, Woychik RP, Wurst W, Yancopoulos GD, Young SG, Zambrowicz B (2004) The knockout mouse project. *Nat Genet* 36: 921–924
- Cha JH, Kosinski CM, Kerner JA, Alsdorf SA, Mangiarini L, Davies SW, Penney JB, Bates GP, Young AB (1998) Altered brain neurotransmitter receptors in transgenic mice expressing a portion of an abnormal human huntington disease gene. *Proc Natl Acad Sci USA* 95: 6480–6485
- Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* 37: 233–242
- Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W, Bleich A, Bogue M, Broman KW, Buck KJ, Buckler E, Burmeister M, Chesler EJ,

- Cheverud JM, Clapcote S, Cook MN, Cox RD, Crabbe JC, Crusio WE, Darvasi A, Deschepper CF, Doerge RW, Farber CR, Forejt J, Gaile D, Garlow SJ, Geiger H, Gershenfeld H, Gordon T, Gu J, Gu W, de Haan G, Hayes NL, Heller C, Himmelbauer H, Hitzemann R, Hunter K, Hsu HC, Iraqi FA, Ivandic B, Jacob HJ, Jansen RC, Jepsen KJ, Johnson DK, Johnson TE, Kempermann G, Kendzioriski C, Kottb M, Kooy RF, Llamas B, Lammert F, Lassalle JM, Lowenstein PR, Lu L, Lusis A, Manly KF, Marcucio R, Matthews D, Medrano JF, Miller DR, Mittleman G, Mock BA, Mogil JS, Montagutelli X, Morahan G, Morris DG, Mott R, Nadeau JH, Nagase H, Nowakowski RS, O'Hara BF, Osadchuk AV, Page GP, Paigen B, Paigen K, Palmer AA, Pan HJ, Peltonen-Palotie L, Peirce J, Pomp D, Pravenec M, Prows DR, Qi Z, Reeves RH, Roder J, Rosen GD, Schadt EE, Schalkwyk LC, Seltzer Z, Shimomura K, Shou S, Sillanpaa MJ, Siracusa LD, Snoeck HW, Spearow JL, Svenson K, et al. (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet* 36: 1133–1137
- Clark RM, Marker PC, Kingsley DM (2000) A novel candidate gene for mouse and human preaxial polydactyly with altered expression in limbs of Hemimelic extra-toes mutant mice. *Genomics* 67: 19–27
- Eaves IA, Wicker LS, Ghandour G, Lyons PA, Peterson LB, Todd JA, Glynne RJ (2002) Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. *Genome Res* 12: 232–243
- Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316: 889–894
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Sun W, Wang H, Wang Y, Xiong X, Xu L, Wayne MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861
- Gross C, Zhuang X, Stark K, Ramboz S, Oosting R, Kirby L, Santarelli L, Beck S, Hen R (2002) Serotonin1A receptor acts during development to establish normal anxiety-like behaviour in the adult. *Nature* 416: 396–400
- Karp CL, Grupe A, Schadt E, Ewart SL, Keane-Moore M, Cuomo PJ, Kohl J, Wahl L, Kuperman D, Germer S, Aud D, Peltz G, Wills-Karp M (2000) Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat Immunol* 1: 221–226
- Keays DA, Tian G, Poirier K, Huang GJ, Siebold C, Cleak J, Oliver PL, Fray M, Harvey RJ, Molnar Z, Pinon MC, Dear N, Valdar W, Brown SD, Davies KE, Rawlins JN, Cowan NJ, Nolan P, Chelly J, Flint J (2007) Mutations in alpha-tubulin cause abnormal neuronal migration in mice and lissencephaly in humans. *Cell* 128: 45–57
- Lettec LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12: 1725–1735

- Li H, Chen H, Bao L, Manly KF, Chesler EJ, Lu L, Wang J, Zhou M, Williams RW, Cui Y (2006) Integrative genetic analysis of transcription modules: towards filling the gap between genetic loci and inherited traits. *Hum Mol Genet* 15: 481–492
- Makino S, Kunimoto K, Muraoka Y, Mizushima Y, Katagiri K, Tochino Y (1980) Breeding of a non-obese, diabetic strain of mice. *Jikken Dobutsu* 29: 1–13
- Mather K, Jinks JL (1982) *Biometrical genetics: the study of continuous variation*. Chapman & Hall, London
- Meng H, Vera I, Che N, Wang X, Wang SS, Ingram-Drake L, Schadt EE, Drake TA, Lusis AJ (2007) Identification of *Abcc6* as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. *Proc Natl Acad Sci USA* 104: 4530–4535
- Mott R, Schultz J, Bork P, Ponting CP (2002) Predicting protein cellular localization using a domain projection method. *Genome Res* 12: 1168–1174
- Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A (2005) Critical assessment of methods of protein structure prediction (CASP) – round 6. *Proteins* 61(Suppl 7): 3–7
- Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database. *Nucleic Acids Res* 35: D224–D228
- Munafò MR, Matheson IJ, Flint J (2007) Association of the DRD2 gene Taq1A polymorphism and alcoholism: a meta-analysis of case-control studies and evidence of publication bias. *Mol Psychiatry* 12: 454–461
- Nejentsev S, Howson JM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, Hulme J, Maier LM, Smyth D, Bailey R, Cooper JD, Ribas G, Campbell RD, Clayton DG, Todd JA, Burton PR, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Donnelly P, Barrett JC, Davison D, Easton D, Evans D, Leung HT, Marchini JL, Morris AP, Spencer CC, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Taylor NC, Walters GR, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, St Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskvina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Nicol Ferrier I, Ball SG, Balmforth AJ, Barrett JH, Bishop DT, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, et al. (2007) Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature* 450: 887–892
- Ray MK, Fagan SP, Brunnicardi FC (2000) The Cre-loxP system: a versatile tool for targeting genes in a cell- and stage-specific manner. *Cell Transplant* 9: 805–815
- Solberg LC, Valdar W, Gauquier D, Nunez G, Taylor A, Burnett S, Arboledas-Hita C, Hernandez-Pliego P, Davidson S, Burns P, Bhattacharya S, Hough T, Higgs D, Klenerman P, Cookson WO, Zhang Y, Deacon RM, Rawlins JN, Mott R, Flint J (2006) A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. *Mamm Genome* 17: 129–146
- Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, Wallace C, Howson JM, Guja C, Ionescu-Tirgoviste C, Simmonds MJ, Heward JM, Gough SC, Dunger DB, Wicker LS, Clayton DG (2007) Robust

- associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 39: 857–864
- Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* 38: 879–887
- Walter NA, McWeeney SK, Peters ST, Belknap JK, Hitzemann R, Buck KJ (2007) SNPs matter: impact on detection of differential expression. *Nat Methods* 4: 679–680
- Wang SS, Schadt EE, Wang H, Wang X, Ingram-Drake L, Shi W, Drake TA, Lusis AJ (2007) Identification of pathways for atherosclerosis in mice: integration of quantitative trait locus analysis and global gene expression data. *Circ Res* 101: e11–e30
- WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678
- Yalcin B, Willis-Owen SA, Fullerton J, Meesaq A, Deacon RM, Rawlins JN, Copley RR, Morris AP, Flint J, Mott R (2004) Genetic dissection of a behavioral quantitative trait locus shows that *Rgs2* modulates anxiety in mice. *Nat Genet* 36: 1197–1202
- Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316: 1336–1341
- Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. *Nat Rev Genet* 5: 89–100

SECTION 4

Functional annotation of proteins

CHAPTER 4.1

Resources for functional annotation

A. J. Bridge¹, A.-Lise Veuthey¹ and N. J. Mulder²

¹Swiss-Prot, Swiss-Institute of Bioinformatics, Centre Medical Universitaire, Geneva, Switzerland

²EMBL Outstation – European Bioinformatics Institute Hinxton, Cambridge, UK

1 Introduction

The continued success of genome sequencing projects has led to an explosion in the availability of sequence data. The Genomes On Line Database (GOLD) currently lists more than 2000 ongoing and completed genome projects, and this number is continuously increasing (Liolios et al. 2008). In order for this sequence information to be useful in the formulation and testing of biological hypotheses, these genome sequences must be adequately annotated.

The process of genome annotation begins with the identification of all predicted gene sequences, promoters and regulatory regions within the genome using a variety of computational techniques, which are discussed in greater detail in Chap. 2 of this book. Following the identification of putative genes, the next step is the definition of the proteome, which is the complement of all possible protein sequences encoded by the genome in question. The complexity of the proteome is enhanced by the possibility of alternative splicing and other modifications to the predicted protein sequences such as proteolytic processing. Once the genome has been annotated and the proteome defined, the next step is to provide functional annotations for the protein sequences. Due to the sheer volume of protein sequence data involved, it is unlikely that most predicted protein sequences will be experimentally characterized in the near future. Therefore most functional annotation of new protein sequences proceeds by the identification of related proteins, or modular protein domains or motifs within the protein sequence, followed by the transfer of their associated annotations to the protein sequence of interest. The accuracy of the final annotations obtained therefore depends on the methods employed for sequence classification and the sources of functional annotations used.

Complete functional annotation of a protein requires a precise description of the biochemical or biological function of the individual protein itself. It also requires a

Corresponding author: Alan J. Bridge, Swiss-Prot, Swiss-Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211-Geneva 4, Switzerland (e-mail: alan.bridge@isb-sib.ch)

description of how individual proteins and other biological entities interact with each other to form macromolecular assemblies, and how proteins and macromolecular assemblies interact to form biological pathways. Information about the precise 3-dimensional (3D) structure of a protein can help elucidate protein function, its mode of regulation and interactions, while proteins can be regulated by post-translational modifications, which must also be described. For proteins implicated in disease susceptibility, complete annotation must include a precise description of the disease and of disease-associated mutations in the protein concerned. In the next section of this chapter, a selection of the available resources providing such annotations will be described. We will then discuss some of the available resources for protein classification and the identification of protein domains and motifs.

2 Resources for functional annotation – protein sequence databases

The first and absolute prerequisite for accurate functional annotation is a correct protein sequence. This may seem obvious but in fact many available protein sequences contain inaccuracies due to technical errors in sequencing or due to difficulties in sequence interpretation (such as failing to accurately determine the boundaries of predicted exons or genes). Without correct protein sequence information, it is impossible to precisely locate sequence features such as individual domains, functionally important residues, modifications, or sites of interaction with other biological macromolecules, nor can related proteins or protein domains be identified with confidence. Correct sequences are therefore of primary importance.

A number of databases provide access to protein sequence information. These can be broadly divided into two categories: simple repositories of protein sequences and annotated protein sequence databases. Sequence repositories provide users with rapid access to newly obtained sequence data, but do not perform correction of erroneous sequences and add little or no annotation to the protein sequences. They may also exhibit high levels of redundancy, where multiple records describe the same protein sequence. One example of a sequence repository is the GenBank Gene Products Databank (GenPept), produced by the National Centre of Biotechnology Information (NCBI) (Benson et al. 2003). The entries in GenPept are derived from translations of sequences contained in the International Nucleotide Sequence Database which is jointly maintained by the DNA Databank of Japan (DDBJ) (Miyazaki et al. 2003), the European Molecular Biology Laboratory (EMBL) (Stoesser et al. 2003), and GeneBank (Benson et al. 2003). Another sequence repository is the NCBI Entrez Protein database, which also contains sequences translated from the International Nucleotide Sequence Database as well as sequences from other sources including the manually annotated UniProtKB/Swiss-Prot database (described below) and the Protein Data Bank (Berman

et al. 2007). Entrez Protein records may contain annotations extracted from annotated databases, although manual curation of the records themselves is not performed.

Annotated protein sequence databases enhance the basic content of sequence repositories by the addition of relevant information from the literature and other sources, including external databases and computational sequence analysis. They may also enhance the quality of the sequences themselves by manual sequence correction and reduce redundancy by merging related sequences into single records. Individual annotated protein sequence databases vary in their scope and coverage; many focus on particular protein types and families or on particular species or taxonomic groupings, while some aim for universal coverage of all protein space. One such universal annotated protein database is the Universal Protein Knowledgebase, or UniProtKB, which is produced by the Universal Protein Resource (UniProt) consortium. The activities of the UniProt consortium, and of UniProtKB, are described in the following sections.

3 UniProt – The Universal Protein Resource

The Universal Protein Resource (UniProt) provides a freely available central resource on protein sequences and functional annotation (UniProt Consortium 2007). UniProt is produced by the UniProt consortium, which was formed in 2002 by the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute for Bioinformatics (SIB).

UniProt is composed of four major components, each of which is optimized for a different use. The first component is the UniProt Knowledgebase (UniProtKB), which is the most comprehensively annotated protein sequence database in existence. UniProtKB provides protein sequence entries with extensive annotation and cross-references, and will be described in more detail below. The second component of UniProt is the UniProt Archive (UniParc), which is the most exhaustive publicly available non-redundant protein sequence database (Leinonen et al. 2004). UniParc obtains sequence data from a number of sources including UniProtKB, EMBL (Stoesser et al. 2003), the partially annotated sequence database RefSeq (Pruitt et al. 2007), Ensembl (Hubbard et al. 2007), and the International Protein Index, or IPI (Kersey et al. 2004). Searching UniParc is therefore equivalent to performing a search against all the source databases simultaneously. The third component of UniProt is the UniProt Reference Clusters (UniRef), which cluster protein sequences from UniProtKB and UniParc at three levels of sequence identity – 50%, 90% and 100% (Suzek et al. 2007). The UniRef clusters are designed to facilitate sequence searches and analysis of the results. For example, by searching the UniRef90 clusters rather than a redundant protein sequence database, one may avoid long lists of very similar high-scoring matches, which would group within a single UniRef90 cluster. The fourth component of UniProt is the UniProt Metagenomic and Environmental Sequences, or UniMES, which was set up to accommodate sequences

derived from environmental samples of unknown taxonomic origin (Suzek et al. 2007). The rate of production and availability of such sequences is likely to grow rapidly in the future, as more and more groups aim to extend our knowledge of sequence space beyond that of laboratory-cultivated organisms.

In the following section, we will describe in more detail the major features of UniProtKB, which many consider to be the central component of UniProt.

4 The UniProt Knowledgebase (UniProtKB)

The UniProt Knowledgebase, or UniProtKB, provides an integrated and uniform presentation of data on protein sequence and function (UniProt Consortium 2007). UniProtKB consists of two separate and quite distinct components; UniProtKB/Swiss-Prot, which contains manually annotated protein sequence records, and UniProtKB/TrEMBL, which contains automatically annotated protein sequence records. UniProtKB/TrEMBL serves as the source of entries for UniProtKB/Swiss-Prot; when entries from UniProtKB/TrEMBL are manually annotated, they are subsequently incorporated into UniProtKB/Swiss-Prot.

The information in UniProtKB/Swiss-Prot and UniProtKB/TrEMBL records is supplemented by the provision of links to over one hundred external specialist databases. These include databases storing sequence and gene model information or providing links to such information, such as EMBL (Stoesser et al. 2003) and RefSeq (Pruitt et al. 2007), databases storing genome annotation such as Ensembl (Hubbard et al. 2007) and Genome Reviews (Kersey et al. 2005), species-specific databases such as WormBase (Bieri et al. 2007), FlyBase (Crosby et al. 2007), and the *Saccharomyces cerevisiae* Genome Database (SGD) (Christie et al. 2004), and protein domain and family databases grouped by InterPro (Mulder et al. 2007). Some of these resources will be described in more detail later. Integration of UniProtKB with other data resources is further enhanced by the extensive annotation of UniProtKB entries with terms from the standardized vocabulary of the Gene Ontology, commonly known as GO (Camon et al. 2004). Biological ontologies such as GO provide controlled terminologies for the assignment of consistent functional annotations to gene products. This facilitates complex database queries and allows the identification of related items in distinct databases with different formats.

4.1 UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot is the manually annotated component of the UniProt Knowledgebase. UniProtKB/Swiss-Prot is the most extensive manually annotated universal protein database in existence; at the time of writing it contains 356,194 records describing proteins from 11,290 species (release 55.0 of 26th February 2008).

Each UniProtKB/Swiss-Prot entry contains several compulsory elements (see Fig. 1). The “Entry information” and “Names and origin” sections of the entry contain an entry identifier, an accession number, the protein name, taxonomic data, and a precise summary of the level of experimental evidence supporting the existence of the protein. Accession numbers are stable and provide a unique means of identifying an entry. Bibliographical reference(s) are displayed in the “References” section, while the protein sequence itself is shown in the “Sequences” section.

This minimum level of annotation is enhanced with information extracted from the literature and external databases, from related UniProtKB entries, and from manually evaluated computational analysis of the protein sequence (see Fig. 2). A precise summary of available biological knowledge is provided in the “General annotation (comments)” section, while annotations pertaining to defined positions within the protein sequence are provided in the “Sequence annotation (Features)” section. Figure 2 shows an extract from a sample entry from UniProtKB/Swiss-Prot, including both “General annotations” and “Sequence annotations”. Specific keywords are used to summarize various aspects of protein biology including function, location, domain content, sequence modifications, ligands, and involvement in disease processes, where appropriate. Keywords provide a convenient means of retrieving lists of related entries; they are stored in the “Ontologies” section, along with terms from the Gene Ontology.

For those annotations which do not derive directly from experimental analysis of the protein concerned, three non-experimental qualifiers are used to indicate their source or the level of confidence associated with them. The qualifier “By similarity” indicates that the particular annotation has been transferred from a homologous protein, where it has been experimentally demonstrated. The qualifier “Probable” indicates that a particular annotation has been inferred based on common biological knowledge or on experimental evidence that is extremely suggestive but not absolutely conclusive. The qualifier “Potential” indicates that a particular feature or annotation is derived solely from computational analysis, such as the presence of predicted transmembrane domains or secretory signals. The use of non-experimental qualifiers allows users to filter entries containing annotations which are not experimentally proven for the protein concerned. As a general rule, annotations derived from experimental analysis are greatly outnumbered by those which are predicted, transferred by homology, or inferred by annotators. For example the current release of UniProtKB/Swiss-Prot contains 83,674 annotated glycosylation sites, of which 76,888 sites were annotated as “Potential”, while 586 sites were annotated as “Probable” and 1887 sites were annotated “By similarity” to a proven site.

Many bioinformatics resources have taken advantage of the high-quality annotations provided by UniProtKB/Swiss-Prot. The manually curated sequences have been used for functional annotation of a number of genomes (see Chaps. 4.2 and 4.3 for examples), and in protein clustering and family classification systems (see Chaps. 4.4 and 4.5). UniProtKB/Swiss-Prot has also supplied datasets for the training and evaluation of sequence feature prediction tools based on statistical and machine

a

Entry information		Hide Top
Entry name	ASF1_CAEEL	
Accession	Primary (citable) accession number: Q19326	
Entry history	Integrated into April 17, 2007 UniProtKB/Swiss-Prot Last sequence update: May 2, 2006 Last modified: April 8, 2008 This is version 49 of the entry and version 2 of the sequence. [Complete history]	
Entry status	Reviewed (UniProtKB/Swiss-Prot)	
Annotation project	Caenorhabditis annotation project	

b

Names and origin		Hide Top
Protein names	Probable histone chaperone asf-1 <i>Also known as:</i> Anti-silencing function protein 1	
Gene names	Name: asf-1 ORF Names: F10G7.3	
Organism	Caenorhabditis elegans [Complete proteome]	
Taxonomic identifier	6239 [NCBI]	
Taxonomic lineage	Eukaryota › Metazoa › Nematoda › Chromadorea › Rhabditida › Rhabditoidea › Rhabditidae › Peloderinae › Caenorhabditis	
Protein existence	Evidence at transcript level.	

c

References		Hide Top
◀ Hide large scale references		
[1]	"Genome sequence of the nematode C. elegans: a platform for investigating biology." The C. elegans sequencing consortium Science 282 2012-2018(1998) [PubMed 9851916] [Abstract] Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA]. Strain: Bristol N2.	
[2]	WormBase consortium Submitted (MAR-2006) to the EMBL/GenBank/DDBJ databases Cited for: SEQUENCE REVISION.	

d

Sequences		Hide Top	
Sequence	Length	Mass (Da)	Tools
<input type="checkbox"/> Q19326-1 [UniParc]. FASTA Last modified May 2, 2006. Version 2. Checksum: 3D468D108F723038	275	31,240	Blast <input type="text" value=""/> <input type="button" value="go"/>
<pre> 10 20 30 40 50 60 MASRVNIVQV QILDNPAMFV DFKFLEITFE VFEHLPHDLE WELVYVGSST SRDFDQVLDS 70 80 90 100 110 120 ALVGPPEGR HKFVFDADHP DISKIPVDDI VGVSVLLLRG KYNDQEF INH GWFVANEYTE </pre>			

learning techniques (Chap. 5.2). Within the framework of the BioSapiens project, several manually curated datasets of sequences with experimentally verified features have been made available to the bioinformatics community (at <http://biosapiens.isb-sib.ch>). These include sets of over 15,000 proteins located in defined subcellular compartments and sorted according to their taxonomic origin. Other datasets include sequences or sub-sequences carrying various types of post-translational modifications. These data can be used to benchmark existing prediction methods and to develop ever more efficient and reliable tools for sequence analysis.

4.1.1 Sequence curation in UniProtKB/Swiss-Prot

We have previously emphasized the absolute importance of correct sequence information for accurate functional annotation. UniProtKB/Swiss-Prot performs extensive manual curation of protein sequences to provide users with the most correct protein sequence possible. During the creation of a single UniProtKB/Swiss-Prot entry, all available UniProtKB/TrEMBL entries pertaining to the same protein in the same species are identified by sequence similarity searches and merged into a single entry. UniProtKB/Swiss-Prot is therefore essentially a non-redundant protein knowledgebase, although it may contain identical sequences when they are derived from homologous genes in related species.

During the merge process, all sequence discrepancies are identified and analyzed. Individual sequences may differ due to biological events such as alternative splicing, alternative promoter usage or alternative translation initiation site usage. The identification and annotation of alternatively spliced protein isoforms is an essential prerequisite for the definition of a complete proteome set. Sequence discrepancies may also arise due to technical problems such as the presence of frameshifts in the underlying nucleotide sequence, the presence of contaminating vector sequence, or simple sequencing errors. Many protein sequences are derived from computational gene model predictions in genomic DNA, and these predictions are not always correct (as discussed in Chap. 2).



Fig. 1 The compulsory elements of a UniProtKB/Swiss-Prot entry. (a) The “Entry information” section contains the entry identifier (ASF1_CAEEL) and the unique stable accession number of the entry (Q19326), plus history, status, and version information. It also includes the name of the annotation project responsible for maintenance of the entry (the *Caenorhabditis* annotation project in this case). (b) The “Names and origin” section contains the protein name and synonyms, gene name and synonyms (when available), taxonomic data, and a precise summary of the level of experimental evidence supporting the existence of the protein. Here the existence of the protein is inferred from homology to other existing protein sequences. (c) The “References” section stores bibliographical reference(s). The information extracted from each reference is listed under “Cited for:”. This entry contains only two references, which describe the original sequence and a subsequent sequence revision. (d) The “Sequences” section shows the actual protein sequence and related information such as the sequence version and protein length and mass. This section also contains a menu that displays tools for sequence analysis, such as BLAST

Chapter 4.1: Resources for functional annotation

a

General annotation (Comments) Hide Top	
Function	Has A-to-I RNA editing activity on extended dsRNA: edits RNA-binding protein Rnp4F. A-to-I editing of pre-mRNAs acts predominantly through nervous system targets to affect adult nervous system integrity, function and behavior. Essential for adaptation to environmental stresses, such as oxygen deprivation, and for the prevention of premature neuronal degeneration, through the editing of ion channels as targets.
Tissue specificity	Expressed in embryonic nervous system, late stage 13 sees ventral nerve cord expression which spreads to brain by stage 16. Expression is maintained through to adulthood.
Developmental stage	Expressed throughout development, highest expression is during pupal stage. Isoforms A, C and D have lowest expression levels.

b

Sequence annotation (Features) Hide Top				
Feature key	Position(s)	Length	Description	Graphical view
Molecule processing				
<input type="checkbox"/> Chain	1 – 676	676	Double-stranded RNA-specific editase Adar	
Regions				
<input type="checkbox"/> Domain	61 – 127	67	DRBM 1	
<input type="checkbox"/> Domain	197 – 272	76	DRBM 2	
<input type="checkbox"/> Domain	348 – 672	325	A to I editase	
Sites				
<input checked="" type="checkbox"/> Active site	374	1	Proton donor By similarity	
<input checked="" type="checkbox"/> Metal binding	372	1	Zinc By similarity	
<input checked="" type="checkbox"/> Metal binding	430	1	Zinc By similarity	
<input checked="" type="checkbox"/> Metal binding	493	1	Zinc By similarity	

c

Ontologies Hide Top	
Keywords	
Biological process	mRNA processing
Coding sequence diversity	Alternative initiation Alternative splicing RNA editing
Domain	Repeat
Ligand	Metal-binding Zinc
Molecular function	Hydrolase
Technical term	Complete proteome
Gene Ontology (GO)	
Biological process	Adenosine to inosine editing Ref.1 Ref.7 Inferred from direct assay. Source: UniProtKB Adult locomotory behavior Ref.2 Ref.7 Inferred from mutant phenotype. Source: FlyBase

Each of these possibilities is investigated using multiple alignments of all candidate protein sequences and by examination of the proposed coding sequence CDS in the context of comprehensive assemblies of cDNA- and EST-genome alignments using tools such as the BLAST-Like Alignment Tool BLAT (Kuhn et al. 2007). Annotators also make extensive use of Ensembl, which provides predicted gene sets (including alternatively spliced forms) based on mRNA and protein sequences for the genomes of a range of vertebrate species and model organisms (Hubbard et al. 2007). Following manual sequence analysis, all erroneous gene model predictions, incorrect CDS assignments, and errors such as frameshifts are corrected, and all splice variants are explicitly described (Fig. 3). Most of the sequence analysis and proteomic analysis tools available at the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) will automatically incorporate all possible annotated splice isoforms in UniProtKB/Swiss-Prot in their searches (Gasteiger et al. 2003). Thus, the extensive manual curation of sequences performed by UniProtKB/Swiss-Prot results in a comprehensive resource of high quality protein sequences, including alternative protein isoform sequences where available.

4.1.2 Computational sequence annotation in UniProtKB/Swiss-Prot

Following sequence analysis and correction, a variety of computational tools are used to analyse the corrected protein sequence and predict potential features of interest. Computational tools facilitate protein classification and may provide information about the putative function of uncharacterized proteins or those that have no similarity to existing characterized protein sequences. The presence of specific modular domains or family membership is determined using InterPro resources (which are discussed later in this chapter) available through InterProScan (Zdobnov and Apweiler 2001). Other features such as transmembrane domains, secretory signals and organellar targeting signals are also predicted. All predicted features, except for specific domains and repeats, are flagged as “Potential”, to indicate their origin from computational analysis.

4.1.3 Functional annotation in UniProtKB/Swiss-Prot

Following sequence curation and analysis, each protein is annotated using information manually extracted from the literature and from specialist databases. We previously



Fig. 2 Annotation of a UniProtKB/Swiss-Prot entry. Selected annotations from the UniProtKB/Swiss-Prot entry Q9NII1 (ADAR_DROME). (a) An extract from the “General annotation (Comments)” section showing information on protein function and expression patterns. (b) An extract from the “Sequence annotation (Features)” showing specific domains, active sites, and ligand-binding residues. (c) An extract from the “Ontologies” section showing UniProtKB keywords (displayed under specific category headings according to their keyword type) and terms from the controlled vocabularies of the Gene Ontology (GO) (displayed under specific category headings corresponding to the GO sub-ontologies)

a

Sequence caution [AAM89124.1](#) sequence differs from that shown. Reason: Frameshift at position 686.

Cross-references Hide | Top

Sequence databases

EMBL [AF395746 mRNA. Translation: AAM89124.1. Frameshift](#)
[BC083546 mRNA. Translation: AAH83546.1.](#)

b

Alternative products Hide | Top

This entry describes 7 isoforms produced by **alternative splicing** and **alternative initiation**. [\[Align\]](#) [\[Select\]](#)

Isoform C (identifier: [Q9NII1-1](#))
 Also known as b,
This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

Isoform A (identifier: [Q9NII1-2](#))
 Also known as a,
The sequence of this isoform differs from the canonical sequence as follows:
 154-191 GIENLSSSKMFEIIQTMLTEKLSNPTSLEQPTFCMSQN → D

Sequences Hide | Top

Sequence	Length	Mass (Da)	Tools
<input type="checkbox"/> Isoform C (b) [UniParc] Last modified February 1, 2005 Version 2 Checksum 6570C16D25D1C9F4 Show >	676	74.978	<input type="text" value="Blast"/> <input type="button" value="go"/>
<input type="checkbox"/> Isoform A (a) [UniParc] Checksum B1FD5629F6836008	639	70.832	<input type="text" value="Blast"/> <input type="button" value="go"/>

```

10      20      30      40      50      60
MKFDSRVMLN SANNNSPOHP VSAPSDINMN GYNRKLQKR  GYEMPKYSDP KKKHCKERIP
70      80      90     100     110     120
QPKNTVAMLN ELRHLGIYKL ESQTGPVHAF LFTISVEVDG QKYLGGGRSK KVARIEAAAT
130     140     150     160     170     180
ALRSF IQFKD  GAVLSPLKPA GNLDFTSDEH LENDVSRSAI TVDGGKQKVPD KGPVHLLLYEL
    
```

Fig. 3 Sequence curation in UniProtKB/Swiss-Prot. (a) A “Sequence caution” describing a CDS that differs from the sequence displayed due to a frameshift in the underlying EMBL/GenBank/DDBJ entry. The precise position of the frameshift is given. The cross-reference to the original EMBL entry is flagged to indicate the error type. (b) An extract from an “Alternative products” subsection describing seven protein isoforms produced by a mixture of alternative splicing and alternative initiation. Each description lists all differences between the isoform sequence and that displayed. Due to space limitations, only two descriptions are shown. An extract from the corresponding “Sequences” section, which contains all isoform sequences, is also shown. Each individual sequence can be displayed or hidden according to the wishes of the user

defined functional annotation as a synthesis of information on individual protein function (including pathological functions associated with disease states), structure, modifications, interactions with other proteins, and participation in higher order

biological pathways. UniProtKB/Swiss-Prot aims to capture information on all these aspects of protein biology from a variety of sources. These are detailed in the following sections.

4.1.4 Annotation of protein structure in UniProtKB/Swiss-Prot

The three dimensional (3D) structure of a protein can reveal an enormous amount of mechanistic and functional information about the protein concerned. 3D structures shed light on the architecture of individual proteins and protein assemblies. They provide detailed information about the interactions of proteins and their ligands (substrates, ions, cofactors or regulatory molecules), and contribute to the elucidation of mechanisms of enzyme catalysis and the identification of active site residues. They also show post-translational modifications and can serve to demonstrate the precise effects of disease causing mutations.

UniProtKB/Swiss-Prot records contain manually annotated information from protein structure databases such as the Molecular Structure Database (e-MSD) (Tagari et al. 2006), one of the three partners of the world wide Protein Data Bank (wwPDB), which includes protein structures determined by X-ray crystallography, NMR and 3D electron microscopy (Berman et al. 2007). UniProtKB/Swiss-Prot extracts a variety of information from wwPDB records including the positions of sites of interaction with cofactors, metal ions, and other ligands or interacting proteins. UniProtKB maintains explicit links to all source wwPDB records and to other resources for protein 3D-structures, such as the Swiss Model Repository (SMR) (Kopp and Schwede 2006). The SMR provides access to over one million homology-based models of 3D-structure for proteins that share significant sequence similarity to at least one experimentally determined 3D protein structure (Kopp and Schwede 2006). This allows users easy access to protein models when experimentally determined structures are not available.

4.1.5 Annotation of post-translational modifications in UniProtKB/Swiss-Prot

Post-translational modifications (PTMs) regulate a variety of aspects of protein biology including cellular location (e.g. protein lipidation), protein-protein interactions (e.g. N-glycosylation) or the activation state of the protein (e.g. phosphorylation). Over 300 types of protein modification are currently known, and new types are reported each year. UniProtKB/Swiss-Prot annotates PTMs described in the literature, including small-scale experiments and high-throughput mass spectrometry studies that permit the identification of PTMs on hundreds or thousands of proteins (Olsen et al. 2006). Data is also extracted from the 3D-structure database wwPDB (Berman et al. 2007), and from specialist PTM databases such as GlycoSuiteDB

(Cooper et al. 2003). Specific predictors are used to identify potential sites for several PTMs, which are tagged as “Potential”. Individual PTMs are annotated using a strict controlled vocabulary with mappings to that of the RESID Database of Protein Modifications (Garavelli 2004). RESID is a comprehensive resource of annotations and structures for protein modifications, including systematic nomenclature, atomic formulas and masses, structural models and source annotations for UniProtKB (Garavelli 2004).

4.1.6 Annotation of protein interactions and pathways in UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot stores information on protein interactions from the literature and from structural databases such as wwPDB (Berman et al. 2007). Protein interaction data is also imported directly from the IntAct database, a central repository for storing and accessing information on protein interactions (Kerrien et al. 2007). IntAct and UniProtKB maintain reciprocal links between records describing proteins and the interactions in which they participate.

Individual proteins and assemblies of proteins form higher order biological pathways, such as metabolic and signal transduction pathways. UniProtKB/Swiss-Prot curators annotate biological pathways using the specific data model and controlled vocabularies of the UniPathway database (<http://www.grenoble.prabi.fr/obiwarehouse/unipathway>). UniPathway stores information on metabolic pathways, reactions, and the chemical entities involved in them. UniProtKB records precise information about the biological pathway and particular step(s) at which each protein participates. UniProtKB also maintains links to Reactome, a knowledgebase of pathways authored by expert biological researchers (Vastrik et al. 2007) and to BioCyc, a collection of pathway/genome databases (PGDBs) plus the BioCyc Open Chemical Database (Karp et al. 2005). Reactome focuses on human proteins but also projects human pathways onto orthologs in other organisms.

4.1.7 Annotation of human sequence variants and diseases in UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot places a special emphasis on the annotation of human sequence variants. Genetic variation plays a crucial role in determining phenotypic variation and disease susceptibility. The human genome has around 10 million “polymorphisms”, which are genetic variants in which the minor gene forms occur at least once in at least 1% of the population. Cataloguing these variants, their associations and effects, provides a means of identifying common risk factors for human diseases and may ultimately improve human health.

UniProtKB/Swiss-Prot imports data on human variants from the NIEHS Environmental Genome Project (EGP) (<http://egp.gs.washington.edu/>), the Seattle SNP program (<http://pga.gs.washington.edu/>), and from dbSNP, the Single Nucleotide Polymorphism database maintained by the NCBI (<http://www.ncbi.nlm.nih.gov/SNP/>). dbSNP incorporates data from a large number of sources including individual researchers and large consortia such as HAPMAP (The International HapMap Consortium 2005). Every SNP imported into UniProtKB/Swiss-Prot is assigned a unique identifier, which facilitates the implementation of reciprocal links to these and other resources. All human variants from UniProtKB/Swiss-Prot are automatically mapped onto the 3D structures of similar sequences (where these are available), allowing the likely effect of a given variant on the 3D structure of a protein to be determined (Yip et al. 2004).

All human sequence entries are cross-linked to the online version of MIM, the Mendelian Inheritance in Man database (McKusick 2007). MIM is an extensive catalogue of human genes and genetic disorders, and is considered to be one of the most comprehensive resources in the field of human genetics. Cross-links are also provided to GeneCards, a database of human genes, their products and their involvement in diseases (Safran et al. 2002) and to GeneLynx (Lenhard et al. 2001). Both resources are meta-databases providing extensive hyperlinks to human gene-specific information in diverse resources.

4.2 UniProtKB/TrEMBL

Due to the labor-intensive nature of manual curation, UniProtKB/Swiss-Prot cannot hope to keep pace with the current rate of production of protein sequences. UniProt therefore provides UniProtKB/TrEMBL, a computer annotated supplement to UniProtKB/Swiss-Prot.

UniProtKB/TrEMBL accommodates all protein sequence entries that are awaiting manual annotation and entry into UniProtKB/Swiss-Prot. It contains translations of all coding sequences present in the EMBL/GenBank/DDBJ nucleotide sequence databases, the sequences of PDB structures and data derived from amino acid sequences directly submitted to UniProtKB or extracted from the literature (UniProt Consortium 2007). Small fragments, synthetic sequences, non-germline immunoglobulins and T-cell receptors and most patent sequences are excluded from UniProtKB/TrEMBL. Sequences that are derived from the same organism and that are 100% identical over their entire length are merged. The current release of UniProtKB/TrEMBL (release 38.0 of 26th February 2008) contains 5,395,414 entries from 155,282 species. UniProtKB/TrEMBL is therefore currently around fifteen times larger than its manually annotated counterpart, UniProtKB/Swiss-Prot.

Each UniProtKB/TrEMBL record is enriched with high-quality annotation and classification that is performed using automatic annotation systems. These utilize manually annotated UniProtKB/Swiss-Prot entries as a source of annotations which

can be propagated to similar sequences in UniProtKB/TrEMBL (Fleischmann et al. 1999; Kretschmann et al. 2001; Wu et al. 2004). Together these processes raise the level of annotation of UniProtKB/TrEMBL above that of a simple sequence repository, and close to that of the manually annotated gold standard UniProtKB/Swiss-Prot.

5 Protein family classification for functional annotation

While many biologists work on single genes or proteins at a time, with the increase in genome sequence data available, it is now possible to study multiple proteins or even whole genomes as well as protein families across different genomes. Grouping related proteins into families has, for a long time, provided a means for not only making sense of large sequence datasets, but also for using these groupings for inferring function. If the biological function of one or more family members has been studied, this function can be inferred for other closely related protein sequences. Therefore the process of protein family classification is important in functional annotation. In this section we describe the methods of protein family classification, some example databases, and how these can be used for functional annotation.

5.1 Protein signature methods and databases

Traditionally, protein family members were derived through sequence similarity searches, however, though powerful, these have their limitations. They are not ideal for finding distantly related sequences, are influenced by database size and bias, and may provide inaccurate results for multifunctional proteins. Sequence similarity searches are now used as the starting point for more sophisticated methods of sequence classification, such as protein signatures. Protein signatures are mathematical descriptions of protein families that are derived from multiple sequence alignments of known members of a protein family. When related sequences are aligned, areas of conservation can be highlighted and emphasized in the creation of the signatures. Protein signatures may be generated using different methods, e.g. regular expressions, profiles or hidden Markov models.

5.1.1 Regular expressions and PROSITE

Regular expressions are computational tools for searching and retrieving data that can be defined very specifically. In protein signatures, regular expressions are used to describe small, highly conserved regions. The expression provides information on which amino acids should be present at the different positions of the conserved motif based on a given alignment. PROSITE (Hulo et al. 2006) is a database of both regular expressions (also known as patterns) and profiles and has a primary focus on signatures

for the annotation of UniProtKB/Swiss-Prot proteins. The regular expressions cover a number of important and well-characterized active sites and binding sites, and many have a corresponding profile covering a larger region of the protein family or domains. Regular expressions, while useful for identifying highly conserved sites, have limitations arising from their lack of adequate flexibility. A match is either positive or negative and thus how much of the region was changed is not reflected in the results. This is the reason that PROSITE develop profiles to provide additional confirmation of pattern matches (Hulo et al. 2006).

5.1.2 Profiles and the PRINTS database

Profiles overcome the limitations of regular expressions by covering a larger region of the sequence alignment, and enabling flexibility in conservation across that region. A profile is a table of position-specific amino acid weights and gap costs. The table describes the probability of finding an amino acid at a given position in the sequence (Gribskov et al. 1990), and these probabilities are used to calculate similarity scores between a profile and a sequence for a given alignment. For each profile, a threshold score is calculated and used to determine a positive match. The flexibility of profiles enables sequences with low scores in some amino acids to still be matched when the scores are high enough across the remaining sequence.

As mentioned above, profiles are used by the PROSITE database, but a variation on this method is also used by the PRINTS database (Attwood et al. 2003). PRINTS generates “fingerprints”, or sets of position-specific scoring matrices covering the most conserved regions diagnostic of each protein family. The PRINTS database has a strong focus on developing methods for classifying proteins on different levels of the protein family hierarchy, and provides some family hierarchies with several levels of depth. The database has good coverage of G-protein coupled receptors, ion channels and other families of pharmaceutical interest.

5.1.3 Hidden Markov Models (HMM) and HMM databases

A Hidden Markov Model (HMM), is a statistical model based on probabilities. It describes a set of states, which each have a probability distribution associated with it. HMMs are generated from multiple sequence alignments, and, like profiles, have threshold scores associated with them to discriminate between positive and negative matches (Krogh et al. 1994; Durbin et al. 1998). HMMs are the most common method for generating protein signatures due to their accuracy and ability to identify distantly related members of a protein family. They can cover conserved domains or the entire protein sequence. Examples of protein signature databases using HMMs include Pfam (Finn et al. 2006), SMART (Letunic et al. 2006), TIGRFAMs (Selengut et al. 2007), PIRSF (Wu et al. 2004) and PANTHER (Mi et al. 2007).

Pfam and SMART focus on generating signatures for protein domains. The latter has a specific focus on cell signaling, extracellular and chromatin-associated domains, while Pfam covers all areas of biology, including proteins of unknown function. PIRSF and PANTHER develop HMMs to describe protein families and try to cover the full sequence. A requirement of PIRSF HMMs is that they cover the full sequence and group proteins with identical domain compositions. TIGRFAMs generate both family and domain HMMs, although focus more on the former, and are tailored for use in the annotation of microbial genomes.

5.1.4 Structure-based protein signature databases

The protein signature databases described above use sequence-based families for developing their protein signatures. There are also, however, two protein signature databases that derive their families or domains from protein structural families. The SCOP (Andreeva et al. 2004) and CATH (Greene et al. 2007) databases generate protein structural family hierarchies from solved structures deposited in the Protein Data Bank (PDB) (Berman et al. 2007). Corresponding protein sequences from members of these structural families are then used as the basis for generating protein signatures. The SUPERFAMILY database (Wilson et al. 2007) creates HMMs from SCOP superfamilies, while Gene3D (Yeats et al. 2006) HMMs are based on CATH superfamilies. SUPERFAMILY and Gene3D HMMs often correlate well with some of the sequence-based HMM databases, such as Pfam, but some families differ due to the wider variation in sequence conservation for structural compared to sequence-based families.

5.1.5 ProDom sequence clustering method

In addition to the protein signature databases and methods described above, an alternative method of protein family classification is through sequence clustering. This is achieved through pairwise sequence similarity calculation and subsequent clustering using one of a variety of clustering algorithms. ProDom (Bru et al. 2005) is an example of a database that provides protein families and domains via sequence clustering. Although databases such as ProDom achieve very high coverage, the resulting clusters usually have little or no biological annotation and no manual intervention.

5.2 InterPro – integration of protein signature databases

To overcome some of the limitations of individual protein signature methods and databases, the groups described here (Pfam, PRINTS, PROSITE, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D and PANTHER) have collaborated to integrate their data into a single resource, InterPro (Mulder et al. 2007). InterPro (<http://www.ebi.ac.uk/interpro>), an integrated resource for protein families,

domains and functional sites, tries to rationalize where protein signatures from the different databases are describing the same family, domain or functional site, and merges these into single InterPro entries. Where the overlap in proteins matched is not exact, relationships are created between InterPro entries to provide family hierarchies (parent-child relationship) and to describe domain composition (contains-found in relationship). InterPro entries are annotated with a name, short name, and abstract describing the protein family/domain. Cross-references are provided to the original protein signatures, GO terms (Gene Ontology Consortium 2006), protein-protein interaction data, protein structural information, and specialized protein function databases.

All protein sequences in UniProtKB are run through the protein signatures in InterPro to provide protein match lists for each InterPro entry. The set of protein sequences matched can be viewed from the entry in a number of different formats. In the detailed graphical view, each protein signature hit is displayed, colored by originating database, and showing the positions of the matches along the sequence. The graphical overview shows the condensed matches to each InterPro entry, and the table view provides a tabular display of the matches to the signatures within a single InterPro entry. An InterPro Domain Architectures view provides a summary of the different architectures present in each entry. The graphical views also display structural matches, showing the positions on the sequence of solved and modeled structures and structural families from SCOP and CATH.

5.3 Using InterProScan for sequence classification and functional annotation

5.3.1 InterProScan

The classification of proteins into families begins by identifying which protein signatures match a sequence or set of sequences, and this is achieved using the InterProScan package (Quevillon et al. 2005). InterProScan integrates the algorithms from the member databases into a single package and provides the results in a single format with additional information derived from the InterPro entries matched. There are different versions of InterProScan, depending on the user requirements. The package is available for searching single sequences at a time through a web interface at <http://www.ebi.ac.uk/InterProScan/>. The default output display is a graphical view of the matches showing the signatures and corresponding entries hit and providing links to these entries (see Fig. 4a). The results are also available in raw and XML format and as a tabular view. The latter includes additional information such as InterPro relationships and GO terms (Fig. 4b).

For users requiring bulk searches there are two additional options. If confidentiality is required, the user can download and install InterProScan to run locally. Alternatively,

bulk InterProScan runs can be submitted programmatically *via* SOAP-based web services at the EBI. For more information see: <http://www.ebi.ac.uk/Tools/webservices/services/interproscan>.

5.3.2 Interpreting InterProScan results

A basic knowledge of InterPro is required to enable easy interpretation of InterProScan results. In the example shown in Fig. 4, it is clear that the query protein is a kind of DNA helicase, and one can follow the links to the InterPro entries to find out more about the protein family or domain hit (Example entry shown in Fig. 5). There are three InterPro entries matches by the query sequence as well as two unintegrated SUPERFAMILY signatures. These unintegrated signatures are awaiting manual integration into InterPro, but are still available in InterProScan. Of the three entries matched, IPR007692

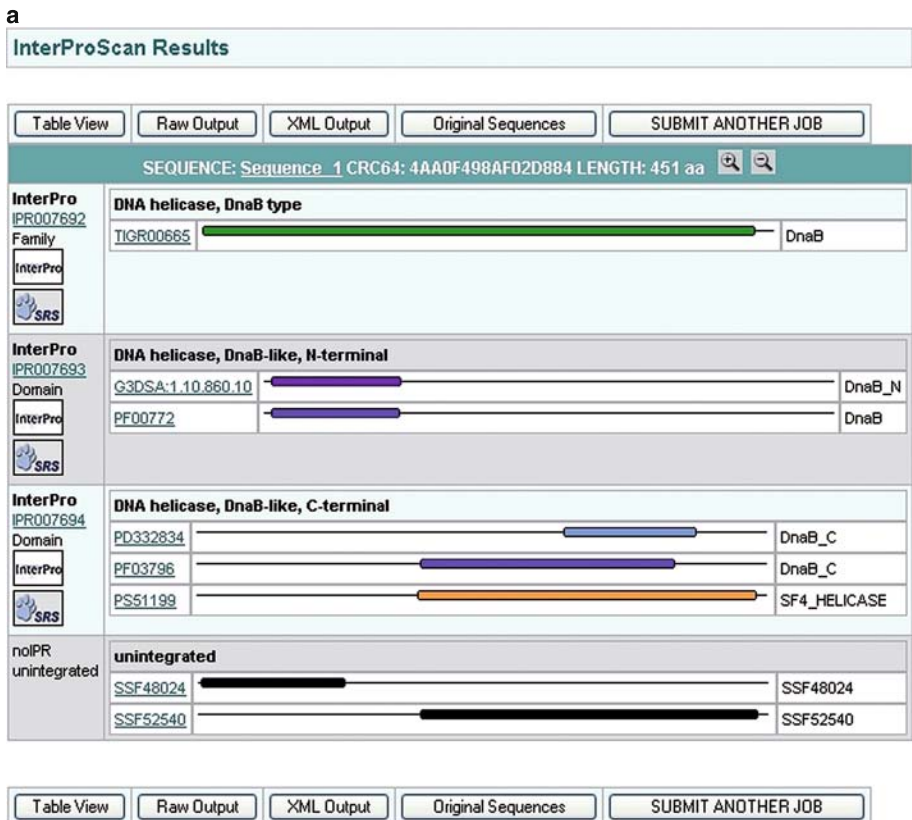


Fig. 4 Example output of the InterProScan package. Results are shown in (a) graphical and (b) table views. A query sequence was run through all the protein signatures in InterPro; matches were identified to InterPro entries relating to DNA helicases

b







InterProScan Results				
Picture View Raw Output XML Output Original Sequences SUBMIT ANOTHER JOB				
SEQUENCE: Sequence_1 CRC64: 4AA0F498AF02D884 LENGTH: 451 aa				
InterPro IPR007692	DNA helicase, DnaB type			
Family	TIGRFAMs	TIGR00865	<i>DnaB</i>	0.0 [3-436]T
				
				
Parent	no parent			
Children	no children			
Found in	no entries			
Contains	IPR003593 IPR007693 IPR007694			
GO terms	Molecular Function: DNA binding (GO:0003677) Molecular Function: DNA helicase activity (GO:0003678) Molecular Function: ATP binding (GO:0005524) Biological Process: DNA replication (GO:0006260)			
InterPro IPR007693	DNA helicase, DnaB-like, N-terminal			
Domain	GENE3D	G3DSA:1.10.860.10	<i>DnaB_N</i>	7.69998480533761E-33 [8-110]T
	PFAM	PF00772	<i>DnaB</i>	6.499979505417751E-42 [8-109]T
				
Parent	no parent			
Children	no children			
Found in	IPR007692			
Contains	no entries			
GO terms	Molecular Function: DNA helicase activity (GO:0003678) Molecular Function: ATP binding (GO:0005524) Biological Process: DNA replication (GO:0006260)			
InterPro IPR007694	DNA helicase, DnaB-like, C-terminal			
Domain	PRODOM	PD332834	<i>DnaB_C</i>	0.0 [291-395]T
	PFAM	PF03796	<i>DnaB_C</i>	1.09999184471398E-98 [179-378]T
	PROFILE	PSS1199	<i>SF4_HELICASE</i>	0.0 [176-443]T
Parent	no parent			
Children	no children			
Found in	IPR007692			
Contains	IPR003593			
GO terms	Molecular Function: DNA helicase activity (GO:0003678) Molecular Function: ATP binding (GO:0005524) Biological Process: DNA replication (GO:0006260)			
noIPR unintegrated	unintegrated			
	SUPERFAMILY	SSF48024	<i>SSF48024</i>	1.7999975402237803E-35 [5-118]T
	SUPERFAMILY	SSF52540	<i>SSF52540</i>	5.30001985225048E-44 [178-442]T
Parent	no parent			
Children	no children			
Found in	no entries			
Contains	no entries			
GO terms	none			

Fig. 4 (continued)

InterPro: IPR007692 DNA helicase, DnaB type

Protein matches

Overview: sorted by AC, of known structure, proteins with splice variants
 Detailed: sorted by AC, of known structure, proteins with splice variants
 Table: For all matches, sorted by AC, of known structure

UniProtKB Matches: 760 proteins
 Accession List

Accession: IPR007692 DNA_helicase_DnaB
Secondary: IPR001198

Type: Family

Database ID Name Proteins
 TrEMBL TR060864 DnaB 760

InterPro Relationships

Contains: IPR003593 AAA+ ATPase core
 IPR007693 DNA helicase, DnaB-like, N-terminal
 IPR007694 DNA helicase, DnaB-like, C-terminal

GO Term annotation:
 Process: GO:0006260 DNA replication
 Function: GO:0003737 DNA binding
 GO:0003738 ATPase activity
 GO:0005524 ATP binding

InterPro annotation

Abstract: This family includes the replicative DNA helicase, helicase DnaB, which exhibit DNA-dependent ATPase activity. Helicase DnaB is a homohexameric protein required for DNA replication. The homohexameric can form a ring around a single strand of DNA near a replication fork. An inborn of more than 400 residues is found at a conserved location in DnaB of *Synechocystis* PCC6803, *Rhodothermus marinus* (both experimentally confirmed), and *Mycobacterium tuberculosis*. The inborn removes (seal) by a self-splicing reaction. Replication protein, GP12, from *Escherichia coli* also belongs to this family.

Structural links: CATH: 1.10.860.10.1, 7.170.16.10.5
 SCOP: 4.1.1.1, 4.86.12

Database links: EBI-EMBL, EBI-ENZYME
 GUST: ALCVALL322476.2
 Enzyme: EC:3.6.1

Taxonomic coverage

4

Overlapping InterPro entries

InterPro ID	Numbers of overlapping proteins	Average numbers of overlapping amino acids
IPR007692	0	N/A
% Overlap: 100	760	518
IPR007693	1	N/A
% Overlap: 99	759	368

describes the DnaB-type DNA helicase family, which contains the DnaB-like DNA helicase N- and C-terminal domains. These relationships are displayed in the entries and in the table view of the InterProScan output. These sources also provide potential GO annotations through the InterPro to GO links (Fig. 4b). In this way, InterProScan provides an automatic means of GO annotation.

By following the links to the InterPro entries matched by the query sequence, one can identify other proteins matching the same entry, other proteins with the same domain architecture, and proteins within the family/domain that have a solved structure. The entry also provides the taxonomic range of its members and has links, in this example, to the Enzyme database and relevant publications. In this way, InterProScan and its links can provide a host of functional information on the query sequence and its family members.

5.3.3 Large-scale automatic annotation

The use of InterPro for automatic annotation of UniProtKB/TrEMBL is mentioned in Sect. 2.5 above. In summary, InterPro entries/matches are used as a means of grouping proteins with related functions. For all UniProtKB/Swiss-Prot proteins matching an entry or set of entries, the common annotation is determined. Annotation rules are generated, in which the criteria that must be met are the InterPro entries (or their signatures) matched, and the rule applied is the addition of the common annotation to the query proteins. This kind of automatic annotation is also used by groups outside the EBI such as the genome annotation projects. In these projects, InterProScan results are used for automatic annotation to supplement and confirm sequence similarity methods and to provide annotation where no relevant annotated matches were found in the similarity searches.

In addition to functional annotation based on UniProtKB/Swiss-Prot information, InterPro and InterProScan are used for large-scale GO annotation. Manual GO annotation, though of very high quality, is slow and time-consuming. The annotation requires curators to read all literature on the protein in question and assign the most appropriate GO terms, together with evidence codes, which indicate the evidence for the annotation. Examples of manually curated evidence codes include “Traceable author statement”, “Inferred from direct assay”, etc. Manual GO annotation is done by some of the UniProt curators, by curators in the Gene Ontology Annotation group at the EBI (specific focus on human proteins), and by curators at the model organism databases, however, the total number of proteins with manual GO annotation is still very low. In order to overcome the problem of low coverage of manual GO annotation and



Fig. 5 Example of an InterPro entry. This entry, IPR007692, describes the DnaB-type DNA helicase family. Only the top of the entry is shown, the rest of the entry includes example proteins and a list of publications cited in the abstract

provide a high-throughput prediction of GO annotations, automatic methods use GO mapping files. InterPro2GO links are an example of GO mappings and other examples include mappings between Swiss-Prot keywords and GO terms, EC numbers and GO terms, etc. Any GO terms assigned using these mapping files are given the GO evidence code “Inferred from Electronic Annotation”, to indicate that the inference was via electronic means. InterPro2GO and other electronic GO mapping account for the largest number of GO annotations world-wide. The integration of GO mapping in the widely-used InterProScan tool facilitates automatic functional annotation of large sets of new protein sequences.

The value of GO annotation is enormous, as it enables not only additional functional annotation of proteins (the user can see at a glance the protein’s function, the biological process it is involved in and the cellular component in which it performs the function), but also the easy retrieval and comparison of annotated data.

6 From genes and proteins to genomes and proteomes

Annotated protein-centric databases such as UniProtKB provide a detailed summary of protein function for individual proteins, while InterPro covers protein families and domains. Numerous annotation pipelines draw on resources such as UniProtKB and InterPro to extend the functional annotation of individual protein sequences and domains to the level of whole genomes or proteomes. Two such annotation pipelines are Genome Reviews, which provides an integrated view of complete genome and proteome data from archaea, bacteria, bacteriophage, and selected eukaryota (Kersey et al. 2005), and Restauro-G, which performs a similar function for bacteria (Tamaki et al. 2007). Both resources are founded on the premise that the quality of annotations associated with original genome sequence submissions to the International Nucleotide Sequence Database can be enhanced by supplementary annotations drawn from external databases such as UniProtKB.

Genome Reviews supplements completed genome sequences from EMBL with standardized annotations from UniProtKB and GO terms and removes sequences identified as erroneous by UniProtKB/Swiss-Prot (Kersey et al. 2005). Restauro-G employs both UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entries as annotation sources, drawing supplementary information from other resources such as Pfam (Finn et al. 2006). Data from Genome Reviews can be interrogated using the Integr8 portal, which offers a variety of statistical analyses of individual complete proteomes as well as numerous pre-computed proteome comparisons (Kersey et al. 2005). For each proteome users can obtain lists of the most common InterPro families, domains or repeats, and a list of high level GO annotations from GO-Slim (Kersey et al. 2005). In addition to Genome Reviews, Integr8 uses data from non-redundant sets of UniProtKB entries representing each complete proteome. For some eukaryotic organ-

isms, appropriate proteome sets may be obtained by filtering UniProtKB using information from model organism databases such as Flybase (Crosby et al. 2007). For higher eukaryotes, proteome sets are derived from the International Protein Index (IPI), which combines data from UniProtKB (UniProt Consortium 2007), Ensembl (Hubbard et al. 2007) and RefSeq (Pruitt et al. 2007) to form non-redundant sets of protein sequences (Kersey et al. 2004). For each individual proteome set additional information is added from a host of resources including HAMAP (Gattiker et al. 2003), InterPro (Mulder et al. 2007), and the Gene Ontology Annotation database (Camon et al. 2004). By combining these resources, Integr8 provides easy access to integrated information about complete genomes and their corresponding proteomes.

7 Summary

Most functional annotation of new protein sequences proceeds by the identification of related proteins, or modular protein domains or motifs within the protein sequence, followed by the transfer of their associated annotations to the protein sequence of interest. In order for this approach to be successful, annotation pipelines require comprehensive sources of functional annotation and accurate means of protein classification. Universal manually annotated protein sequence databases such as UniProtKB/Swiss-Prot provide a comprehensive source of functional annotations for newly identified protein sequences, while InterPro integrates the major protein signature resources to create a powerful tool for protein and domain classification. Together these resources are integrated into a variety of methods for automatic functional annotation that allow the ever-growing avalanche of genome sequence data to be usefully exploited by biologists.

References

- Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32: D226–D229
- Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 31: 400–402
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2003) GenBank. *Nucleic Acids Res* 31: 23–27
- Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 35: D301–D303
- Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C, Canaran P, Chan J, Chen N, Chen WJ, Davis P, Fiedler TJ, Girard L, Han M, Harris TW, Kishore R, Lee R, McKay S, Muller HM, Nakamura C, Petcherski A, Rangarajan A, Rogers A, Schindelman G, Schwarz EM, Spooner W, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Durbin R, Stein LD, Sternberg PW, Spieth J (2007) WormBase: new content and better access. *Nucleic Acids Res* 35: D506–D510

- Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res* 33: D212–D215
- Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Res* 32: D262–D266
- Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res* 32: D311–D314
- 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
- Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res* 35: D486–D491
- Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK
- Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34: D247–D251
- Fleischmann W, Moller S, Gateau A, Apweiler R (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics* 15: 228–233
- Garavelli JS (2004) The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics* 4: 1527–1533
- 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
- Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* 34: D322–D326
- Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 35: D291–D297
- Gribkov M, Luthy R, Eisenberg D (1990) Profile analysis. *Methods Enzymol* 183: 146–159
- Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E (2007) Ensembl 2007. *Nucleic Acids Res* 35: D610–D617
- Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ (2006) The PROSITE database. *Nucleic Acids Res* 34: D227–D230
- Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 33: 6083–6089

- Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thornycroft D, Zhang Y, Apweiler R, Hermjakob H (2007) IntAct – open source resource for molecular interaction data. *Nucleic Acids Res* 35: D561–D565
- 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–D302
- Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 4: 1985–1988
- Kopp J, Schwede T (2006) The SWISS-MODEL repository: new features and functionalities. *Nucleic Acids Res* 34: D315–D318
- Kretschmann E, Fleischmann W, Apweiler R (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* 17: 920–926
- Krogh A, Brown M, Mian IS, Sjolander K, Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 235: 1501–1531
- Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pedersen JS, Hsu F, Hinrichs AS, Harte RA, Diekhans M, Clawson H, Bejerano G, Barber GP, Baertsch R, Haussler D, Kent WJ (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res* 35: D668–D673
- Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R (2004) UniProt archive. *Bioinformatics* 20: 3236–3237
- Lenhard B, Hayes WS, Wasserman WW (2001) GeneLynx: a gene-centric portal to the human genome. *Genome Res* 11: 2151–2157
- Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34: D257–D260
- Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 36: D475–D479
- McKusick VA (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* 80: 588–604
- Mi H, Guo N, Kejariwal A, Thomas PD (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res* 35: D247–D252
- Miyazaki S, Sugawara H, Gojobori T, Tateno Y (2003) DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Res* 31: 13–16
- Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database. *Nucleic Acids Res* 35: D224–D228
- Olsen JV, Blagoev B, Gnadt F, Macek B, Kumar C, Mortensen P, Mann M (2006) Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell* 127: 635–648
- Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35: D61–D65

- Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. *Nucleic Acids Res* 33: W116–W120
- Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Olender T, Chalifa-Caspi V, Lancet D (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* 18: 1542–1543
- Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O (2007) TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* 35: D260–D264
- Stoesser G, Baker W, van den Broek A, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, Nardone F, Stoehr P, Tuli MA, Tzouvara K, Vaughan R (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res* 31: 17–22
- Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23: 1282–1288
- Tagari M, Tate J, Swaminathan GJ, Newman R, Naim A, Vranken W, Kapopoulou A, Hussain A, Fillon J, Henrick K, Velankar S (2006) E-MSD: improving data deposition and structure quality. *Nucleic Acids Res* 34: D287–D290
- Tamaki S, Arakawa K, Kono N, Tomita M (2007) Restauro-G: a rapid genome re-annotation system for comparative genomics. *Genomics Proteomics Bioinformatics* 5: 53–58
- The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437: 1299–1320
- UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res* 35: D193–D197
- Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 8: R39
- Wilson D, Madera M, Vogel C, Chothia C, Gough J (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res* 35: D308–D313
- Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, Ledley RS, Suzek BE, Arminski L, Chen Y, Zhang J, Cardenas JL, Chung S, Castro-Alvarez J, Dinkov G, Barker WC (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res* 32: D112–D114
- Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S, Orengo CA (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res* 34: D281–D284
- Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum Mutat* 23: 464–470
- Zdobnov EM, Apweiler R (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17: 847–848

CHAPTER 4.2

Annotating bacterial genomes

C. Médigue^{1,2} and A. Danchin³

¹CNRS UMR8030, Génomique Métabolique, Evry Cedex, France

²Commissariat à l'Energie Atomique (CEA), Direction des Sciences du Vivant, Institut de Génomique, Genoscope, Laboratoire de Génomique Comparative, Evry Cedex, France

³Génétique des Génomes Bactériens, URA2171 CNRS, Département Génomes et Génétique, Institut Pasteur, Paris, France

1 Background

Since the mid-eighties, laboratories world-wide have endeavoured to determine the complete sequence of genomes from all kinds of living organisms. The first complete sequence of DNA bacteriophage Φ X174 appeared in 1978 (5386 nt (Sanger et al. 1978)), followed by that of bacteriophage lambda, using a shotgun technology, published in 1982 (48,502 bp) (Sanger et al. 1982) and that of a short bacterial genome (*Haemophilus influenzae* 1,830,138 bp, using a scaling up of the same shotgun technique) was published in 1995 (Smith et al. 1995). Curiously, while enormous amounts of funding were involved in sequencing, only comparatively small effort and support has been devoted to the creation of high-quality genome annotations, in particular in the creation of the very important link between experimental validation of *in silico* (Danchin et al. 1991) predictions and annotations. Indeed, explicit reference to experimental validation of annotation has only been recently introduced at the International Nucleotide Sequence Database Collaboration (INSDC <http://www.insdc.org/>).

In silico studies can be broadly divided into four branches, all pertaining to some feature of the annotation procedure:

- *Data acquisition*: this includes signal treatment and image analysis (for sequencing, and in 2D electrophoresis proteins studies, for example). Reference to processes involved at this early step may have to be annotated;

Corresponding author: Antoine Danchin, Génétique des Génomes Bactériens, URA2171 CNRS, Département Génomes et Génétique, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France (e-mail: antoine.danchin@normalesup.org)

- *Data analysis*: this is the most important part, in terms of scientific goals, if not in terms of personnel needed to achieve the goals. This is where annotation is of prime importance, reflecting the outcome of analyses produced by an ever increasing number of approaches;
- *Data management*: this is a domain that is often overlooked. However it is easy to understand that as we are flooded by data, data management is essential. It is utterly impossible to explore the data without a deep understanding of the way to manage and mine them. This domain is usually not a domain of interest for biologists (it can hardly lead to a biological question), although used to create resources that are heavily used, but this is a domain of importance for computer scientists and annotation is deeply rooted in the variety of databases, for example, that collect sequence related data;
- *Man-machine interaction*: presentation of the data, and of the results of analyses performed by computers has to be adapted to the human mind. This is not a trivial process. As for the preceding point, this is not a common domain of interest for biologists (except for those interested in the central nervous system and in psychology) but this is of major importance, and usually neglected. The best annotation platforms have this step in mind.

While much emphasis has been placed on the Human Genome Project, the development of genome programmes owed very much initially to the study of microbes. Indeed, the very first discovery of genomics, that contrary to expectation a considerable proportion of genes are of unknown function, was a contribution of the European Union (Elounda, Greece, EU meeting 1991) devoted to the study of two microbes, *Saccharomyces cerevisiae* and *Bacillus subtilis*. This immediately demonstrated that a considerable effort would be needed to try and annotate genes as new genome sequences would be produced (for a brief history, see (Danchin 2003)). At the time of writing, there are 680 publicly listed complete bacterial and archaeal genomes in the GOLD database (<http://www.genomesonline.org/>). In parallel, novel sequencing technologies (454 <http://www.454.com/>, Solexa-Illumina <http://www.illumina.com/>, SOLiD <http://solid.appliedbiosystems.com/>, Helicos <http://www.helicosbio.com/>, etc. . .) deliver a huge number of new sequences, both finished and draft genomes asking for continuous improvement of genome annotation procedures.

Beyond additional species, multiple strains of some bacteria are being sequenced, opening up the opportunity for detailed studies of genome evolution over the smallest time scales. Due to the public perception of the importance of infectious diseases and the desire to maximize benefits for human health and wealth, genome sequencing is considerably biased towards pathogens and organisms of economic consequence (ca. 80% according to the GOLD database). This bias is now being steadily balanced by interest in isolates from the environment, as well as

projects that aim at covering the tree of life more extensively and discovering new biological functions. The effort in collecting a large sample of genome sequences from a particular niche – metagenomics (Handelsman et al. 1998) – no longer aims at identifying the complete sequence of a given genome, but rather samples from many genomes simultaneously. This trend is now taken into account in annotation platforms.

In the present context “annotation” refers to the creation of explicit comments (organized along a data schema that needs to be specified: gene name, gene function, enzyme identifier, bibliographic reference, experimentally identified feature, etc.) locally associated to the name, or label of genomic objects (genes in particular).

Genome annotation first requires identification of genomic objects (which are defined by specific biological properties: e.g. promoters, genes, terminators, or structural properties: e.g. nucleotide repeats, curved regions, etc.). Furthermore, to be useful to investigators, annotation needs to create and collect indications about the intimate nature of biological systems, relationships between objects (Danchin 2003). In its most efficient form, annotation is meant to provide a help for inductive reasoning (making inferences), permitting investigators to make connections between objects and concepts, and to create experiments meant to validate them. “Making connections” is effective at a background level when one exploits to its most extended consequences the idea of “neighbourhood”. A first example of this

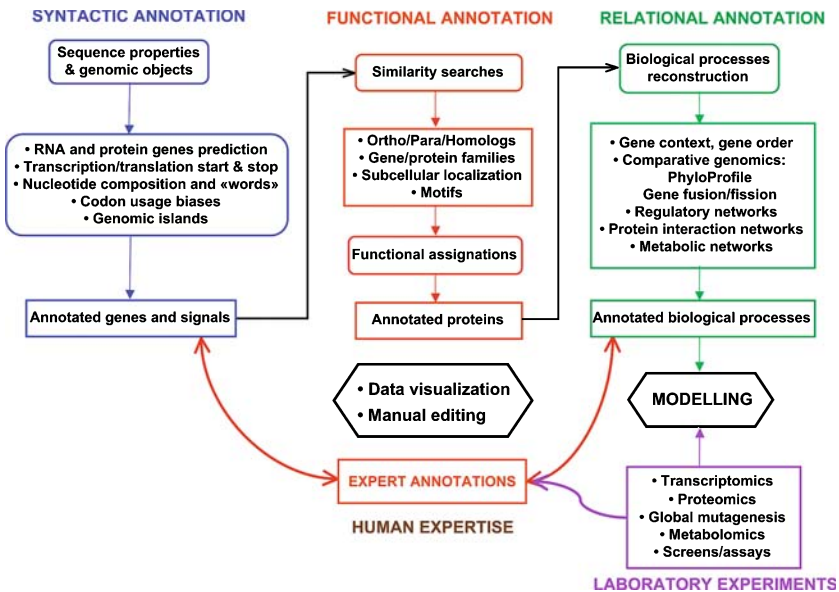


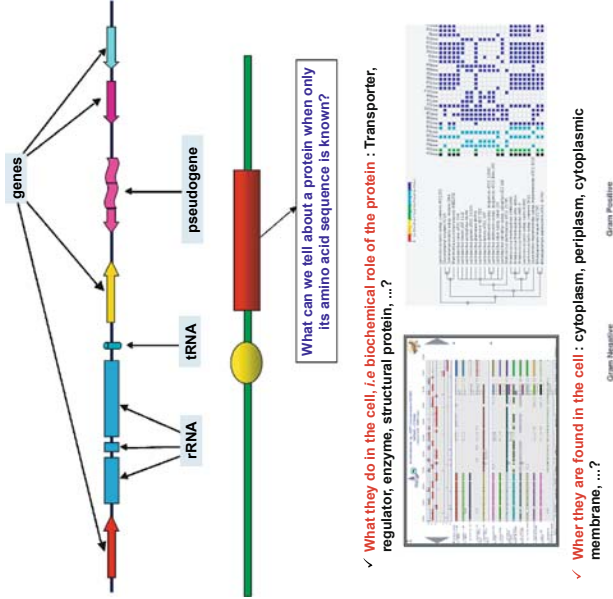
Fig. 1 Genome annotation workflow

Table 1 Genome-wide annotation

First round of annotation: *in silico* procedures

- Identification of protein-coding genes
- Identification of RNA genes
- Identification of approximate repeats in DNA
- Identification of genomic islands
- Computation of similarities of all predicted genes with those contained in Sequence/ Domains and motifs/Protein families/ Enzyme/Gene Ontology databases
- Computation of Bi-directional Best Hits (BBHs) between protein-encoding genes
- Detection of synteny blocks
- Phylogenetic profile analysis
- Prediction of subcellular localisation, transmembrane helices and signal peptide cleavage sites

- <http://exon.gatech.edu/genemark/>
- <http://www.genomics.jhu.edu/Glimmer/>
- <http://lowelab.ucsc.edu/TFNAscan-SE/>
- <http://www.cbs.dtu.dk/services/RNAmmer/>
- <http://www.wabi.snv.jussieu.fr/~public/RepSeek/>
- <http://www.pathogenomics.sfu.ca/islandpath/>
- <http://www.ebi.ac.uk/blast/>
- <http://www.ebi.ac.uk/InterProScan/>
- <http://www.ncbi.nlm.nih.gov/COG/old/xognotor.html>
- <http://bioinfo.genopole-toulouse.prd.fr/priam/>
- <http://bips.u-strasbg.fr/GOAnno/>
- <http://www.biocompare.com/products/pedantpro.php>
- <http://www.genoscope.cns.fr/agc/mage/>
- <http://www.igs.cnrs-mrs.fr/phydbac/>
- <http://www.psорт.org/psорт/>
- <http://www.cbs.dtu.dk/services/TMHMM/>
- <http://www.cbs.dtu.dk/services/SignalP/>



Metabolic networks reconstruction

<http://biocyc.org/> (pathologic)

<http://compbio.mcs.anl.gov/puma2/>

<http://wit.integratedgenomics.com/> (ERGO)

<http://www.igs.cnrs-mrs.fr/FusionDB/>

<http://string.embl.de/>

<http://www.ebi.ac.uk/intact/site/>

Gene fusion/fission Predicted protein-protein interaction

✓ When larger processes they participate in : metabolic pathways, signaling cascade, ...?



Second round of annotation: Manual annotation steps

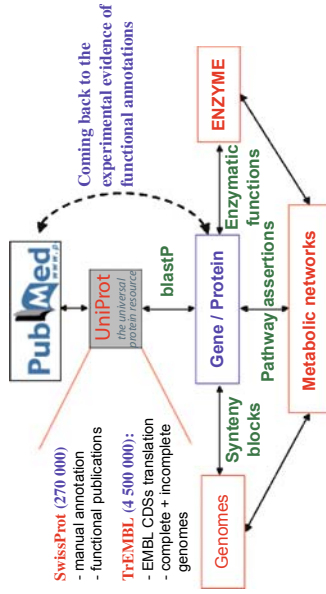
Identification of inconsistencies in the CDS start point

Evaluation of questionable assignments and cases of weak homology:

- Manual inspection of the differences between automatic functional assignments and annotations generated by motif/domain, COGs, etc databases
- History of each functional annotation to avoid percolation of annotation errors: is it associated to experimentally validated data?
- Examination of gene context and co-localisation of genes in several genomes (synteny blocks) to find previously unrevealed functional relations between genes

Detailed function annotations:

- Manual inspection of biological subsystems according to the organism 'life style'
- Improvement of the accuracy of genome annotation and identification of missing functions using neighbourhoods



Combinaison of cellular pathways and gene context analysis:

if most functions in a given pathway are connected to genes which are neighbors on the chromosome, they may yield a functional clue for ORFs without assigned function in this neighborhood.

fruitful idea was developed in the Entrez software at the NCBI (Benson et al. 1994), followed by many types of approaches combining phylogeny (sequence kinship, syntenies), with all kinds of proximities (local nucleotide or amino acid composition, isoelectric points, common metabolic pathways, co-occurrence in scientific articles etc) (Moszer et al. 1995; Nitschké et al. 1998; Marcotte et al. 1999; Huynen et al. 2000; Overbeek et al. 1999).

Annotation proceeds in several steps (Fig. 1). A good practice is to begin with analysis of the global properties of the sequence under investigation. This allows the annotator to place the collection of sequences, usually a genome sequence, in a particular context where it is possible to get a general idea of the biological features of the organisms in relation with their sequence. A second step will identify genomic objects, in particular Coding DNA sequences (CDSs) which encode proteins. Subsequently, functions will be associated to these objects via a variety of methods (Médigue and Moszer 2007). As their number is usually very large, it is convenient to proceed via the use of automatic procedures. In-depth annotation will follow, often combined with manual analysis of the most important objects, in a recursive way (Fig. 1 and Table 1). This is at this stage that functions are associated to relevant objects, which permit investigators to explain the behaviour of the organism (or collection of organisms) or to predict properties that can be submitted to experimental validation.

2 Global sequence properties

A first general analytical study of the global genome features needs to be implemented when annotating a new genome. This preliminary overview is useful to identify possible biases in the overall features of the genome, and it allows the annotator to locate loci that would be interesting for further in-depth studies. This step, which also permits one to have some insight about the quality of the sequence under annotation (remember that no biological experiment is error-free), is essential to place the genome in proper biological context, an important feature to construct internally consistent annotations.

A common first step in the study of the genomic context analyzes the GC content of the sequences as well as the distribution of words and motifs (Necsulea and Lobry 2007). It is essential here to proceed in a recursive way, comparing the real sequence with a realistic model constructed from previous knowledge. This allows identification of new signals, that are incorporated into a second level realistic model, and to progressively improve on the overall description of the sequence, permitting the annotator to get access to the identification of relevant genomic objects (see for example (Hénaut et al. 1996, 1998)). This analysis being recursive it progressively benefits from finer grain exploration of the sequence. In particular, as discussed below, CDSs do not constitute a homogeneous class in terms of codon usage biases (Médigue et al. 1991; Bailly-Béchet

et al. 2006), and this particular feature needs to be included when constructing progressively more realistic models of what a genome sequence is.

In this first exploration, the number of chromosomes and plasmids is determined as well as their origins and termini. This permits investigators to have an idea of replication biases and gives a first hint about the existence of recombination events or hot spots. Local variation in average properties can be the signature of the presence of genomic islands (GIs) and mobile elements (phages, integrons, insertion sequences, transposons) that have important biological roles. A number of methods to detect GIs have been developed (Hsiao et al. 2005). They rely mostly on the presence of atypical features of the sequence composition (GC-content, codon usage biases, repeats, etc.) or associated annotation features (for example tRNA and integrase genes). More recently, a new method based on compositional biases using variable order motif distributions has been published (Vernikos and Parkhill 2006). Comparative genomics can also be used: two or more sequences are aligned to identify unique genomic regions that are putative GIs. Once identified, the nature of the GIs (pathogenicity islands, prophages, symbiosis islands, metabolic islands, or other), and the nature of the exchanged genome, need to be characterised. A strategy calculates the local signature of identified GIs and look for similar signatures in a database of genomic signatures (Dufraigne et al. 2005); another makes use of phylogenetic tools to date the transfer event during evolution and to identify the origin of the suspected alien DNA fragments (Poptsova and Gogarten 2007). However, the identification of GIs' origin remains a difficult task and few examples have been validated.

An important discovery of the genomic era is that the gene pool of several strains of a given microbial species is far larger than the number of genes present in any single genome. General analysis of genomes, together with the ever-growing flow of new genes as one sequences new genomes has led to introduce the term of “pan-genome” to refer to all genes that may be found in a group of genomes, species, genus or clade (Tettelin et al. 2005). This general definition asks for a further in-depth analysis of the structure of genomes, taking into account their role in the life style of the organism. This can be summarized as follows. A genome separates persistent functions coded by genes that allow life to proceed from a large number of genes that permit life to develop in a particular environment. It is important to note that the class of persistent function comprises functions that are ubiquitous, but can be fulfilled by a variety of different objects recruited during the course of evolution either by *de novo* creation or by horizontal gene transfer, precluding extraction of their catalogue as the intersect of orthologs that would be present in all free living bacterial species (Fang et al. 2005b). These genes are connected in such a way that they could be seen as depicting a scenario of the origin of life, for which reason they have been named the paleome (palaios, ancient, in Greek). The corresponding functions permit the cell to express life, as well as fight weathering (aging processes). A second category of genes, apparently not limited in number, distributes among many genomes of the species. These genes, which permit life

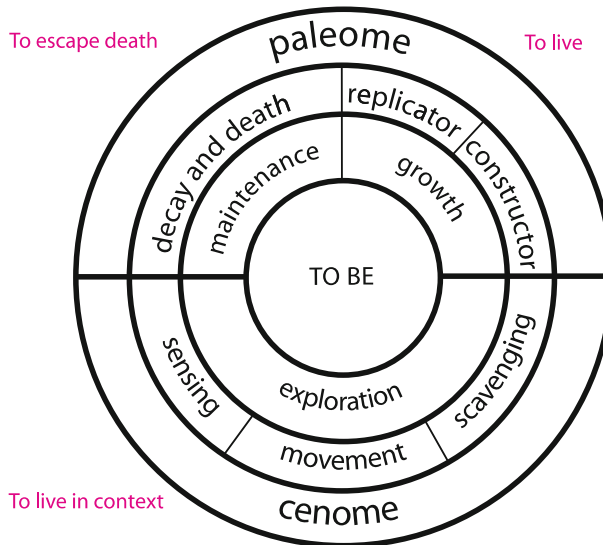


Fig. 2 Organization of bacterial genomes. The genome is split into two major components. The paleome drives all functions required to sustain life, while the cenome permits life in context. The paleome is split into two parts. The first part, comprising a constructor and a replicator, is made of genes that are essential even under laboratory growth conditions. The second part permits aged cells to synthesize young cells, therefore perpetuating life

in context, form the cenome (coenos, common, in Greek, used by ecologists in “biocenose”) (Danchin 2007) (Fig. 2). It is convenient to have an idea of these categories at the onset of the annotation procedure in order to proceed with an efficient functional annotation of the genes. The BioSapiens effort produced lists of genes in the paleome for at least two large classes of organisms, the Firmicutes and the Gamma-proteobacteria (Danchin et al. 2007).

3 Identifying genomic objects

Annotation, broadly speaking, is the extraction of biological knowledge from raw nucleotide sequences. Sequencing DNA samples produces long stretches of nucleotide sequences where the investigator needs to uncover a variety of genomic objects: regions coding for proteins, RNA genes, promoters, transcriptional terminators and miscellaneous features such as insertion sequences, repeated regions, etc. The process of sequencing and annotating bacterial genomes has become highly automated in the past few years and many genome sequences are deposited in databases associated to purely automatic annotation with no further manual check.

In the initial step of automatic annotation, several *in silico* analytical methods are linked up to predict the location of genes and to describe the cellular function of gene products (Table 1 and Fig. 1). First, gene prediction programs are executed to find regions that are likely to encode proteins or functional RNA products. Contrary to the general opinion, it is still somewhat difficult to properly identify CDSs. This is reflected in the ubiquitous, but unfortunate mixing up between terms describing Open Reading Frames (an ORF is a sequence of multiple of three nucleotides between two translation termination codons) and CDSs. Identification of the translation initiation codon is still difficult. Typically in Bacteria, it is preceded by a particular genomic object, a ribosome binding site, which often comprises a GGAGG sequence (Laursen et al. 2005). Several software tools have been created to help in identifying CDSs (Suzek et al. 2001; Yada et al. 2001; Besemer and Borodovsky 2005; Makita et al. 2007). It must be stressed that gene calling programs are still liable to miss small genes or genes of atypical nucleotide composition. To overcome this limit, statistical analysis of intra-genomic variations (see above), can be used to derive multiple gene models which take into account the compositional diversity of genes within a genome (Bocs et al. 2002; Cruveiller et al. 2005). Finally, an increasing number of genomes are being released in “draft” form (i.e. before the finishing stage of a sequencing project) or using techniques that are error-prone in some regions (e.g. typically the 454 sequencing technique is awkward with “homopolymers”, runs of identical nucleotides) with a high sequencing error rates, thus leading to errors in gene predictions. It is indeed important at this point to remind annotators that even the best sequences may contain errors, as any outcome of experimental set ups: we tend to be “nominalists” and are quick to think that what is named and written is right (Eco 1983). The consequence is that automatic annotation will result in a significant number of spurious genes or gene starts, and miss many small genes. In turn, for example, wrong gene start identification often results in by-passing identification of signal peptide signatures such as those identified in PSort (Nakai and Horton 2007) or SignalP (Emanuelsson et al. 2007), and lead to wrong functional assignment of a gene product. This must be borne in mind when reaching the further step of manual annotation (see below).

Transfer RNA molecules can be identified fairly easily using software such as tRNAScan (Schattner et al. 2005). Many other types of small untranslated RNAs are now uncovered as highly relevant to regulatory properties of metabolic networks in bacteria. Furthermore some small RNAs have a catalytic function (ribozymes, such as the RNA moiety of RNase P, which matures tRNAs). The nomenclature describing these molecules is still quite variable, as one finds current acronyms such as sRNA, ncRNA, sncRNA, “Non-coding” is ambiguous, and it would probably be better to stress simply the fact that they are real gene products (hence, coding) that are not translated (small untranslated RNAs, suRNAs or small regulatory RNAs, srRNAs). Because the rules permitting RNA folding are not well understood, these suRNA genes are difficult to predict. Several programs can be run to identify the corresponding objects (Vogel and

Sharma 2005). Their targets can also be identified using software such as TargetRNA (Tjaden et al. 2006) or more evolved approaches such as the one described by Vergassola and co-workers (Mandin et al. 2007) (see <http://www.ncRNA.org>).

Miscellaneous features, such as transcription promoters and terminators (d'Aubenton Carafa et al. 1990; de Hoon et al. 2005), insertion sequences (Siguier et al. 2006), repeats (Achaz et al. 2007) etc., can be annotated using a variety of software. There is no general procedure of annotation for those features, and the annotator has to write scripts to chain relevant procedures when needed. The most important feature in this respect would be identification of promoters, as this would permit better identification and annotation of genes (transcription is needed to express RNAs and precedes translation). Unfortunately however, this remains, till the present time, an impossible task. Prediction of operons can be performed using a combination of criteria (Yan and Moulton 2006) and a database such as RegulonDB may help in identifying some promoters, via connection to experimentally validated examples (Salgado et al. 2006).

4 Functional annotation

Once the global features of the genome have been analyzed, it becomes important to proceed to annotation of CDSs, associating functional properties to the corresponding sequence (Table 1 and Fig. 1). In many cases, as just stated, investigators rely on automatic annotation platforms and combine scripts with relevant data input and data consistency analyses to release an “annotated” sequence to one of the entry points of the INSDC, with the corresponding format (which differs between DDBJ, EMBL-EBI and GenBank). The accuracy of this step depends not only on the software used for the automatic annotation (i.e. on inferences produced by running comparison algorithms (usually BLAST, sometimes FASTA)), on the quality of the sequence itself but above all, on the quality of the primary resources i.e. the annotation already stored in the INSDC databases. This standard approach is therefore a fairly inaccurate way of annotating genomes as the simple fact that a common annotation is associated to a gene will make it seem relevant, resulting in the percolation of annotation errors (Gilks et al. 2002, 2005). Writing parsers implementing simple biological consistency checks driven by knowledge of constraints associated to the ecological niche of the organism may improve annotation by flagging obvious inconsistencies. For example prediction of a lactose permease in a member of the Cyanobacteria should raise questions, or prediction of an outer membrane protein in a member of the Firmicutes should generate an error message.

To increase the value of functional annotation, some automatic procedures give a priority to the similarity results obtained with reference annotations of model organisms. This is important to remember as this places model organisms at the root of propagation of inferences. It is therefore of the utmost importance to verify that the

annotation pipeline is geared to extract the most recent annotations coming from the models. As a good practice the annotation pipelines should incorporate some kind of quality factor to annotations used for inferences, with the highest values corresponding to experimentally validated data, often associated to model genomes.

Whenever possible the annotator should use comparison with data that have been validated either via in-depth *in silico* analysis or experimentally. The SwissProt protein database, now extended as the UniProt knowledge base is the best reference that should be used for general annotation (The-Uniprot-consortium 2008). This database is constantly improved but it cannot match the extraordinary speed of data production, so that it has extended to a parallel database TrEMBL, that extends annotation automatically to all regions annotated as CDSs (often simple ORFs, unfortunately) in published sequences. The accuracy and depth of annotation in UniProt depends on the protein, on the organism and on its domain of action. The detail of annotation and degree of validation is highly variable. The HAMAP project, or ‘High-quality Automated and Manual Annotation of microbial Proteomes’, aims to integrate manual and automatic annotation methods in order to enhance the speed of the curation process while preserving the quality of the database annotation (Gattiker et al. 2003) and it is an indispensable reference to annotate bacterial protein genes.

The core of functional annotation deals with protein coding genes. It is essential, at this step, to have some consistent view of what a “function” is. Unfortunately, this concept is very fuzzy, and not consistently used by biologists (Allen et al. 1998). This led groups of investigators to create dictionaries of terms and so-called “ontologies”. The word “ontology”, which has a very specific meaning in philosophy, has curiously been diverted in health care sciences from its original meaning (Herbert 1995) to refer to a particular structured vocabulary describing knowledge associated to a patchwork of biological data, objects, sequences, biological functions and functionalities and other general features of biological processes. It refers to different ways of considering living organisms and several ontologies are used to characterize the life of a particular set of organisms. In the context of genome annotation, an ontology will rest on a particular data model that can be used to organize specialized databases. The step of definition of the exact meaning of a particular vocabulary to describe features of genomic objects is an essential prerequisite for annotation. Several ontologies are used in this respect, in particular the GO ontology (Gene-ontology-consortium 2001, 2008). This classification, although not originally defined for bacterial genome annotation, is useful for the consideration of individual proteins in the context of the cell: what they do, i.e. the molecular function that describes the biochemical role of the protein (transporter, regulator, enzyme, structural protein, etc.); where they are found in the cell, i.e. their subcellular localisation (cytoplasm, periplasm, cytoplasmic membrane, etc.); and what larger processes they participate in, i.e. the biological function that describes the role of the protein in the cell (metabolic pathway, signalling cascade, etc.). Even when one creates her or his own data schema, it is useful to have a correspondence, whenever

possible, with that particular ontology. In practice the features of genomic objects considered in Gene Ontology should be systematically annotated (when possible, if not, a qualifier “unknown” should be used).

Proteins are often enzymes. An important feature, which will permit metabolic reconstruction, is the assignment of an EC number to enzymes (or predicted enzymes). The Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes by the reactions they catalyse ascribes a reference number (EC number) made of four digital numbers separated by a dot, corresponding to a hierarchical classification of enzyme activities (e.g. EC 1.1.1.1 refers to alcohol dehydrogenase). This nomenclature was initiated in 1956 to normalize naming of enzymes as the uncontrolled naming by biochemists resulted in a jungle of names almost impossible to explore, with the consequence that it was not possible to know whether a given activity had already been identified. Unfortunately, while this endeavour is extremely important for the future of annotation, it stumbled on many difficulties as new structures of enzymes and intricacies in catalytic reactions were uncovered (typically, a given protein can catalyze quite different functions, a given reaction can require several different proteins, some shared with different enzymes, for example). Today, the largest publicly available enzyme information system is certainly the Brenda database (Barthelmes et al. 2007). Among the tools that are useful for annotation there exists prediction systems that use the sequence of a protein and compares it to a database profile of identified proteins of known activity and predicts its EC number (PRIAM, Claudel-Renard et al. 2003).

Finally, in addition to the general prediction of gene functions, annotation pipelines can provide other types of information about the encoded products, in general proteins: chemical and structural properties (e.g. isoelectric point and molecular mass are important information for proteomic studies), subcellular localisation (which has implications for both the function of the protein and its interactions with other proteins), and modular organisation (formation of complexes). It is indeed important for the different domains of a protein to be characterised so as to avoid a well-known annotation error: a function is transferred to another protein that only shares one single module. Consequently, sequence similarity search tools for thematic databases, such as motif/pattern or enzyme family databases, are also used.

5 A recursive view of genome annotation

Once a first round of annotation has been produced, the annotator has a much better view of the global knowledge they are associating to the sequence. Inferences can be made using the existing round of annotation to create new knowledge (Fig. 1 and Table 1).

At this stage, annotation will therefore be modified recursively. Analysis of neighbourhoods is being employed to improve the accuracy of genome annotation

(co-occurrence, gene order, gene fusion, but also results of multivariate analyses, such as those found in the “R” package (<http://www.r-project.org/>). They provide information for functional characterisation of genes from their genome environment, e.g. by allowing a more confident identification of orthologues when sequence similarities are low. In addition, the analysis of co-localised genes provides clues about the functional interactions between the corresponding proteins, which is a first step towards the description of protein interaction networks. This dynamic view of genome annotation is useful to correct or to increase the specificity of the assigned functional annotations (Fig. 1 and Table 1).

As a case in point, metabolic-network reconstruction is an active field of research, and some annotation platforms include automatic prediction of this kind of network (see Annex). The annotated proteins that are characterised by an EC number and/or an enzyme name, are used to identify steps that ought to be present (or should be missing, knowing the biological properties of the organism), permitting investigators to explore functions of tentatively annotated genes as fulfilling the missing reactions. These predictions are performed by comparing the enzymes within a given genome against the known set of reference pathways stored in metabolic database such as KEGG annotation, the (Aoki-Kinoshita and Kanehisa 2007) and BioCyc (Caspi et al. 2008). The value of this similarity-based reconstruction is highly dependent on the quality of completeness of the metabolic database, and the criteria used for assessing the presence of a pathway. While the very large KEGG metabolic maps are mosaics that combine pathways and reactions from many organisms, MetaCyc pathways describe single metabolic routes that have been experimentally elucidated in specific organisms. This latter metabolic resource is obviously more accurate than the KEGG database. However, in terms of completeness, the KEGG maps are sometimes useful to propose hypotheses on existing alternative metabolic pathways. Although the automated reconstructions provide an overview of the metabolic capabilities of the studied microorganisms, detailed assessment of the validity of these networks remains essential. Indeed, issues potentially leading to error include incorrect substrate specificity, multifunctional enzymes, reaction reversibility, association to coenzymes or prosthetic groups, and missing known reactions that have no assigned gene.

Proceeding to the reconstruction of as much as possible of the metabolic capacities of the organism will monitor the consistency of the annotation. Proximity of genes with functions belonging to the same pathway will help in making relevant inferences (Table 1). While this can be partially automated, it is clear that the annotator expertise is required at this step. Furthermore, the risk of annotation errors percolation advocates for an annotation process that combines automatic with manual annotation. Once a first round of annotation has been performed the team of investigators involved in curating the process will look for global properties of the genes, then go down to more and more specific features.

A first analysis will try to find inconsistencies in the CDS start point, by comparing predictions with what is known of cognate proteins from related organisms. After this step, analysis of the addressing of proteins will substantiate a significant fraction of the predictions by predicting signal peptides, which should display fairly consistent features and split into at least two classes, cleaved signal peptides, and lipoprotein signal peptides. A further analysis of gene product compartmentalization will use prediction of integral inner membrane proteins, based on the discrepancy in the way amino acid residues are distributed in this class as compared to the bulk of proteins (Pascal et al. 2006).

All these steps need to be integrated in a common platform to permit efficient curation/correction of annotations. This gives an important role to the user-friendliness of the platform and requires deep reflection on the human/machine interaction, where the end user is perceived as a biologist, often an experimental biologist. Much work needs to be devoted to improve the existing platforms. One way forward might be to go for the use of Widgets, as advocated by Valencia and colleagues in this book.

Contributing to the dispute about automatic vs. expert annotation, Raes and collaborators recently published a surprising observation: they estimated that, in completely sequenced genomes, the fraction of proteins to which at least some functional features can be automatically assigned is close to 75% using similarity searches alone, and 85% if genomic context methods are also used (Raes et al. 2007b). However, for genes which do not belong to the paleome (often referred to as “house-keeping” genes), these automatic annotations might often be erroneous and of limited use for biologists (e.g. unknown enzyme/transporter substrate, very short domain, etc.): manual curation undoubtedly remains necessary.

6 Improving annotation: parallel analysis and comparison of multiple bacterial genomes

In parallel with improvement of annotation procedures, annotations of specific genomes need to be reprocessed on a regular basis to take into account the identification of newly characterized functions. Furthermore, large-scale functional analyses produce additional data that contribute to the interpretation of genomic data. Every annotation project should monitor re-annotation of related projects: although a number of annotation sources are very accurate, continual updating of genome annotations for a large number of species is not straightforward. Indeed, databases and computational methods are constantly evolving and processing again automatic functional annotations should be performed on a regular basis. In addition, new experimentally-derived functional information is being continually generated, and can prove useful, for example, for modifying the annotation of genes of “putative” or unknown function. This requires systematic exploration of bibliographic references (<http://www.pubmed.gov/>), an element of paramount importance for collecting sound fundamental knowl-

edge about model organisms. While we had mostly convenient access to the content of the title an abstract of articles in Medline at PubMed, the new PubMed Central database allows the annotator to explore using a combination of keywords the whole content of articles in Open Access (Walport and Kiley 2006). This needs to be used as much as possible in manual annotation (Table 1).

The transfer of the reliable up-to-date reference annotation to validated orthologues in the newly sequenced genomes is essential. Annotation platforms already exist that include procedures which rely on the ability to cluster proteins from related genomes into orthologous groups (a first attempt was provided by the Cluster of Orthologs, COGs (Tatusov et al. 2003)), and new interfaces allowing annotators to view evidence associated with each protein in the group and make annotation decisions about the group as a whole. Care must be taken to investigate the fine details of the clusters of orthologs, as structure and function are often not related in the expected way.

Comparative genomics is based on the development of novel methods, databases and graphical interfaces for organizing and extracting biological information from the comparison of a large collection of complete and unfinished genome sequences.

Techniques based on the knowledge of the genomic context use the co-localisation of genes in several genomes at various levels of proximity (chromosomal, metabolic, co-citation, etc.) and do not require sequence similarity for the genes to be annotated. Especially, genes co-localized in the chromosome tend to be functional neighbours, either in terms of expression patterns or network neighbourhood. Combined with similarity-based predictions, such information can be used to elucidate protein function. This technique has been exploited recently and proved valuable for the accurate identification of candidates for the missing genes of the lysine fermentation pathway in anaerobic organisms (Kreimeyer et al. 2007). Combined with experimental validation, as in the case of lysine fermentation, it has also been used in the complete identification of the methionine salvage pathway in a variety of organisms, where misannotation of several proteins precluded construction of a consistent chemical pathway for the recycling of the methylthioadenosine produced in the course of polyamine biosynthesis (Sekowska et al. 2004). Furthermore, inference from annotation can help identification of unexpected objects where it can be surmised that a function must exist, while no object that would permit it has been yet found (Danchin 1999). This was illustrated with the identification of a new pretein for oligoribonuclease, which had no common structural counterpart in Firmicutes (Mechold et al. 2007).

Analysis of genomes from closely related species can help in the identification of novel genes and other features such as gene fusion/fission and pseudogenes that only become apparent in a comparative genomics context. These phenomena also include lineage-specific genes which can be characterised efficiently if many related sequences are exploited. Identification of genetic differences between entire genomes allows correlation of the differences with biological function, providing insight into selective evolutionary pressures and patterns of gene transfer or loss. This has proven to be

particularly relevant for virulence analysis of pathogenic strains. For example, a comparative analysis of several extraintestinal pathogenic *Escherichia coli* (ExPEC) strains has shown that the ability to accumulate and express a variety of virulence-associated genes distinguishes ExPEC from many commensals and that different way exist among ExPEC to cause disease (Brzuszkiewicz et al. 2006).

A multi-strain annotation project might also involve Single Nucleotide Polymorphism (SNPs) analyses to address evolutionary issues. SNPs are very short sequence differences between closely homologous sequences; they may affect either coding or non-coding sequences. SNPs are important features of a genomic sequence: for example, they may reveal genes that contribute to the adaptation of the bacteria to different environmental stimuli, allowing them to shift from commensalism to pathogenicity (Wei et al. 2006). SNP analysis, however, needs to be aware of the possibility of sequencing errors in the data of interest.

There is finally the interesting issue of handling projects related to the annotation of bacterial genomes that are evolutionarily distant, and very different, from the minuscule fraction of microbial species we know today. Examples are studies of prokaryotic species belonging to a novel bacterial genus (e.g. *Herminiimonas arsenicoxydans*, which is able to metabolize arsenic and to efficiently colonise toxic environments (Muller et al. 2007)), and some metagenomics projects enabling the reconstruction of complete genomes. The case of an anaerobic ammonium oxidation (anammox) community dominated by *Kuenenia stuttgartiensis* is interesting and illustrative: genome annotation has led to novel candidate genes for hydrazine and ladderane metabolism (Strous et al. 2006). Obviously, it is impossible to implement in a computer the rules that a manual annotator would follow because the biology of such organisms presents numerous exceptions or novel features. The meticulous work of expert annotation is thus very often the only way to discover novelties.

7 Perspectives: new developments for the construction of genome databases, metagenome analyses and user-friendly platforms

Reference databases, computational methods and knowledge that form the basis of annotation pipelines keep being developed at many places in the world, making the process difficult to update. In addition, the rapid increase of new sequence data has necessitated the evolution of software resources from functional annotation of a single genome towards simultaneous analysis of information from multiple genomes. There is now a natural shift towards the creation of tools for viewing and manipulating data in a comparative genomics context.

Novel methods are emerging for the annotation of sets of functionally-related genes across a set of genomes, rather than the usual “gene-by-gene” annotation of one genome

at a time. Indeed, exploration of biological processes is more effective when developed at a global scale, and gathering biological knowledge about several organisms simultaneously allows biologists to detect discrepancies and identify exceptions (e.g., the lack of a key enzymatic reaction in a pathway in several organisms may suggest the existence of an alternative route). Tools such as SEED (Overbeek et al. 2005), and “Genome Properties” (Haft et al. 2005) define a set of biological processes (e.g. metabolic pathways, secretion systems) and a set of functional roles that are essential in the completion of a biological process. For each process, a two-dimensional matrix is obtained in which columns describe roles and rows describe organisms. The status of a cell in the table defines which gene(s) encode(s) a particular step in the process in a given organism. This matrix can be used as a starting point to identify variants concerning a process by gathering organisms sharing the same profile (i.e. the same functional roles). In addition, Genome Properties proposes rules, mainly based on role essentiality, to determine automatically whether or not a process exists in a given organism. The integration of these approaches into annotation platforms is not yet implemented. This should improve annotations such that automatic analysis of functional variants is possible, including mapping missing genes and locating gene candidates for experimental validation.

New advances in sequencing technologies have given rise to the field of metagenomics (Riesenfeld et al. 2004). A metagenome (environmental genome, community genome) is a sample of an aggregate genome collection of all community members obtained directly from the natural environment, without a preliminary cultivation step. Since most (>98%) naturally occurring bacteria cannot be cultured, metagenomic studies provide us with a mechanism for analyzing previously unknown organisms. Various sampling and sequencing strategies can help in answering different questions about the diversity and abundance of community members, their metabolic potential, and the complex interactions between members of a specific environment.

Many novel computation tools have recently been developed to analyze this metagenomics data, starting with new algorithms for sequence assembly to increase fragment contig length and assembly quality. The development of metagenome gene prediction software is still an ongoing endeavour: new tools are necessary to deal with fragmented genes, phylogenetic diversity and lower end-quality of sequences, systematically resulting in frameshifts (Raes et al. 2007a). The use of BLASTX-derived annotation of reads avoids gene prediction problems, but limits the analysis to what is known in the metagenome. Then, starting from this large amount set of genes and gene fragments, other methods and tools are used to establish the taxonomic identity of community members (i.e, diversity and abundance of microbial communities).

In parallel with the unfolding of this new era of genomics, software tools are continuously being developed to analyze the functional and metabolic potential of microbial communities to examine differences between their collective genome sequence datasets in terms of functional categories and/or metabolic pathways. Such differential

functional analysis is based on comparing the frequency of genes that may code for specific functions across metagenome or isolated genome datasets (Raes et al. 2007a). The future of genomics will considerably owe to novel developments of automatic annotation platforms allowing investigators to cope with the flood of sequence data which is now following the considerable decrease in DNA sequencing costs.

8 Annex: databases and platforms for annotating bacterial genomes

Genome annotation relies on inferences stemming from a wealth of data collections that gather knowledge on nucleotide and protein sequences, metabolites and pathways, interactions etc. The International Nucleotide Sequence Databases Collaboration (INSDC: DDBJ/EMBL-EBI/GenBank <http://www.insdc.org/>) is the reference archive for the construction of derived repositories containing a subset of sequences, or for computational analysis based upon DNA information. This reference library, established almost 30 years ago, is now facing new challenges in parallel with the exponential increase of sequences as well as of the associated knowledge. In particular, the huge number of associated features can hardly be efficiently queried, and updating annotation is not straightforward. Parallel resources have therefore been developed with the aim of collecting and retrieving information from a subset of the global database. Genome Reviews at the EBI presents an up-to-date, standardized and comprehensively annotated version of complete genomes (Kersey et al. 2005; Sterk et al. 2006). RefSeq at the NCBI provides an integrated and non-redundant set of nucleotide and protein sequences for organisms widely used in research (Pruitt et al. 2007). These major databases are mostly driven by automatic procedures and do not tackle specialized requests. Hence, organism-specific databases have been constructed as an important data resource for the annotation of new bacterial genomes and the re-annotation of “old” genomes. These specialized databases are sometimes carefully curated, with an emphasis on the most relevant features of the organism of interest (the ecological niche, with its corresponding genome, in particular). They include support from manual investigation of the relevant literature. In this context, model organisms are prominent references for investigation of related species: this is especially true for *Escherichia coli* and *Bacillus subtilis*, the most extensively studied Gram-negative and Gram-positive bacteria, respectively. Several microbial genome web resources provide consistent and comprehensive sets of annotations, together with a series of tools for genome querying, analysis, and comparative studies. Features range from simple selection of organisms and genes to complex multi-genome queries, e.g. queries allowing the user to identify specific or shared proteins among a set of selected genomes. Examples of such specialised software environments include the Integrated Microbial Genomes (IMG) (Markowitz et al. 2006) and the Comprehensive Microbial Resource (CMR) databases

(Peterson et al. 2001). Recently, the Colibri and SubtiList specialized databases were extended to a collection for model organisms, GenoChore, using a novel data model and MySQL as the database management system (<http://bioinfo.hku.hk/genochore.html>) (Fang et al. 2005a) and the GenoList genome browser (<http://genolist.pasteur.fr/GenoList>) was upgraded to provide an intuitive yet powerful multi-genome user interface, primarily designed to address biologists' requirements, and including original functionalities such as subtractive proteome analysis (Lechat et al. 2008).

A large collection of databases is available for proteins. The most widely accepted reference is UniProtKB, which originated from a merger between the Swiss-Prot and PIR protein databases (The-Uniprot-consortium 2008). The quality targets of this knowledgebase are numerous: expertly curated annotations, access to the relevant literature, numerous cross-references, etc. However, the exponential increase of sequence data has made the curation of all that information an impossible task. Two separate sections have therefore been defined in the knowledgebase: UniProtKB/Swiss-Prot contains protein entries that are still manually validated, and UniProtKB/TrEMBL provides access to computationally annotated records. In parallel, the Swiss Institute of Bioinformatics has initiated the High-quality Automated and Manual Annotation of Microbial Proteomes (HAMAP) project meant to create rules allowing semi-automatic annotation of orthologous proteins in the Swiss-Prot database that are part of well-conserved families in prokaryotes (Gattiker et al. 2003).

Protein are further annotated *via* identification of motifs and domains using a variety of approaches, creating specialized databases. Most of these organize proteins into families sharing motifs or domains. The InterPro resource brings together many of these databases, providing a unified resource to search for protein signatures (Mulder et al. 2007). Other levels of classification are defined using clustering procedures that lead to groups of proteins: for instance, the COG (Clusters of Orthologous Groups of proteins) resource was built upon systematic BLASTP comparison of all proteins against all (from selected genomes), and subsequent construction of clusters containing at least three proteins from distant species (Tatusov et al. 2003).

Many proteins are enzymes, hence the annotation of enzymatic information provided by databases such as BRENDA is of particular interest (Barthelmes et al. 2007). In the same way, the explicit abstract description of metabolic pathways – either computationally predicted or experimentally described – is also of considerable importance. It is best maintained in the KEGG (Kanehisa et al. 2008) and BioCyc/MetaCyc databases (Caspi et al. 2008). Interaction data are also of major importance, and despite the high level of noise in the techniques permitting their identification, they can be useful in annotation when used with care (e.g., the STRING online resource gives access to protein interaction data, by integrating known and predicted interactions from a large variety of organisms (Huynen et al. 2000)). Finally, recent large-scale endeavours in structural genomics are enhancing data collections such as the Protein Data Bank (PDB), used to store protein three-dimensional structures of individual proteins or

complexes (Berman et al. 2007). PDB is now automatically referred to in UniProt (Martin 2005).

As described, the genome annotation process requires complex bioinformatics support. This includes at least three main elements: a pipeline for fully automated sequence annotation, using a large spectrum of bioinformatics tools, a consistent data management system (i.e. advanced biological data models and integrated databases), and several interactive graphical interfaces to organize and present the results in a user-friendly manner. These features are rarely present together in the most common annotation platforms.

The first automated pipelines, MAGPIE (Gaasterland and Sensen 1996) and GeneQuiz (Scharf et al. 1994) were developed more than ten years ago. They provided entirely automatic annotation, based on artificial intelligence approaches that provided “reasoning” capabilities to the software, thereby permitting the combination of analyzes produced by the different tools, and the assignment of a specific biological function. They also provided a quality factor for assessment of the annotation accuracy (Andrade et al. 1999). However, resulting from the chaining of extraction of automatic analyzes, percolation of annotation errors was pervasive (Gilks et al. 2002, 2005).

Since this early time, automatic software with new functionalities continued to be developed world-wide. As examples, the AutoFACT pipeline takes a single FASTA formatted genomic sequence file as input and proceeds to assign each individual sequence to one among six annotation functional protein classes, combining multiple BLAST reports from several user-selected databases (Koski et al. 2005). BASys, a web server performs completely automated annotation of bacterial chromosomes, combining more than 30 programs to determine approximately 60 annotation subfields for each gene, including gene/protein name, GO function, COG function, possible paralogues and orthologues, molecular mass, isoelectric point, operon structure, subcellular localization, signal peptides, transmembrane regions, secondary structure, 3D structure, reactions and pathways (Van Domselaar et al. 2005).

Web services clearly provide the most convenient tools for research groups who lack the computing resources and expertise required to install or implement the software necessary for bacterial genome annotation. However, in most of these systems, user-friendly interfaces, which are essential for allowing manual input of expertise superimposed on automatic predictions to ensure high-quality annotation, are not available. This observation led to the development of annotation browsers and editors such as Artemis (Rutherford et al. 2000). This system is a useful tool for reviewing and editing annotation, based on annotations collected using other programs.

Most of the existing systems offer two complementary functionalities: they generate automatic annotations, and provide graphical facilities for subsequent manual review of the predictions (see the general features described in Table 1). Examples of comprehensive annotation platforms include commercial or private systems, such as ERGO (Overbeek et al. 2003), Pedant-Pro (<http://www.biomax.com/products/pedantpro.php>,

successor of PEDANT (Frishman et al. 2001; Riley et al. 2007)), GNARE (Sulakhe et al. 2005), SMIGA (Lu et al. 2006), and open-source systems, such as GenDB (Meyer et al. 2003), Manatee (TIGR, unpublished), SABIA (Almeida et al. 2004), and AGMIAL (Bryson et al. 2006). However, installing these systems on a local computer is not straightforward because they make use of many bioinformatics tools that must be installed independently. Finally, the availability of a large collection of genomes led more recently to the development of comparative annotation and analysis environments, such as MaGe (Vallenet et al. 2006), which enables the annotation of microbial genomes using genomic context and synteny results obtained using known bacterial genome sequences. Indeed, the predictive power of chromosomal clustering has proven to be very effective for assigning putative functions to genomic objects.

References

- Achaz G, Boyer F, Rocha EP, Viari A, Coissac E (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics* 23: 119–121
- Allen C, Bekoff M, Lauder G (eds.) (1998) *Nature's purposes: analyses of function and design in biology*. MIT Press, Cambridge, MA
- Almeida LG, Paixao R, Souza RC, Costa GC, Barrientos FJ, Santos MT, Almeida DF, Vasconcelos AT (2004) A system for automated bacterial (genome) integrated annotation – SABIA. *Bioinformatics* 20: 2832–2833
- Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C (1999) Automated genome sequence analysis and annotation. *Bioinformatics* 15: 391–412
- Aoki-Kinoshita KF, Kanehisa M (2007) Gene annotation and pathway mapping in KEGG. *Methods Mol Biol* 396: 71–92
- Bailly-Bechet M, Danchin A, Iqbal M, Marsili M, Vergassola M (2006) Codon usage domains over bacterial chromosomes. *PLoS Comput Biol* 2: e37
- Barthelme J, Ebeling C, Chang A, Schomburg I, Schomburg D (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res* 35: D511–D514
- Benson DA, Boguski M, Lipman DJ, Ostell J (1994) GenBank. *Nucleic Acids Res* 22: 3441–3444
- Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 35: D301–D303
- Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 33: W451–W454
- Bocs S, Danchin A, Médigue C (2002) Re-annotation of genome microbial coding-sequences: finding new genes and inaccurately annotated genes. *BMC Bioinformatics* 3: 5
- Bryson K, Loux V, Bossy R, Nicolas P, Chaillou S, van de Guchte M, Penaud S, Maguin E, Hoebeke M, Bessieres P, Gibrat JF (2006) AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. *Nucleic Acids Res* 34: 3533–3545
- Bruszkiewicz E, Bruggemann H, Liesegang H, Emmerth M, Olschlager T, Nagy G, Albermann K, Wagner C, Buchrieser C, Emody L, Gottschalk G, Hacker J, Dobrindt U (2006) How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. *Proc Natl Acad Sci USA* 103: 12879–12884
- Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD (2008) The MetaCyc database of metabolic

- pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 36: D623–D631
- Claudel-Renard C, Chevalet C, Faraut T, Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* 31: 6633–6639
- Cruveiller S, Le Saux J, Vallenet D, Lajus A, Bocs S, Médigue C (2005) MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. *Nucleic Acids Res* 33: W471–W479
- d'Aubenton Carafa Y, Brody E, Thermes C (1990) Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J Mol Biol* 216: 835–858
- Danchin A (1999) From protein sequence to function. *Curr Opin Struct Biol* 9: 363–367
- Danchin A (2003) *The Delphic boat. What genomes tell us.* Harvard University Press, Cambridge, Mass, USA
- Danchin A (2007) Archives or palimpsests? Bacterial genomes unveil a scenario for the origin of life. *Biol Theor* 2: 52–61
- Danchin A, Fang G, Noria S (2007) The extant core bacterial proteome is an archive of the origin of life. *Proteomics* 7: 875–889
- Danchin A, Médigue C, Gascuel O, Soldano H, Hénaut A (1991) From data banks to data bases. *Res Microbiol* 142: 913–916
- de Hoon MJ, Makita Y, Nakai K, Miyano S (2005) Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comput Biol* 1: e25
- Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res* 33: e6
- Eco U (1983) *The name of the rose.* Harcourt Brace Jovanovich, Orlando, FL, USA
- Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using Target P, Signal P and related tools. *Nat Protoc* 2: 953–971
- Fang G, Ho C, Qiu Y, Cubas V, Yu Z, Cabau C, Cheung F, Moszer I, Danchin A (2005a) Specialized microbial databases for inductive exploration of microbial genome sequences. *BMC Genomics* 6: 14
- Fang G, Rocha E, Danchin A (2005b) How essential are nonessential genes? *Mol Biol Evol* 22: 2147–2156
- Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW (2001) Functional and structural genomics using PEDANT. *Bioinformatics* 17: 44–57
- Gaasterland T, Sensen CW (1996) MAGPIE: automated genome interpretation. *Trends Genet* 12: 76–78
- 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
- Gene-ontology-consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res* 11: 1425–1433
- Gene-ontology-consortium (2008) The gene ontology project in 2008. *Nucleic Acids Res* 36: D440–D444
- Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* 18: 1641–1649
- Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA (2005) Percolation of annotation errors through hierarchically structured protein sequence databases. *Math Biosci* 193: 223–234
- Haft DH, Selengut JD, Brinkac LM, Zafar N, White O (2005) Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics* 21: 293–306
- Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 5: R245–R249

- Hénaut A, Lisacek F, Nitschké P, Moszer I, Danchin A (1998) Global analysis of genomic texts: the distribution of AGCT tetranucleotides in the *Escherichia coli* and *Bacillus subtilis* genomes predicts translational frameshifting and ribosomal hopping in several genes. *Electrophoresis* 19: 515–527
- Hénaut A, Rouxel T, Gleizes A, Moszer I, Danchin A (1996) Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. *J Mol Biol* 257: 574–585
- Herbert SI (1995) Informatics for care protocols and guidelines: towards a European knowledge model. *Stud Health Technol Inform* 16: 27–42
- Hsiao WW, Ung K, Aeschliman D, Bryan J, Finlay BB, Brinkman FS (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet* 1: e62
- Huynen M, Snel B, Lathe W 3rd, and Bork P (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 10: 1204–1210
- Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36: D480–D484
- 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) Integrate and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res* 33: D297–D302
- Koski LB, Gray MW, Lang BF, Burger G (2005) AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinformatics* 6: 151
- Kreimeyer A, Perret A, Lechaplais C, Vallenet D, Médigue C, Salanoubat M, Weissenbach J (2007) Identification of the last unknown genes in the fermentation pathway of lysine. *J Biol Chem* 282: 7191–7197
- Laursen BS, Sorensen HP, Mortensen KK, Sperling-Petersen HU (2005) Initiation of protein synthesis in bacteria. *Microbiol Mol Biol Rev* 69: 101–123
- Lechat P, Hummel L, Rousseau S, Moszer I (2008) GenoList: an integrated environment for comparative analysis of microbial genomes. *Nucleic Acids Res* 36: D469–D474
- Lu Q, Hao P, Curcin V, He W, Li YY, Luo QM, Guo YK, Li YX (2006) KDE Bioscience: platform for bioinformatics analysis workflows. *J Biomed Inform* 39: 440–450
- Makita Y, de Hoon MJ, Danchin A (2007) Hon-yaku: a biology-driven Bayesian methodology for identifying translation initiation sites in prokaryotes. *BMC Bioinformatics* 8: 47
- Mandin P, Repoila F, Vergassola M, Geissmann T, Cossart P (2007) Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets. *Nucleic Acids Res* 35: 962–974
- Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285: 751–753
- Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavromatis K, Kunin V, Garcia Martin H, Dubchak I, Hugenholtz P, Kyrpides NC (2006) An experimental metagenome data management and analysis system. *Bioinformatics* 22: e359–e367
- Martin AC (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics* 21: 4297–4301
- Mechold U, Fang G, Ngo S, Ogryzko V, Danchin A (2007) YtqI from *Bacillus subtilis* has both oligoribonuclease and pAp-phosphatase activity. *Nucleic Acids Res* 35: 4552–4561
- Médigue C, Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol* 158: 724–736
- Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* 222: 851–856
- Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Puhler A (2003) GenDB – an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* 31: 2187–2195
- Moszer I, Glaser P, Danchin A (1995) SubtiList: a relational database for the *Bacillus subtilis* genome. *Microbiology* 141(Pt 2): 261–268

- Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database. *Nucleic Acids Res* 35: D224–D228
- Muller D, Médigue C, Koechler S, Barbe V, Barakat M, Talla E, Bonnefoy V, Krin E, Arsene-Ploetze F, Carapito C, Chandler M, Cournoyer B, Cruveiller S, Dossat C, Duval S, Heymann M, Leize E, Lieutaud A, Lievreumont D, Makita Y, Mangenot S, Nitschké W, Ortet P, Perdrial N, Schoepp B, Siguier P, Simeonova DD, Rouy Z, Segurens B, Turlin E, Vallenet D, Van Dorsseleer A, Weiss S, Weissenbach J, Lett MC, Danchin A, Bertin PN (2007) A tale of two oxidation states: bacterial colonization of arsenic-rich environments. *PLoS Genet* 3: e53
- Nakai K, Horton P (2007) Computational prediction of subcellular localization. *Methods Mol Biol* 390: 429–466
- Necsulea A, Lobry JR (2007) A new method for assessing the effect of replication on DNA base composition asymmetry. *Mol Biol Evol* 24: 2169–2179
- Nitschké P, Guerdoux-Jamet P, Chiappello H, Faroux G, Hénaut C, Hénaut A, Danchin A (1998) Indigo: a World-Wide-Web review of genomes and gene functions. *FEMS Microbiol Rev* 22: 207–227
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33: 5691–5702
- Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol* 1: 93–108
- Overbeek R, Larsen N, Walunas T, D'Souza M, Pusch G, Selkov E Jr, Liolios K, Joukov V, Kaznadzey D, Anderson I, Bhattacharyya A, Burd H, Gardner W, Hanke P, Kapatral V, Mikhailova N, Vasieva O, Osterman A, Vonstein V, Fonstein M, Ivanova N, Kyrpides N (2003) The ERGO genome analysis and discovery system. *Nucleic Acids Res* 31: 164–171
- Pascal G, Médigue C, Danchin A (2006) Persistent biases in the amino acid composition of prokaryotic proteins. *Bioessays* 28: 726–738
- Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O (2001) The comprehensive microbial resource. *Nucleic Acids Res* 29: 123–125
- Poptsova MS, Gogarten JP (2007) The power of phylogenetic approaches to detect horizontally transferred genes. *BMC Evol Biol* 7: 45
- Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (Ref-Seq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35: D61–D65
- Raes J, Foerster KU, Bork P (2007a) Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* 10: 490–498
- Raes J, Harrington ED, Singh AH, Bork P (2007b) Protein function space: viewing the limits or limited by our view? *Curr Opin Struct Biol* 17: 362–369
- Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 38: 525–552
- Riley ML, Schmidt T, Artamonova II, Wagner C, Volz A, Heumann K, Mewes HW, Frishman D (2007) PEDANT genome database: 10 years online. *Nucleic Acids Res* 35: D354–D357

- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 16: 944–945
- Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 34: D394–D397
- Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchison CA 3rd, Slocombe PM, Smith M (1978) The nucleotide sequence of bacteriophage phiX174. *J Mol Biol* 125: 225–246
- Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB (1982) Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* 162: 729–773
- Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C, Sander C (1994) GeneQuiz: a workbench for sequence analysis. *Proc Int Conf Intell Syst Mol Biol* 2: 348–353
- Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 33: W686–W689
- Sekowska A, Denervaud V, Ashida H, Michoud K, Haas D, Yokota A, Danchin A (2004) Bacterial variations on the methionine salvage pathway. *BMC Microbiol* 4: 9
- Siguier P, Filee J, Chandler M (2006) Insertion sequences in prokaryotic genomes. *Curr Opin Microbiol* 9: 526–531
- Smith HO, Tomb JF, Dougherty BA, Fleischmann RD, Venter JC (1995) Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* 269: 538–540
- Sterk P, Kersey PJ, Apweiler R (2006) Genome reviews: standardizing content and representation of information about complete genomes. *Omics* 10: 114–118
- Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW, Horn M, Daims H, Bartol-Mavel D, Wincker P, Barbe V, Fonknechten N, Vallenet D, Segurens B, Schenowitz-Truong C, Médigue C, Collingro A, Snel B, Dutilh BE, Op den Camp HJ, van der Drift C, Cirpus I, van de Pas-Schoonen KT, Harhangi HR, van Niftrik L, Schmid M, Keltjens J, van de Vossen J, Kartal B, Meier H, Frishman D, Huynen MA, Mewes HW, Weissenbach J, Jetten MS, Wagner M, Le Paslier D (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* 440: 790–794
- Sulakhe D, Rodriguez A, D'Souza M, Wilde M, Nefedova V, Foster I, Maltsev N (2005) GNARE: automated system for high-throughput genome analysis with grid computational backend. *J Clin Monit Comput* 19: 361–369
- Suzek BE, Ermolaeva MD, Schreiber M, Salzberg SL (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics* 17: 1123–1130
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41
- Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci USA* 102: 13950–13955
- The-Uniprot-consortium (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Res* 36: D190–D195

- Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, Storz G (2006) Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res* 34: 2791–2802
- Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, Médigue C (2006) MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* 34: 53–65
- Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS (2005) BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res* 33: W455–W459
- Vernikos GS, Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* 22: 2196–2203
- Vogel J, Sharma CM (2005) How to find small non-coding RNAs in bacteria. *Biol Chem* 386: 1219–1238
- Walport M, Kiley R (2006) Open access, UK PubMed Central and the Wellcome Trust. *J R Soc Med* 99: 438–439
- Wei W, Cao Z, Zhu YL, Wang X, Ding G, Xu H, Jia P, Qu D, Danchin A, Li Y (2006) Conserved genes in a path from commensalism to pathogenicity: comparative phylogenetic profiles of *Staphylococcus epidermidis* RP62A and ATCC12228. *BMC Genomics* 7: 112
- Yada T, Totoki Y, Takagi T, Nakai K (2001) A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Res* 8: 97–106
- Yan Y, Moulton J (2006) Detection of operons. *Proteins* 64: 615–628

CHAPTER 4.3

Data mining in genome annotation

I. Artamonova^{1,2}, S. Kramer³ and D. Frishman^{1,4}

¹German Research Center for Environmental Health, Institute for Bioinformatics and Systems Biology, Ingolstädter Landstraße, Neuherberg, Germany

²Vavilov Institute of General Genetics RAS, Gubkina, Moscow, Russia

³Institut für Informatik, Technische Universität München, München, Germany

⁴Department of Genome Oriented Bioinformatics, Technische Universität München, Freising, Germany

1 Introduction

Systematic annotation of protein sequences was initiated more than two decades ago by two groups of enthusiasts at the Swiss-Prot (Bairoch and Boeckmann 1991) and PIR (George et al. 1986) databases and is actively continuing today as a centralized UniProt effort (see Chap. 10 in this volume). Until the middle of nineties human experts carefully annotated essentially every protein sequence that became known at the highest quality standards. The advent of high-throughput genome sequencing and the unprecedented growth of sequence databanks radically changed the picture. Over the last decade the percentage of manually annotated proteins fell to probably less than 5% (Frishman 2007), and is rapidly continuing to decrease. Instead, the overwhelming majority of amino acid sequences gets annotated by automated software pipelines that systematically apply similarity based methods and various prediction techniques for functional and structural characterization of proteins. While efficient and cheap, electronic annotation suffers from the notoriously high level of errors made by unsupervised algorithms. Available computational methods are unable to reproduce the complex decision process of a human curator, who, while making a decision on a particular assignment, will survey literature, carefully analyze available alignments, and heavily rely on his specific experience and intuition. Typical sources of annotation errors have been reviewed before (Bork and Bairoch 1996; Galperin and Koonin 1998). In addition to fundamental difficulties in annotation transfer by homology (Devos and Valencia 2000), dubious assignments may be caused by spurious similarity hits stemming from compositionally biased protein sequences and failure to take into consideration

Corresponding author: Dmitrij Frishman, Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany (e-mail: d.frishman@wzw.tum.de)

multidomain organization of proteins. Further complications include wrong gene models and unrecognized pseudogenes. Annotation errors systematically pollute sequence databases, leading to the gradual deterioration of the total corpus of available information and undermining further analysis efforts.

The detection of annotation errors has thus become a necessity due to the intrinsic limitations of automatic annotation procedures and the rapidly increasing gap between the number of available sequences and the number of experimentally studied proteins. Given a certain error probability and the independent assignment of annotation items, unlikely and rare combinations of items will arise in practice. As methods from data mining typically aim for regularities and patterns in data, they naturally lend themselves to finding deviations from those patterns, in our case, uncommon combinations of annotation items. The purpose of this chapter is to review data mining methods for the detection of erroneous annotations in molecular biological databases.

Biological databases today present huge collections of measured, derived, and in many cases interlinked data items and so can serve as proper subjects for a wide range of data mining techniques. A generic set of tasks can be found in almost all life science applications of data mining: detecting patterns, regularities, and deviations from regularities in biological data (pattern mining and deviation/anomaly detection), finding groups of objects belonging together (clustering), finding statistically interesting subgroups (subgroup discovery), classifying biological objects into discrete classes (classification), predicting numeric quantities (regression), discovering links among objects (link discovery), and predicting the time course of biological events (time series analysis and process modeling). In the context of protein annotation, typical problems include (a) the discovery of new biological knowledge based on unexpected patterns in sequence and other confirmed (non-derived) data, (b) the generation of new biological annotations by supervised learning methods trained on already annotated entries, and (c) annotation error correction based on the inconsistencies in feature combinations. While problems (a) and (b) have attracted a lot of attention in the past, we are just beginning to understand the third problem (c), the detection of annotation errors in large protein databases.

A high-level overview of the general process is shown in Fig. 1. The first step in the process is the computation of sequence features and functional annotations for a set of protein sequences. In the next step, data mining techniques extract patterns, regularities, and deviations from those patterns for the correction of inconsistencies in annotation items. Following the correction of annotation inconsistencies, the next cycle of the process can be started. In the overall process, each update of the annotation database has to be recorded, to be able to explain and revise previous annotation decisions if necessary.

This chapter is organized as follows: In the second section we give a general overview of large annotation databases. As an example for a manually curated database, we will discuss Swiss-Prot in more detail. As an example for an automatic database, the

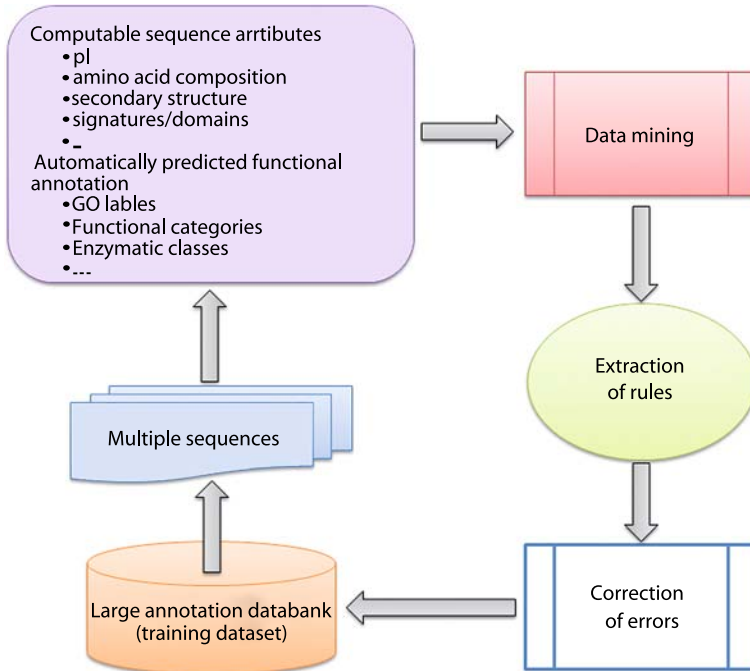


Fig. 1 General overview of the data mining process in genome annotation

PEDANT database system will be presented. In the third section, we review data mining in genome annotation from a technical perspective. Following general remarks on the task, we discuss both supervised and unsupervised learning methods applicable to the problem. Parallel to the presentation in the second section, the following two sections explain the application of association rule mining to Swiss-Prot and PEDANT. The sixth and final section summarizes the main points and lessons learned from applying data mining methods in this area.

2 An overview of large biological databases

2.1 Manually curated vs. automatic databases

Databases contain information on a large set of objects of the same type in form of structured records, or entries. Each record presents information on an individual object via a limited number of pre-defined data fields. The latter may hold one or several terms (attributes), or be empty. The majority of molecular biology databases focus on the

Table 1 Examples of biological databases of interest

Database name	Description	Organisms	Number of entries	Number of data fields	Main fields	Visualization and navigation possibilities	Reference
BRENDA	Manually annotated repository for enzyme data	All (currently 7500)	~1.3 M	40	Enzyme name, suggested name, isolation/preparation, structure, reaction and specificity, stability	Combined search with up to 20 terms; sequence search; EC, TaxTree, Genome and Ontology explorers	(Barthelmes et al. 2007)
PEDANT	Automatic genome analysis	Completely and some partially sequenced genomes (currently 468)	>1.76 M	19	Automatic FunCat, EC numbers, InterPro, GO categories, COG, KOG, known 3D, SCOP, TMHMM, signal peptides, secondary structure	Gene reports, functional and structural categories, genome and protein viewers	(Riley et al. 2007)
Ensembl	Automatic annotation of selected eukaryotic genomes	35 eukaryotic genomes	>650 K (genes)	20–35	Alignments, orthologs/paralogs, protein families, protein features, oligo matches, GO	Genome Browser, DAS servers, BioMart, Blast	(Flicek et al. 2007)
FlyBase	Genomic data for the insect family <i>Drosophilidae</i>	12 completely sequenced <i>Drosophila</i> species	>550 K (genes)	23	Physical maps, cDNA clones and other gene products, synonyms, references, cross-references	Genome, image and interaction browsers, TermLink, annotation viewer/editor Apollo, CytoSearc, Blast, Query builder	(Crosby et al. 2007)
Swiss-Prot/UniProt	A curated protein sequence database and a computer-annotated supplement	All (currently 11074 in Swiss-Prot and 144824 in UniProt)	289473/5035267	18	Keywords, features, comments, references, cross-references	Feature table viewer, feature aligner, multiple cross-links	UniProt (The UniProt consortium 2007)

MEDLINE	World's most comprehensive source of life sciences and biomedical bibliographic information	Non applicable (papers on any organism)	>17M abstracts	31	MeSH terms, abstract, title, authors, journal	Related articles, links to journals, full texts, various types of search	www.pubmed.gov
Protein Data Bank	Depository of three-dimensional structures of large biological molecules	All	>45K structures	15 in summary	Experimental method, molecular description, asymmetric unit, classification, ligand chemical component, SCOP and CATH classifications	Image gallery, KING, Jmol, WebMol, Rasmol and Swiss-PDB viewers, searches by/D, sequence, ligand, models, browsing by EC, MeSH, genome location, SCOP and CATH	(Deshpande et al. 2005)

annotation of genes and proteins. In such databases each record typically corresponds to one gene or protein, and data fields contain attributes pertaining to different aspects of gene function and structure.

How does gene annotation come about? With respect to the information sources, sequence annotation databases may be either automatically generated, or manually curated, or sometimes both. In the automatic databases annotation attributes are generated by computer algorithms in an unsupervised fashion, while curated databases chiefly result from systematic manual collation of experimentally verified facts from literature. Manually curated databases tend to contain fewer entries than automatically produced datasets, but the annotation is more reliable and often serves as the gold standard for benchmarking automatic procedures. By contrast, machine-generated data collections may contain millions of records, but the probability to encounter false information in them is much higher.

There exists an enormous variety of biological databanks, from central depositories of primary information such as GenBank (Benson et al. 2007) to highly specialized resources covering specific biological aspects. The latest database issue of *Nucleic Acids Research* (Galperin 2007) alone lists 968 databases, including datasets as specialized as “a database of orthologous U12-type spliceosomal introns” (Alioto 2007). In Table 1 we give an overview of some of the typical data resources, indicating their size (number of records) and the number of data fields in each record. These databases differ both in their thematic area and the degree of curation. Some of them get thoroughly curated by human experts (BRENDA, Swiss-Prot), others are generated in a completely automatic fashion (PEDANT), others are depositories of experimental information equipped with powerful retrieval and visualization tools (PDB). Many databases are initially produced by automatic software pipelines, and then subsequently get manually curated, verified, and enriched by experimental data (Ensemble, FlyBase). Finally, databases such as MEDLINE are official providers of a give type of data, such as bibliographic references.

2.2 Manually curated databases: the Swiss-Prot example

For a closer look we selected the Swiss-Prot database (The UniProt consortium 2007) which is considered to be the gold standard of protein annotation owing to its high level of manual curation and resulting good data quality. As of the time of writing it contains 289473 protein entries (Release 54.5). As seen in Table 1, Swiss-Prot entries may contain up to 18 types of data fields. Some of them, such as entry identifier, description, and the date of the last modification are obligatory. Other fields may or may not be present dependent on whether the information is available. The most biologically interesting fields – those directly pertaining to the protein function – can contain positional information (e.g. sequence motifs), numerical parameters describing the protein as a whole (e.g., protein length), annotation using controlled vocabulary (e.g. Gene Ontology labels, see below), as well as free-text annotation (Fig. 2).

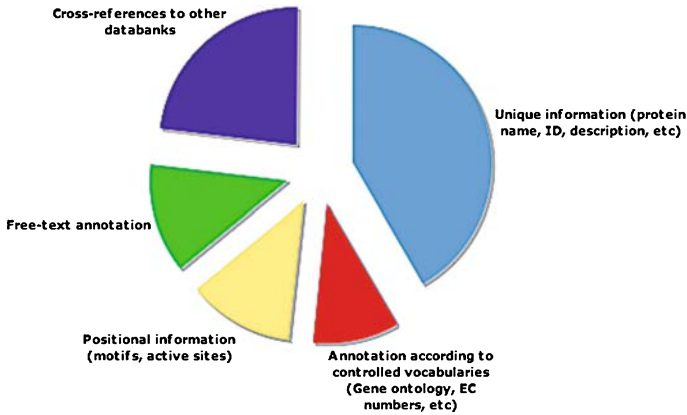


Fig. 2 A breakdown of the Swiss-Prot annotation according to different types of data fields

Positional annotation describing a defined part of the object is stored in the data field called “*Feature table*”. This section contains information on up to 30 different positional features of protein sequences as described in biological literature, such as posttranslational modifications or binding sites. Each term of this type consists of a term name (e.g., TRANSMEM), start and stop positions of the given feature, and a brief description line. Some of these features, while being local in nature, describe general characteristics of the associated amino acid chain. For example, the feature VARSPLIC not only points to the alternative part of the protein but also serves as an indication that the protein is subject to alternative splicing. Although Swiss-Prot is a manually curated database, not all fields are necessarily confirmed by experimental evidence. In some entries the *Feature table* section describes predicted features, in which case they are marked as “hypothetical”.

Numerical parameters often describe physicochemical properties of genes and gene products, theoretically calculated based on their primary structure or measured in experiments. Examples of numerical features are GC-content, protein length, isoelectric point, and biochemical constants.

The annotation relying on a controlled vocabulary is the most suitable part of database annotation for data mining. A given feature of an object can be described by one or several terms from the specially designed list of terms. Usually, every term of the used vocabulary is precisely defined in order to avoid ambiguity. A very interesting and information rich field of this type is *Keyword* which is Swiss-Prot specific. There are almost a thousand individual keywords in Swiss-Prot, such as “Galactose metabolism” or “Kelch repeat”. Database entries may contain more than 20 different keywords, with average being three-four keywords.

Swiss-Prot annotation also includes other general purpose controlled vocabularies, such as Enzyme Nomenclature and Gene Ontology (The Gene Ontology Consortium 2007) assignments. Biological ontologies provide an extremely efficient framework for

structuring and organizing functional information about proteins. They constitute a common language for formalizing knowledge about cellular roles of gene products based on a controlled vocabulary and play a crucial role in streamlining and standardizing annotation work. The Gene Ontology (GO) has become a community standard for annotating genomes of multicellular organisms. GO describes biological roles of genes and gene groups in terms of attributes defined by three major branches: molecular function, biological process, and cellular component. The hierarchical classification of enzymatic reactions known as EC (Enzyme Commission) system is one of the most widely used bio-ontologies. Each EC number is a unique code that describes enzyme activity at four progressively finer levels of detailization. GO terms are available in the *Cross-references* section of any Swiss-Prot entry while the EC numbers are part of the *Description line* in the *Synonyms* subsection.

The free-text annotation can describe a variety of aspects of protein function and structure. Even though it contains concise sentences describing a limited range of object features in a possibly precise fashion, free text annotation is hardly suitable for the majority of data-mining techniques because of the abundance of auxiliary words. The only possibility to extract machine-parseable information from such fields is by sophisticated natural language processing algorithms. In Swiss-Prot the most interesting free-text field is *Comments* where all substantial protein annotation that can not be formalized in other fields is kept. The information is arranged based on 28 topics.

2.3 Automatically generated databases: the PEDANT example

Automatically annotated databases typically contain no free-text information, and the number of computed numerical parameters is higher. It seems more appropriate to distinguish different data fields in automatic annotation by the particular mechanism used to extract or calculate information rather than by their representation. There are three distinct field categories from this point of view.

- Type 1. Features that are definitely known. This group includes either inherent properties of genes and their products, such as their taxonomic origin, or features that can be unambiguously calculated from primary sequences, such as GC content, length, pI value, percentage of low complexity regions, and so on.
- Type 2. Structural and functional properties of proteins predicted directly from their amino acid sequences by *ab initio* computational algorithms (secondary structure, disordered regions, coiled coils, transmembrane segments, signal peptides, cellular localization).
- Type 3. Structural and functional properties of proteins derived by similarity searches against previously characterized gene products. These features include sequence domains, keywords, functional categories, enzyme classes, and functional and structural superfamilies.

As an example of automatic annotation databases let us consider the PEDANT Genome Database (Riley et al. 2007) which contains pre-computed information resulting from bioinformatics analyses of publicly available genomes. The main vehicle for similarity searches is the PSI-BLAST algorithm (Altschul et al. 1997). This method is used for searches against the full non-redundant protein sequence databank as well as against a number of special datasets including the MIPS functional categories (see below) and the COG database (Tatusov et al. 2003). The detection of InterPro domains (Mulder et al. 2007) is performed by profile searches. For those sequences that have significant matches in the UniProt/Swiss-Prot database, the annotation of the respective entries is analyzed and keywords and enzyme classification are extracted. Structural categorization of gene products involves secondary structure prediction as well as PSI-BLAST searches against the sequences with known 3D structure as deposited in the PDB databank (Deshpande et al. 2005) and SCOP database of known structural domains (Andreeva et al. 2004). Other calculated or predicted structural features include molecular weight, pI, low complexity regions (Wootton 1994), membrane regions (Krogh et al. 2001), coiled coils (Lupas 1997), and signal peptides (Bendtsen et al. 2004).

Functional roles of gene products are described in terms of the manually curated hierarchical functional catalog (FUNCAT) (Ruepp et al. 2004). Each of the 16 main classes (e.g., metabolism, energy) may contain up to six subclasses. Correspondingly, the numeric designator of a functional class can include up to six numbers. For example, the yeast gene product YGL237c is attributed to the functional category 04.05.01.04, where the numbers, from left to right, mean transcription, mRNA transcription, mRNA synthesis, and transcriptional control. An essential feature of FUNCAT is its multidimensionality, meaning that any protein can be assigned to multiple categories.

In the current release of the PEDANT database 67% of annotation is of type 3. Information of this type is constituted by the MIPS functional categories, InterPro and SCOP domain assignments, COG families, PIR superfamilies, as well as by keywords and EC numbers transferred by homology to the Swiss-Prot protein entries.

The three categories of automatic annotation outlined above differ in their intrinsic susceptibility to errors. It is obvious that the features of type 1 are unfaultable and cannot generally contain errors (except for incorrectly predicted gene models, typographical errors, or errors caused by software bugs or human error). Features of type 2 are typically predicted with the accuracy in the order of 70% by machine learning techniques, such as neural networks or support vector machines. If no experimental data for a given feature type is available (e.g., known three dimensional structure, experimentally determined cellular localization), such predictions can only rarely be further improved by human curation. Finally, features of type 3 are transferred from one or several previously annotated gene products to the query protein based on a sufficiently significant degree of similarity. These features constitute the main bulk of protein-associated information available in the databases, and it is precisely this part of

protein annotation that is especially prone to errors due to intrinsic limitations of annotation transfer by homology.

3 Data mining in genome annotation

3.1 General remarks

Technically speaking, the automatic assignment of annotation items to a given protein sequence can be done independently for each item, or coupled, in a concerted fashion. In the latter case, unlikely combinations could, in principle, be avoided in the first place. Machine learning algorithms, e.g., for multi-label classification and multi-task learning, could be applied to achieve this goal. Unfortunately, the situation is more complicated in practice, where annotations may originate from manual or automated efforts (e.g., transferred *via* analogy), and inconsistencies may occur due to independently assigned annotation items. Therefore, we have to assume that annotation errors do occur, and methods for their correction (in hindsight) are required.

In the following, we give an overview of data mining techniques for the detection of erroneous genome annotations. Finding errors or unusual events in data is both one of the most useful and least spectacular applications of data mining methods. Work in this area has a long tradition and is typically published under the heading of deviation detection or anomaly detection (see, e.g., the early systems such as KEFIR (Matheus et al. 1996) and WSAR (Wong et al. 2002), many of which focused on temporal trends).

Dependent on the availability of a labeled training set (i.e., a set of sequences with reliably assigned annotation items), three different types of techniques are applicable. In a *supervised learning* approach (such as the one proposed by (Wieser et al. 2004), see below), reliable annotations from a smaller dataset are transferred to a larger unlabeled set. In *semi-supervised learning*, both labeled and unlabeled data are used. However, to the best of our knowledge, semi-supervised learning has not been applied to the problem of mining erroneous genome annotations so far. Finally, *unsupervised learning* approaches detect annotation errors without the help of labeled training sets. The overall goal is to discover unusual combinations of annotation items. Strictly speaking, the problem could be framed as the one of finding all annotation items A_1 to A_k of gene products such that their joint probability $P(A_1, \dots, A_k)$ is smaller than a user-defined threshold. To solve this problem, graphical models like Bayes nets could in principle be used and queried. However, as annotation databases are sometimes quite sparse, it is questionable whether this could be done successfully. Therefore, methods for clustering and association rules have been applied to annotation databases. All of the solutions presented below, however, are just work-arounds, circumventing the problem of determining how likely a given set of annotation items is.

An important aspect of mining annotation databases is the duality of possible annotation errors. Over-annotation causes false positive annotation features, under-annotation causes false negatives. It is important to note that different methods may be necessary to deal with each of the two cases.

The performance requirements on methods applied to databases of genome annotations are in a medium range. As seen in Table 1, molecular biological databases may have millions of entries, making algorithms quadratic in the number of examples impractical. In the following we present a brief overview of supervised and unsupervised learning approaches, recalling the basic notions and presenting selected applications in the context of mining genome annotation.

3.2 Supervised learning

The task in supervised machine learning is to find useful generalizations from observational data in order to make accurate predictions for unseen cases. It is called supervised because the learning setting assumes a teacher who assigns class labels for the learner. Supervised learning is related to the notion of *predictive data mining*, which mostly involves the search for classification or regression models. Traditional machine learning techniques have been devised to find predictive models, that are, at least in principle, comprehensible and human-readable.

In our context supervised learning methods transfer reliable annotations from a smaller dataset to a larger unlabeled set. For example, rules can be learned from a highly curated and reliable database, such as Swiss-Prot, and then used either to further improve annotation in the same database, or in another, automatically generated database, such as TrEMBL. (Kretschmann et al. 2001) applied the C4.5 data mining algorithm to derive decision trees representing the knowledge on Swiss-Prot keywords. Rules obtained in this fashion combined with information on sequence groups gleaned by sequence analysis can be applied both for consistency checks within Swiss-Prot and for generating keywords for new TrEMBL entries with high accuracy. Conversely, exclusion rules for a specific protein group (e.g. sharing the same sequence motif) can be generated by the C4.5 algorithm to detect contradicting annotation items, as implemented in the Xanthippe post-processing system (Wieser et al. 2004).

3.3 Unsupervised learning

In unsupervised learning there is no teacher and no class labels to guide the induction process. It is related to the notion of *descriptive data mining*, where the task is to describe and characterize the data in some way, e.g., by finding frequently occurring *patterns* in the data. Please note that clustering should be categorized as descriptive data mining, although, viewed as density estimation, it could be used for predictive purposes as well.

As no class labels are available, the task is basically much harder, but also much harder to evaluate objectively.

3.4 Clustering

Generally, the task of clustering is to find groups of observations, such that the intra-group similarity is maximized and the inter-group similarity is minimized. There are countless papers and books on clustering, and it is hard to tell the advantages and disadvantages of the respective methods. Part of the problem is that the evaluation and validation of clustering results is to some degree subjective. Clustering belongs to the family of unsupervised learning schemes in the sense that there is no target value to be predicted.

Clustering algorithms can be categorized along several dimensions:

- *Categorical vs. probabilistic*: Are the observations assigned to clusters categorically or with some probability?
- *Exclusive vs. overlapping*: Does the algorithm allow for overlapping clusters, or is each instance assigned to exactly one cluster?
- *Hierarchical vs. flat*: Are the clusters ordered hierarchically (nested), or does the algorithm return a flat list of clusters?

According to this classification, the method by Kaplan and Linal (2005) is a hierarchical agglomerative clustering method which utilizes a measure of similarity between the annotation combinations of each pair of proteins and produces categorical and non-overlapping clusters. Another method, by Kunin and Ouzounis (2005), essentially is a graph-based clustering method producing categorical, non-overlapping, flat clusters. Following the generation of clusters of proteins sharing some degree of annotation similarity, subclusters based on sequence similarity are created.

In the context of our application, annotation mining, clustering methods first group proteins that are thought to be related, e.g., on the basis of sequence information. Subsequently, annotations that are completely distinct or untypical in one of the clusters are inspected, as they indicate possible errors.

More precisely, the basic procedure is to form clusters of proteins based on sequence similarity and/or commonality in annotation items. Errors are then identified and corrected by detecting inconsistencies in the annotation of related proteins forming a sequence cluster. Based on the observation that more than 95% of proteins have more than two annotation attributes, with 10 being the average number (Kaplan et al. 2003) implemented a system that represents protein-keyword relationships in biological databases in the form of a hierarchical graph, each node of which symbolizes proteins sharing unique combinations of keywords. While analyzing protein sets attributed to the same functional category by automated annotation methods, observing certain

proteins occupying areas on the graph that are distinct from the main bulk of the collection clearly points to potential false annotations. More generally, one can define a score that indicates how similar the sets of annotations for any given pair of proteins are. Functionally related proteins are naturally expected to have more similar annotation than unrelated ones. Based on the defined similarity measure, proteins are clustered into groups with homogeneous annotation, the so-called property clusters. This method can be used to detect false positive annotation by any given automatic method aimed, for example, at detection of conserved sequence motifs. The idea is to find those proteins that share the same annotation, e.g., a sequence motif, from the test method, and at the same time form disjointed subsets as a result of clustering in the space of other annotated features. Alternatively, annotation errors can be identified by comparing protein groupings obtained by sequence and annotation clustering. Again, the underlying assumption is that the more sequence-similar proteins are, the higher chance they have to share functional annotation.

3.5 Association rule mining

Finally, methods for association rule mining (Zhang and Zhang 2002) detect certain types of variable dependencies. They return rules of the form $X \Rightarrow Y$, where X and Y are sets of items, in our case annotation items. Consider the set of all possible annotation items I from a database of protein annotations. Then every protein is annotated by a subset $X \subseteq I$ of possible items. Sets of items X are usually called *itemsets* in the literature. From a formal point of view, a *database* D is then defined as a *multi-set* of *itemsets*.

Positive association rules have the form $(A_1 \& \dots \& A_n) \Rightarrow Z$, where A_1, \dots, A_n are the items on the left-hand side (LHS), and Z is a single item on the right-hand side (RHS) of the rule. This rule should be interpreted as: ‘*Database entries that possess all features A_1, \dots, A_n are likely to possess feature Z* ’. Association rules are usually constructed from *frequent itemsets*, i.e., itemsets that occur with a frequency greater than or equal to a user-specified threshold (a parameter known as *support*) in database D . For the computation of frequent itemsets, a vast number of effective methods is known and available today.

A non-standard variant of association rules allows for a limited form of negation in the left-hand side and the right-hand side of rules: *Negative association rules* (Antonie and Zaiane 2004; Wu et al. 2004) are rules of the form $not X \Rightarrow not Y$, $not X \Rightarrow Y$, or, $X \Rightarrow not Y$, where X and Y are again item sets. For the remainder of the chapter, we restrict ourselves to negative association rules with a single negated item on the right-hand side, $(A_1 \& \dots \& A_n) \Rightarrow not Z$, meaning ‘*database entries that possess all features A_1, \dots, A_n are unlikely to possess feature Z* ’. Only the third variant of negative association rules is actually needed for our application of mining annotation databases.

To evaluate positive and negative association rules, we can use several scoring functions. Each rule is characterized by its *coverage*, the number of entries in the

database that satisfy the LHS (possess all features A_1, \dots, A_n), and its *strength* (also known as *confidence*), the fraction of entries that satisfy LHS and RHS among the entries satisfying the LHS. In other words, strength is the probability that an entry will satisfy the RHS given that it satisfies the LHS. An additional very important parameter used to characterize negative rules is *leverage* which is defined as the difference of the rule support and the product of supports of its LHS and RHS. Leverage measures the unexpectedness of a rule as the difference of the actual rule frequency and the probability of finding it by chance with the given frequencies of its RHS and LHS.

When applied to large annotation databases (Artamonova et al. 2005, 2007), methods for association rule mining often detect simple, yet biologically meaningful implications. For instance, in a database of annotated proteins, such as Swiss-Prot, one positive rule is the implication “Nuclear localization \Rightarrow Origin: eukaryota”, i.e., every protein annotated as localized in nucleus has a eukaryotic origin. The rules are not necessarily absolutely strict. For instance, the rule “Alternative splicing \Rightarrow Origin: eukaryota” has exceptions, because viral genes also may be spliced (and alternatively spliced as well). However, this is still a valid rule, because the exceptions comprise a small fraction of the database. Thus, this rule is naturally interpreted as “the majority of proteins with evidence of alternative splicing originate from eukaryotic organisms”. “Many-to-one” rules can also be considered. For instance, “Alternative splicing and Kinase \Rightarrow Origin: eukaryota”. In this example, alternatively spliced proteins are specific to eukaryotic organisms or viruses and kinases belong to either eukaryota or prokaryota. If a protein kinase is annotated as resulting from alternative splicing, then it is an eukaryotic protein.

An example of a trivial biologically relevant *negative association rule* is “Nuclear localization \Rightarrow not bacterial origin”, i.e., every protein annotated as localized in the nucleus cannot have a bacterial origin. As with positive rules, negative rules are not necessarily absolutely strict. For instance, the rule “Operon structure \Rightarrow not eukaryotic origin” has a number of exceptions because bacterial-like operons were described in *Ceanorhabditis elegans* (Blumenthal et al. 2002). Since these exceptions comprise only a small fraction of the annotated genes, this rule may naturally be interpreted as “the majority of genes constituting an operon structure do not originate from eukaryotic organisms”.

On a more abstract level, positive and negative association rules may indicate under-annotation as well as over-annotation in protein databases. The main assumption is that if the database annotations satisfy a rule “ $A \& B \Rightarrow (not) C$ ” with a high support and a very high strength, then such a rule reflects some biological regularity or maybe a peculiarity of the annotation process. If the strength is very close, but not equal, to one, then the rule has a minor number of exceptions. While in some cases such exceptions may reflect biological reality, it is plausible that a significant fraction of them are actual errors in annotation. Hence the strategy of the method developed by Artamonova et al. (2005, 2007) is to find rules of high strength, e.g. in the range (0.95–1), filter them,

identify proteins that are exceptions to such rules, mark the features from the left side of positive rules or from both sides of negative rules and add the right side feature of positive rules to the annotation of such exception proteins.

4 Applying association rule mining to the Swiss-Prot database

The approach of positive rule mining was evaluated on the high-quality Swiss-Prot database, focusing on the most formalized non-overlapping fields of a standard Swiss-Prot entry such as protein length, the highest-level taxon of the protein origin, assignment of InterPro domains, keywords and features from the feature table. The calculation of positive association rules for the annotation set extracted from Swiss-Prot resulted in roughly 3 hundred thousand rules with the strength greater than 0.1 and the minimal coverage count 50. These rules vary greatly in terms of their coverage and strength. For example, the rule “Alternative Splicing & Transmembrane \Rightarrow Eukaryota” extracted from Swiss-Prot has coverage count 1433, support count 1417, and strength 0.989 ($\approx 1417/1433$), indicating that there are 1433 proteins in Swiss-Prot with simultaneously assigned keywords “Alternative splicing” and “Transmembrane” and 1417 of them are of eukaryotic origin. The remaining 16 proteins originate from viruses. On the other hand, the rule “Alternative splicing & Nuclear protein \Rightarrow Repressor” has 129 confirmations from 1288 covered proteins (support count = 129,

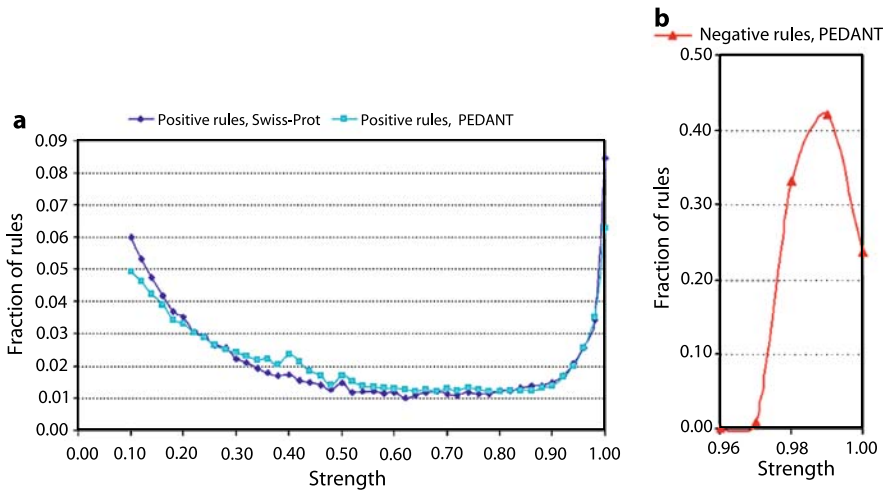


Fig. 3 Distribution of association rules strength for (a) positive rules in the Swiss-Prot and PEDANT annotations and (b) negative rules in the PEDANT annotation. Association rule strength is defined as the probability that a given database entry will satisfy the right side of the rule given that it satisfies the left side of the rule

coverage count = 1288, strength $129/1288 \approx 0.1$), implying that only a small fraction of all nuclear proteins subjected to alternative splicing are repressors, whereas these three keywords taken separately occur frequently among Swiss-Prot protein entries.

A prominent feature of strength distribution of Swiss-Prot rules is the presence of two distinct peaks in the regions of very weak and very strong rules, with rules in the medium strength range being relatively infrequent. Strength distributions of association rules for both databases are shown in Fig. 3a. A large number of weak rules (strength below 0.2) originate from diverse combinations of frequent items, such as the majority of the Swiss-Prot keywords or features. These combinations are typically not wrong, but they do not represent typical associations between items. For example, the Swiss-Prot entry Q6W2J9 (BCoR protein from *Homo sapiens* functioning as transcriptional co-repressor) contains the keywords “Alternative splicing”, “Nuclear protein” and “Repressor” and conforms to the rule “Alternative splicing & Nuclear protein \Rightarrow Repressor”. It has repressor function, localizes in the nucleus, as in the case of the majority of transcription factors, and is subject to alternative splicing. But only a certain fraction of all repressors have multiple alternatively spliced isoforms, and thus this rule is not classified here as a biological regularity.

The other extreme (in Fig. 3a) is constituted by very strong rules with strength values in the range roughly between 0.95 and 1.0. For example, all 1554 proteins annotated with the keyword “G-protein coupled receptor” also have the keyword “Transmembrane” while all 1904 proteins having the feature “MITOCHONDRION” in the FT line also contain the keyword “Transit peptide”.

The approach specifically focuses on the strong rules with the strength over a certain threshold (e.g., 0.95), but below 1.0. These rules are nearly always fulfilled, but exceptions from them do occur in the database. As argued above, such exceptions may constitute annotation errors that can be detected and corrected, or at least flagged, automatically.

Approximately every half year, a new release of the Swiss-Prot database is made available with novel protein entries added as well as some pre-existing entries revised. As one progresses from one release to another, many exceptions to association rules get corrected. One possible way to reveal the corrections introduced by the Swiss-Prot staff is to examine the protein entries that constituted rule exceptions and classify these entries as corrected if in a subsequent database release, either one of the items forming the LHS of the associated rule was deleted from the annotation or the item from RHS was newly introduced or altered.

Figure 4 displays the strength distribution of such corrected rules. In line with the main assumption, the exceptions to the strongest rules get corrected more often. 23.5% of protein entries constituted exceptions to the rules with strength in the range (0.97–1) found in the Swiss-Prot release 44.0 were corrected until the release 47.0, while the average percentage of corrected exceptions to the rules of any strength was 3.47. As an additional test, all 350 exceptions from randomly selected 149 rules in the strength

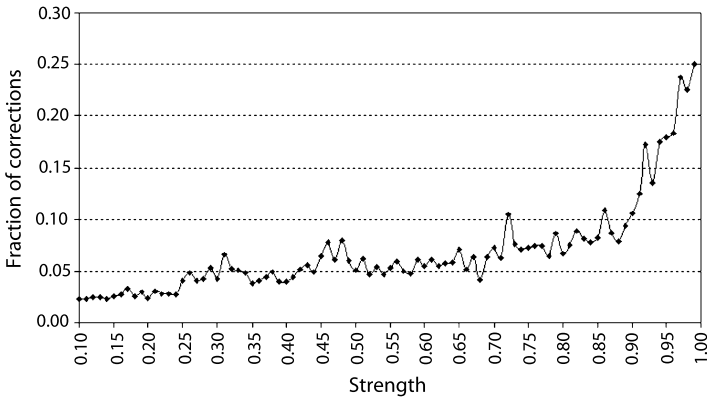


Fig. 4 Fraction of corrected annotations as a function of rule strength ranges in the Swiss-Prot releases 44.0–47.0

range (0.97–1) not corrected by Swiss-Prot staff were subjected to careful manual evaluation by an experienced protein annotator. The exception was classified as an error if one of the items of the LHS of the rule was assigned wrongly to a given protein entry, or the required item on the RHS of the rule was missing. It was found that 24.7% of these exceptions indeed constituted annotation errors. The overall error rate was calculated as (the number of exceptions corrected in subsequent releases + manually verified error rate * number of uncorrected exceptions)/overall number of exceptions. We thus estimate that about 41% of exceptions from strong rules are actually associated with erroneous annotation.

Unfortunately the approach of the negative rule mining cannot be satisfactorily evaluated on the Swiss-Prot data. By design, exceptions from negative association rules can only reveal over-annotation, i.e., erroneous assignment of attributes to protein entries, while under-annotation (missing attributes) cannot be detected. Manually curated databases are typically under-annotated and this approach is not efficient for them. The performance of the method was very low when tested on the Swiss-Prot database.

5 Applying association rule mining to the PEDANT database

The application of the positive association rule mining to the automatically generated PEDANT annotation revealed that statistical properties of the association rules gleaned from the PEDANT database are similar to that of Swiss-Prot whereas the absolute number of positive association rules was much higher for PEDANT than for Swiss-Prot. Overall, almost one and a half million rules were calculated for the strength range

(0.1–1.0). Here, too, strong enrichment of very weak and very strong rules was observed, with rules in the intermediate strength range being relatively rare (Fig. 3a).

Calculation of negative rules for the annotation set extracted from PEDANT resulted in 9591 rules. For example, one of the most trivial rules found was “Bacteria \Rightarrow not Eukaryota”. This rule is satisfied in all possible cases and thus its strength is 1.0 with no exceptions. Much more interesting rules in the context of this approach are those of strength very close, but not equal to 1.0. These rules have a small number of exceptions that may constitute annotation errors. An example of such rules is “Nuclear protein \Rightarrow not Bacteria”. This statement which is obvious from the biological point of view nevertheless does not make an absolute rule; in fact out of all 1808 protein entries annotated by the keyword “Nuclear protein” in the PEDANT database only 1798 actually have eukaryotic origin. The ten proteins constituting exceptions from this rule simply inherit this keyword from their eukaryotic homologs.

Some aspects of negative rule statistics differ significantly from positive association rules due to vastly different item frequencies. Because annotation items themselves are rare, and most items are in fact extremely rare (e.g., the PFAM domain PF01029 is only found in 12 (0.02%) of proteins analyzed), their negations used in negative association rule mining are unavoidably very frequent. This simple circumstance makes the calculation of negative rules computationally much more challenging compared to positive rules and necessitates the application of much stricter thresholds on the rules of interest. While analyzing rule strength distribution only the rules exhibiting strength higher than 0.1 were considered. The number of weaker rules (strength below 0.1) is too high due to the combinatorial explosion caused by random feature combinations, making their analysis computationally prohibitive. However, even in the strength interval 0.1–1.0 the number of negative rules is several orders of magnitude higher than the number of positive rules. To make the task computationally tractable we imposed a threshold on minimal leverage which effectively helps to select only the most ‘non-random’ rules and eliminates all rules with the strength below 0.97. The distribution of negative rule strength is also plotted in Fig. 3b.

Since PEDANT annotation typically never gets corrected manually, and there is thus no release dynamics as in the case of Swiss-Prot, the only way to estimate the amount of errors in the strong rules with exceptions is by manual verification. A randomly selected sample of 144 positive rules in the strength range (0.97–1) was analyzed. About a half of the sample was selected at random, whereas the second half was selected among rules showing at least one exception associated with a protein contained in Swiss-Prot to ensure that this protein is reasonably well documented. The overall number of curated exceptions was 330. The total fraction of exceptions classified as errors in PEDANT was close to 68%.

In contrast to Swiss-Prot only 30% of all errors revealed in PEDANT by manual curation were due to omission of an RHS item. Other errors resulted from false assignment of the items in LHS of the rules, or over-annotation.

In the application of the negative rule mining approach it is interesting to identify errors in the annotation attributes of type 3 transferred by similarity from other proteins. In the PEDANT annotation dataset about 67% of all features were similarity-derived features, more than a half of which were constituted by functional category assignments.

A considerable and arguably the most valuable part of PEDANT annotation involves assignments of functional categories based on the FunCat (Ruepp et al. 2004), a hierarchical catalog of protein functions developed at MIPS. Among all 184 different FunCat labels (2 upper levels of the hierarchy) used in this study 71 were taxon-specific (e.g., fc75.03 – “animal tissue”). It turned out that a very large number of negative rules combined FunCat labels on one side of the rule with the taxon of the protein origin on the other side (here only the highest-level taxons, namely Eukaryota, Bacteria, Archae, and Viruses were used). Thirty three percent of all negative rules had such structure. Homology-based transfer of taxonomic information is highly prone to error. Where a taxonomically specific FunCat label is incompatible with the known gene taxon, it is the FunCat assignment that is guaranteed to be erroneous, since the protein origin is doubtlessly known. We classify such cases as annotation errors according to the general procedure. This simple test resulted in automatic correction of almost 50% of all exceptions of strong negative rules.

To estimate the prevalence of errors among exceptions not corrected by the taxonomy procedure described above we selected randomly a sample of 100 rules and analyzed their exceptions manually. Annotation features of these proteins occurring either in the LHS or in the RHS of the rules were subjected to careful manual analysis by an experienced protein annotator according to the established procedures routinely used at MIPS for genome annotation. An exception was classified as an error if one of the features in the LHS or RHS of the rule was found to be assigned wrongly to the given protein entry. In 96% of examined exceptions at least one of the features constituting the rule was assigned wrongly to the given protein.

The overall specificity of the approach was estimated according to the formula: (percentage of exceptions classified as annotation errors among all manually verified exceptions * number of exceptions in rules not involving ‘taxon specificity’ + number of exceptions from ‘taxon specific’ rules)/overall number of exceptions.

For negative rule approach it was as high as 98%: practically all feature combinations associated with exceptions included at least one annotation error.

The specificity of the negative rules is thus much higher than that in the case of positive rules (68% for the PEDANT database). Over-annotation is a typical problem of many automatic software pipelines, including PEDANT, and the ability to correct this type of errors using negative rule mining is valuable. At the same time, the approach based on exceptions from strong negative rules yields much smaller coverage than positive rule mining. As seen in Fig. 5, negative rule mining allows identifying 11 times fewer annotation features (0.6% for negative rules vs. 6.7% for positive rules) that

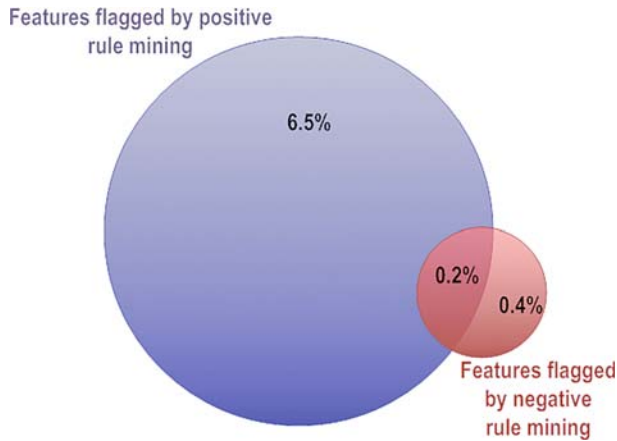


Fig. 5 Coverage of the negative and positive rule mining approaches. The numbers represent the percentage of all annotation features identified as potentially erroneous by each individual method and by both of them

participate in incompatible feature combinations. More than two thirds of these features do not get detected by positive rule mining.

Both of these approaches are designed to flag incompatible feature combinations for subsequent manual inspection rather than to automatically correct annotation errors in an unsupervised fashion. With the exception of taxon-specific rules where FunCat labels incompatible with the taxonomic origin of a protein are guaranteed to be errors, we do not know exactly which feature of a flagged feature combination is wrong. Besides, there always exists a chance that all features constituting an exception from a strong negative rule are nevertheless correctly assigned and that the exception is in fact biologically motivated.

6 Conclusion

Applying a combination of positive and negative rule mining creates an opportunity to enhance the fidelity of genome annotation in two alternative ways. First, insights about the sources of annotation errors gained in this investigation can be used to adjust the automatic annotation pipeline in order to minimize generation of these errors in the future. Examples of such possible modifications include taxon-specific homology-based transfer of functional categories and utilization of individualized similarity thresholds for various features. Second, suspicious features can be visually marked for subsequent inspection by the user. While this approach is better suited for manually curated databases where errors actually get corrected by human experts, it is also useful for

automatic systems such as PEDANT where users get alerted to specific less trusted annotation items that should be used with caution.

References

- Alioto TS (2007) U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res* 35: D110–D115
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402
- Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32: D226–D229
- Antonie M-L, Zaiane OR (2004) Mining positive and negative association rules: an approach for Confined Rules Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2004). Springer, pp 27–38
- Artamonova II, Frishman G, Gelfand MS, Frishman D (2005) Mining sequence annotation databanks for association patterns. *Bioinformatics* 21: iii49–iii57
- Artamonova II, Frishman G, Frishman D (2007) Applying negative rule mining to improve genome annotation. *BMC Bioinformatics* 8: 261
- Bairoch A, Boeckmann B (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* 19(Suppl): 2247–2249
- Barthelme J, Ebeling C, Chang A, Schomburg I, Schomburg D (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res* 35: D511–D514
- Bendtsen JD, Nielsen H, von HG, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340: 783–795
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2007) GenBank. *Nucleic Acids Res* 35: D21–D25
- Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M, Kim SK (2002) A global analysis of *Caenorhabditis elegans* operons. *Nature* 417: 851–854
- Bork P, Bairoch A (1996) Go hunting in sequence databases but watch out for the traps. *Trends Genet* 12: 425–427
- Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res* 35: D486–D491
- Deshpande N, Address KJ, Bluhm WF, Merino-Ott JC, Townsend-Merino W, Zhang Q, Knezevich C, Xie L, Chen L, Feng Z, Green RK, Flippen-Anderson JL, Westbrook J, Berman HM, Bourne PE (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res* 33: D233–D237
- Devos D, Valencia A (2000) Practical limits of function prediction. *Proteins* 41: 98–107
- Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Pric A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S (2007) Ensembl 2008. *Nucleic Acids Res* 36: D707–D714
- Frishman D (2007) Protein annotation at genomic scale: the current status. *Chem Rev* 107: 3448–3466
- Galperin MY (2007) The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Res* 35: D3–D4

- Galperin MY, Koonin EV (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol* 1: 55–67
- George DG, Barker WC, Hunt LT (1986) The protein identification resource (PIR). *Nucleic Acids Res* 14: 11–15
- Kaplan N, Linial M (2005) Automatic detection of false annotations via binary property clustering. *BMC Bioinformatics* 6: 46
- Kaplan N, Vaaknin A, Linial M (2003) PANDORA: keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Res* 31: 5617–5626
- Kretschmann E, Fleischmann W, Apweiler R (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* 17: 920–926
- Krogh A, Larsson B, von HG, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305: 567–580
- Kunin V, Ouzounis CA (2005) Clustering the annotation space of proteins. *BMC Bioinformatics* 6: 24
- Lupas A (1997) Predicting coiled-coil regions in proteins. *Curr Opin Struct Biol* 7: 388–393
- Matheus C, Piatetsky-Shapiro D, McNeil D (1996) Selecting and reporting what is interesting: the KEFIR application to healthcare data advances in knowledge discovery and data mining. AAAI/MIT Press.
- Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database. *Nucleic Acids Res* 35: D224–D228
- Riley ML, Schmidt T, Artamonova II, Wagner C, Volz A, Heumann K, Mewes HW, Frishman D (2007) PEDANT genome database: 10 years online. *Nucleic Acids Res* 35: D354–D357
- Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, Mewes HW (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* 32: 5539–5545
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41
- The Gene Ontology Consortium (2007) The gene ontology project in 2008. *Nucleic Acids Res* 36: D440–D444
- The UniProt consortium (2007) The Universal protein resource (UniProt). *Nucleic Acids Res* 35: D193–D197
- Wieser D, Kretschmann E, Apweiler R (2004) Filtering erroneous protein annotation. *Bioinformatics* 20(Suppl 1): i342–i347
- Wong W-K, Moore A, Cooper G, Wagner M (2002) Rule-based Anomaly Pattern Detection for Detecting Disease Outbreaks Proceedings of the 18th National Conference on Artificial Intelligence. MIT Press.
- Wootton JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 18: 269–285
- Wu X, Zhang C, Zhang S (2004) Efficient Mining of Both Positive and Negative Association Rules. *ACM Trans Inform Syst* 22: 381–405
- Zhang C, Zhang S (2002) Association rule mining. Models and algorithms. *Lecture Notes in Artificial Intelligence*. Springer, Berlin, p 2307

CHAPTER 4.4

Modern genome annotation: the BioSapiens network

C. Yeats¹, Ch. Orengo¹, A. Lise Veuthey², B. Boeckmann², L. Juhl Jensen^{3,5}, A. Valencia⁴, A. Rausell⁴ and P. Bork^{3,6}

¹Structural and Molecular Biology, University College London, London, UK

²Centre Medical Universitaire, Swiss-Prot, Swiss Institute of Bioinformatics, Geneva, Switzerland

³Structural and Computational Biology Unit, EMBL, Heidelberg, Germany

⁴Spanish National Cancer Research Centre (CNIO), Structural Biology and Biocomputing Programme, Melchor Fernandez Almagro, Madrid, Spain

⁵Novo Nordisk Foundation Centre for Protein Research, Panum Institute, University of Copenhagen, Copenhagen, Denmark

⁶Max-Delbrück-Centre for Molecular Medicine, Berlin-Buch, Germany

1 Homologous and non-homologous sequence methods for assigning protein functions

1.1 Introduction

In order to maximise our understanding of biology and evolution, gained from the large scale sequencing projects of the current era, it is necessary to be able to assign detailed biochemical, cellular and developmental functions to as many protein sequences as possible. More than five million distinct proteins can be found in the major public repositories, i.e., UniProt & RefSeq (Pruitt et al. 2007; UniProt Consortium 2007), but detailed laboratory investigations have only been carried out for a tiny fraction. For instance, only ~25,000 proteins have solved structures in the international protein structure repository, the worldwide Protein Data Bank (wwPDB, Berman et al. 2003).

A variety of complementary approaches are being developed to increase the functional annotations of proteins. There are large-scale projects, often on multiple genomes and with bioinformatics researchers working in tandem with laboratory groups, to maximise the amount of functional annotation that can be transferred between proteins. Examples of this include some of the structural genomics initiatives, which aim to provide multiple structures for protein families predicted to have diverse functions. On a more theoretical level bioinformatics researchers are attempting

Corresponding author: Lars Juhl Jensen, Structural and Computational Biology Unit, EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany (e-mail: lars.juhl.jensen@gmail.com)

to build a theoretical understanding of molecular and functional evolution which encompasses genome content together with sequence and structure variation and which will allow the identification of individual functions for proteins as well as predictions of functional associations between proteins.

In the BioSapiens network there is a particular focus on analysing protein sequence and structure data to predict biochemical functions, and on combining the functional annotations provided by different groups participating in the network to improve the coverage and confidence in the assignments. Analysis of a query protein to detect functional sites often involves integrating biochemical data with knowledge of the proteins structure and/or sequence. That is, proteins which have been well characterised experimentally and for which one or more structures are known, can be studied to detect those highly conserved residues likely to be important for function, for instance catalytic residues involved in active sites. Subsequently any insights derived from these analyses can be extended to related sequences.

By comparing protein sequences, using alignment tools like BLAST (Altschul et al. 1990), it is possible to gain a measure of the likelihood that two sequences are evolutionarily related. If two sequences are closely related then it is highly likely that they will exhibit functional and structural similarity. Furthermore, orthologous relatives are much more likely to share similar functions than paralogues (see Sect. 1.2). Several groups have studied the relationship between conservation of enzyme function and sequence similarity and demonstrated a high likelihood of functional conservation for two homologues sharing more than 50% sequence identity (Skolnick 2003).

A more accurate approach than using simple sequence similarity, is to group sequences into homologous families and then subgroup them into functional sub-families (Fig. 1). For each family and subfamily the degree of functional conservation can be estimated and used to generate individual thresholds for transferring functions. As a result many of the current functional inference methods are based around one or more of the various family resources available.

Amongst all the protein family resources that have become publicly available over the last two decades or more, the first immediate distinction is whether they group the whole proteins (i.e. the HAMAP (Gattiker et al. 2003) and eggNOG (Jensen et al. 2007) databases described below) or the domains from which proteins are assembled (i.e. the CATH (Greene et al. 2007) and SMART (Schultz et al. 2007) databases). The existence of distinct domains was initially discovered by studies on three-dimensional structures, and other studies have since confirmed that most proteins are made from combinations of discrete, globular domain units, often with a particular function (ion chelation) or range of functions (kinases). In eukaryotes at least 70% of proteins are thought to comprise multiple domains.

Another distinction between the various protein family resources is whether the families are presented as a hierarchy, with subfamilies representing different levels of similarity and structural and/or functional conservation (i.e. CATH or ProtoNet

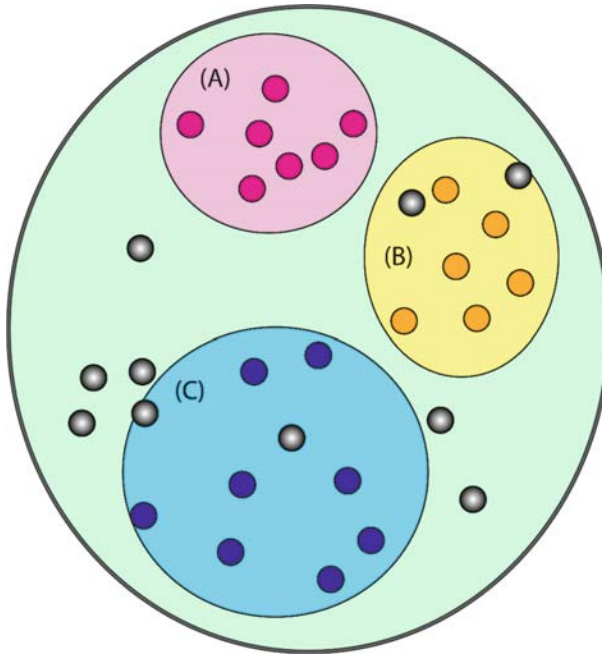


Fig. 1 An idealised view of a protein family is presented. The individual coloured circles are the member proteins and the distances between them reflect their respective sequence similarities. The proteins are coloured according to their enzymatic function, while the uncoloured dots are non-enzymatic homologues. They group according to their function into functional subfamilies (A), (B) and (C)

(Kaplan et al. 2005)) or whether they are ‘flat’ (i.e. SMART or HAMAP). Hierarchical classifications are typically designed to capture different aspects of protein family evolution and hence aid their study. They can also provide a means of choosing alternative similarity levels for safe functional inheritance in different families. However, hierarchical families are often more difficult to create and maintain, and their analysis can sometimes be more problematic, hence the use of single layer classifications by some resources. Where families in these resources are clearly associated with a single function they can be more easily used to provide confident functional assignments.

The third main distinction in protein family resources is in their ‘coverage’ – i.e. the percentage of known proteins that they classify. High coverage largely correlates with a high level of computer automation in the assignment of proteins to the classification and vice versa. For instance, Gene3D (Yeats et al. 2007) essentially uses an all-against-all similarity search and an automated clustering method (Frey and Dueck 2007) to classify most known proteins. In contrast, curators at HAMAP manually construct prokaryotic protein families and assign membership based on detailed analysis and study of the literature.

Somewhere in between these approaches is another common protocol that involves creating manually validated sequence profiles for families, and then applying these profiles automatically together with reliable statistical measures to recruit further sequence relatives into the family (i.e. SMART). Therefore, it is also likely that there is an inverse relationship between coverage and accuracy. It is also worth noting that development of the various types of resource are synergistic, since the automated approaches can use manually curated resources for benchmarking their methods to maximise accuracy, whilst the automatically-produced clusters can help in guiding manual assignments.

This chapter will largely focus on the family-based function prediction resources of the BioSapiens network; however, this is not the only way of exploiting protein sequence similarity information for function prediction and some methods will also be described that use neural networks to infer function from non-homologous proteins, for instance ProtFun (Jensen et al. 2002). The various family resources of the BioSapiens network and their similarities are depicted in Table 1.

1.2 Homologs, orthologs, paralogs. . .

The genomes of species that diverged only recently are highly similar. Chimpanzee and human, for instance, share approximately 98% sequence identity at the nucleotide level. Most human genes have a corresponding gene in chimpanzee and vice versa. These

Table 1 Frequently used terms for homologous genes

Term	Definition
Homologs	Genes of common origin
Orthologs	1. Genes resulting from a speciation event 2. Genes originating from an ancestral gene in the last common ancestor of the compared genomes
Co-orthologs	Orthologs that have undergone lineage-specific gene duplications subsequent to a particular speciation event
Paralogs	Genes resulting from gene duplication
Inparalogs	Paralogs resulting from lineage-specific duplication(s) subsequent to a particular speciation event
Outparalogs	Paralogs resulting from gene duplication(s) preceding a particular speciation event
1:1 orthologs (one-to-one)	Orthologs with no (known) lineage-specific gene duplications subsequent to a particular speciation event
1:n orthologs (one-to-many)	Orthologs of which at least one – and at most all but one – has undergone lineage-specific gene duplication subsequent to a particular speciation event
n:n orthologs (many-to-many)	Orthologs which have undergone lineage-specific gene duplications subsequent to a particular speciation event
Xenologs	Orthologs derived by horizontal gene transfer from another lineage

common genes must have already existed in the last common ancestor shared by the two species, and were subsequently retained after speciation. Genes derived by speciation events are called *orthologs*, and are said to be orthologous to each other. Orthologs of less related species have a lower sequence identity, due to a longer period of separate evolution, although the protein structure and the biologically relevant positions generally remain conserved. The function of orthologous proteins is not expected to change over time, unless the original function becomes obsolete or the function is taken over by another gene product.

In contrast to orthologs, *paralogs* are derived by gene duplication. In some cases, the two initially redundant gene copies are subsequently retained in the genome, which may be explained in two ways (summarized in: 1) neofunctionalization, where the product of one gene copy performs the original function of the ancestral gene, while the product of the other gene copy acquires a new function beneficial to the organism; 2) subfunctionalization, where deleterious mutations in both gene copies enforce the retention of the now complementary functional gene products. However more commonly, one of the genes accumulates deleterious mutations (nonfunctionalization), degrades to a pseudogene (pseudogenization), or is completely lost. More complex models have been developed to explain the ongoing evolutionary process of gene birth and gene death, including the duplication-degeneration-complementation (DDC) model, synfunctionalization, and subneofunctionalization. It is difficult to predict what functional changes may occur after a gene duplication event without a detailed experimental characterization of the duplicates, although paralogs often share a related general biochemical activity. Enzymes and receptors, for instance, can change their substrate specificity and thus contribute to new functional innovations.

Genes originating from a common ancestral gene are called *homologs*, they are *homologous* to each other (see Fig. 2a). According to this definition, homologous groups can include both, orthologs and paralogs. Therefore if you do not know precisely how proteins are related to each other – whether they are orthologs or paralogs – then the term homolog will always be correct – provided of course that they do share a common ancestor. Genes predicted to be homologous are generally grouped into gene families.

In the context of comparative genomics corresponding genes are clustered into 1:1 (one-to-one) orthologs, 1:n (one-to-many) orthologs and n:n (many-to-many) orthologs, according to the number of copies detected in the complete genomes of the given species. *Xenopus tropicalis* and *X. laevis*, for instance, share many 1:2 orthologs due to a whole genome duplication (WGD) in *X. laevis* that occurred about 30–40 mya. A number of species from distinct taxonomic branches have experienced a WGD, including the yeast *Saccharomyces cerevisiae* and there is evidence that ancestral vertebrates have undergone as many as 2 rounds of WGDs. Taking into account the above mentioned WGDs, an ortholog of the last common ancestor of fungi and metazoa theoretically gives rise to a 1:2:4:8 orthologous relationship in e.g. *Schizosaccharomyces pombe*, *S. cerevisiae*, *X. tropicalis* and *X. laevis*. Indeed, it is unlikely to detect

such cases in the relevant genomes, as many gene copies resulting from WGD are lost over time and further lineage-specific gene duplications are added.

Critical for the development of evolutionary genomics was the development of terms to describe relationships between orthologs and paralogs. Three of these terms became especially popular and are described in Table 1: *Inparalogs* originate from lineage-specific duplications subsequent to a given speciation event – they are thus paralogs within a group of orthologs; inparalogs of a particular orthologous group are *co-orthologous* to all other members of this group; gene duplications prior to the radiation of the species under consideration give rise to *outparalogs*. Such relationships are visualized in Fig. 2b: the paralogous vertebrate genes A and B are co-orthologs of the *Drosophila* gene AB; the paralogous vertebrate genes A and B are inparalogs of the

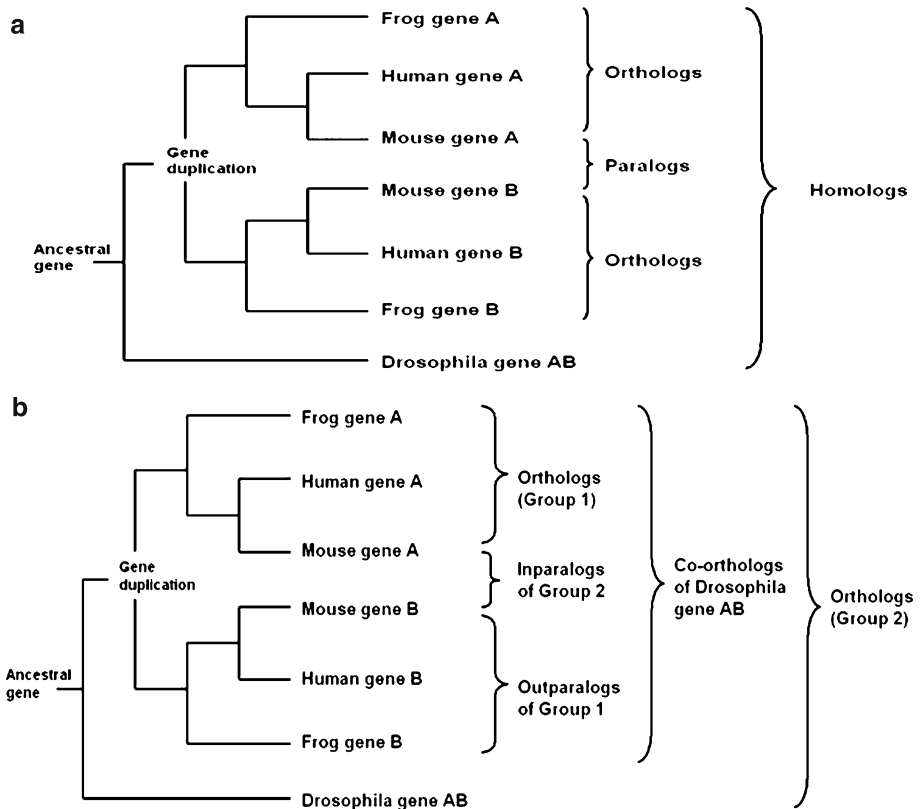


Fig. 2 Schematic illustration of relationships between homologous genes. a) Orthologues and paralogues; b) relationships between orthologues and paralogues. The ancestral gene presents the homologue of the last common ancestor of the compared species

larger orthologous group 2, while the vertebrate genes B are outparalogs of the smaller orthologous group 1 (genes A), etc. Although such relations might occur confusing at first glance, the terms are of great convenience when analyzing large gene families with manifold gene duplications in distinct taxonomic branches.

It remains to be mentioned that there exist 2 different understandings of orthologous protein groups in the scientific community: some scientists do not accept inparalogs in orthologous groups and thus split homologs at each known gene duplication event, resulting in many small orthologous groups (as indicated in Fig. 2a), while others accept inparalogs and form larger groups at the level of interest (Fig. 2b). The latter definition seems to have become widely accepted by many researchers in the field.

In addition to gene duplications, there are many other evolutionary processes leading to the creation of new genes including gene fusion and gene fission. The appropriate presentation of such events is a phylogenetic network, which takes into account multiple parents and the linkage of branches. No specific terms exist that refer to relations between original genes and novel genes.

Genes that are transferred between species *via* horizontal gene transfer (HGT) are called *xenologs*. Horizontal gene transfer events are most commonly found to occur between species that share a common habitat or between hosts and parasites. Evidence of HGT can be detected in a phylogenetic tree, when genes of a given species seem to arise from an ancestral species which is known to be unrelated; for example, it may be observed that a gene from a eukaryotic genome occurs in the branch of an alpha-proteobacterial clade. In this case, it is likely that the gene was derived from the mitochondria, which – according to the endosymbiotic hypothesis – originates from bacteria.

The concept of orthology and paralogy has been developed to define the relationships between genes, but today the same terms are often used for proteins in the context of protein function prediction, where the term ‘ortholog’ generally implies that a group of proteins share the same or a similar function. But is this correct? Regarding proteins, a tree-based classification does not always coincide with a function-based classification, nor does it reflect the fact that most genes produce multiple products and/or biochemically altered forms with distinct functions, such as protein isoforms derived by alternative splicing, post-translational modifications, protein complexes. A hierarchical ortholog classification provides a valuable framework for function prediction, but it is likely to require further graduation for an accurate function assignment of the individual gene products.

1.3 The HAMAP resource for the annotation of prokaryotic protein sequences and their orthologues

Since the publication of the first completely sequenced bacterial genome in 1995, new and cheaper methods to sequence prokaryotic genomes have been developed. It is now

possible to sequence and assemble the whole genome of a bacterium in one day. At the end of 2007, more than 620 bacterial and archaeal genomes had been sequenced and these sequencing projects are the main contributors to the exponential growth of protein databases. Prokaryotic sequences represent the majority of these new sequences, and the quality of both the sequences themselves and the prediction of coding regions is very high. However, the quality of the functional annotation derived from these sequences is very variable.


This influx of sequences from genome projects has drastically changed databases: fifteen years ago most of the available sequences had been submitted by wet labs that were intensively studying a particular protein or a particular region of a gene or operon, and most of these sequences were accompanied by at least some information regarding function, expression, localization in the cell, etc. Most of the sequences available today come from large-scale sequencing projects and will never be experimentally characterised. Nevertheless, with the advent of all the “omics” of the post-genomic era (Proteomics, Metabolomics, etc.) and the development of the concept of systems biology, the demand for corrected and annotated complete proteome sets is greater than ever.

Within UniProt, this large influx of new sequences without experimental characterization has challenged the way proteins have been annotated for at least ten years. Manual curation of protein sequences is a time-consuming, laborious effort and the fast pace of sequencing of prokaryotic genomes was causing a tremendous increase in the backlog of sequences waiting to be manually annotated. In order to try to increase the number of annotated prokaryotic sequences, UniProt implemented **HAMAP** (**H**igh-quality **A**utomated and **M**anual **A**nnotation of microbial **P**roteomes) in 2000, which has greatly changed the way protein sequences are annotated in the Swiss-Prot section of the UniProt Knowledgebase (UniProtKB) (Gattiker et al. 2003).

HAMAP comprises a manual procedure and an automated pipeline. The aim is to automatically transfer annotation based on manually built family rules so that there is no decrease in the quality of annotation. In this way, for each protein family, annotators perform a thorough analysis of the existing literature in order to assess conservation of function and other features and gather the important experimental information that is available for this protein family. They also correct frameshifts, start sites and other problems in the submitted sequences, look for missing proteins in the submitted genome. Finally, they manually define protein families based on similarity searches and the available literature. The automated pipeline is responsible for finding the new members of these manually defined protein families, and for propagating the annotation to them. This set-up allows for greater speed in the annotation of newly submitted sequences without decrease in quality, since the bulk of the annotation work is still performed manually.

Rules are manually created to annotate proteins belonging to well defined protein families or sub-families. These rules contain the information that can be propagated to all orthologues, or to a subset of them (Fig. 3). The use of conditions (for example:

Home Proteomes Families Documents Downloads Links



HAMAP annotation rule: MF_01260

[?] General information about the entry

Accession	MF_01260
Dates	26-JUL-2005 (Created) 28-JUN-2007 (Last updated, Version 8)
Data class	Protein, auto
Predictors	HAMAP: MF_01260, [a substitution of match scores in UniProt-B] [seed alignment for MF_01260]
Identifier	bioH
Description	Carboxylesterase bioH (EC 3.1.1.1) (Biotin synthesis protein bioH)
Gene name	bioH

[?] Comments

- **FUNCTION:** Shows carboxylesterase activity with a preference for short chain fatty acid esters (acyl chain length of up to 6 carbons). Also displays a weak thioesterase activity. Can form a complex with CoA, and may be involved in the condensation of CoA and pimelic acid into pimeloyl-CoA, a precursor in biotin biosynthesis (By similarity)
- **CATALYTIC ACTIVITY:** A carboxylic ester + H₂O = an alcohol + a carboxylate.
- **PATHWAY:** Cofactor biosynthesis, biotin biosynthesis
- **SUBUNIT:** Monomer (By similarity)
- **SUBCELLULAR LOCATION:** Cytoplasm (By similarity)
- **SIMILARITY:** Belongs to the AB hydrolase superfamily, Carboxylesterase bioH family

[?] Cross-references

Pfam	PF00561, Abhydrolase_1_1
PROSITE	PRO0111, ABHYDROLASE_1
TrEMBL	TGR01730, bioH_1

[?] Keywords and Gene Ontology

- **Keyword:** Cytoplasm
- **Keyword:** Biotin biosynthesis
- **Keyword:** Hydrolase
- **Keyword:** Serine esterase
- GO:0004091, Molecular function: carboxylesterase activity
- GO:0009102, Biological process: biotin biosynthetic process

[?] Features

Key	From	To	Description	Condition	[?] Gene
From: B30H_EC01 (P13901)					
ACT_SITE	81	82	Hydrophobic (By similarity)	D	
ACT_SITE	207	207	By similarity	D	
ACT_SITE	235	235	By similarity	D	

[?] Characteristics

- Size range: 239-265 amino acids
- Related UniRules: None
- Template: P13901, Q9K197, Q9GHL1
- Fusion: Nbr: None, Ctr: None
- Duplicate: None
- Plasmid-encoded: None

[?] Comments on the rule

Longer C-terminal in HEIMA, sequence not shown in alignment and not taken into account in size range

[?] Sets of member sequences

Bacteria	[35]
All	[30]

[?] Taxonomic distribution of member sequences in complete prokaryotic proteomes

Fig. 3 An example of the family rule used for automatic annotation. The first section contains annotation propagated to new family members. The features (FT lines) are computed based on their position in the template sequence, and the presence of the specific residues that are crucial to the activity

restriction on the propagation of the annotation to a taxonomic group, dependence on the presence of a conserved active-site amino acid residue at a certain position, etc.) helps to limit the extension of the propagation when it is not safe to assume that the same function, subunit, cofactor, etc. apply to all members of a protein family. If more information becomes available, the rules are updated and the propagation of the annotation is calibrated accordingly.

Each family rule is composed of several sections:

1. Annotations regarding protein name, the name of the gene that encodes for this protein, function, catalytic activity, cofactor, pathway, quaternary structure (i.e., subunit), localization of the protein in the bacterium, etc., and keywords.

2. Position of sequence features, i.e., regions of interest in the sequence, such as active sites, metal-binding sites, domains, post-translational modifications, etc.
3. Cross-references to other databases, and other miscellaneous information, such as size range, presence of multiple genes encoding the protein in a certain organism, whether the protein is encoded on a plasmid in certain bacteria, etc.
4. A manually curated alignment of the representative members of the family. This “seed alignment” is used to automatically generate a profile that detects other possible family members by scanning the UniProtKB (Swiss-Prot and TrEMBL). The alignment is also used to propagate sequence features to newly annotated entries.

The structure of the pipeline is outlined in Fig. 5. After a complete genome is submitted to DDBJ/EMBL/GenBank, its translated CDSs enter TrEMBL, the unreviewed section of UniProtKB. The HAMAP automatic pipeline is then used to annotate additional members, in the following way: protein sequences from UniProtKB/Swiss-Prot and UniProtKB/TrEMBL are scanned against the HAMAP profile collection on a daily basis. True matches are annotated using the corresponding family rule. Many checks are performed in order to prevent the propagation of wrong annotation and to spot problematic cases, which are channeled to manual curation. The results of this annotation are integrated into UniProtKB/Swiss-Prot.

At the end of 2007, 1450 family rules were available on the HAMAP web site (<http://www.expasy.org/sprot/hamap/>); these rules were used to annotate approximately 40% of all UniProtKB/Swiss-Prot entries, and this number shows the efficiency of the approach. The HAMAP pipeline is being used to annotate proteins from Bacteria, Archaea and from plastids (i.e. chloroplasts, cyanelles, apicoplasts, non-photosynthetic plastids), and a similar pipeline is being developed to annotate eukaryotic proteins.

The essential point that must be emphasized is that the vast flow of data arising from complete genome projects does not have to cause a decrease in the quality standards of curated databases in order to hastily incorporate the sequence information. On the contrary, meaningful classifications and interpretation of the data require a reliable core of verified and structured knowledge. Still, databases are under pressure to keep up the pace and provide their services to the community without delay. The HAMAP pipeline stands as an example that high quality standards and high sequencing throughput are not incompatible.

1.4 CATH, Gene3D & GeMMA

The CATH database (Greene et al. 2007; <http://cathwww.biochem.ucl.ac.uk/latest/index.html>) was one of the first protein domain structure classifications to become publicly available, with all the domains identified from three-dimensional data deposited in the Protein Databank (PDB). CATH curators systematically ‘chop’ the

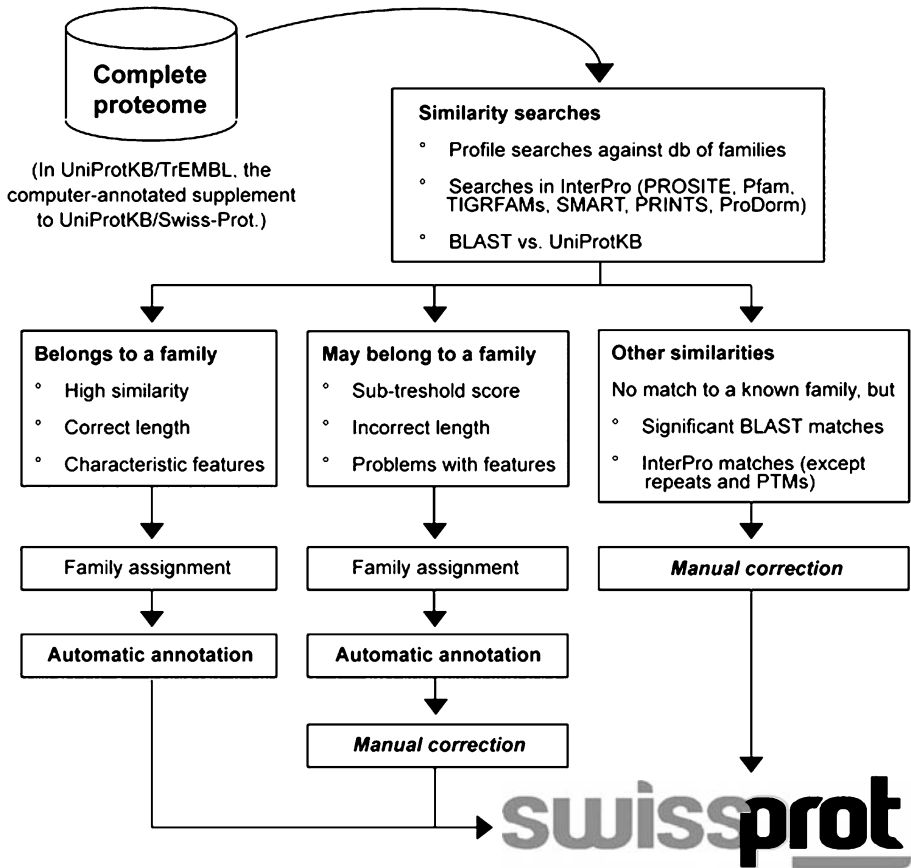


Fig. 4 The HAMAP pipeline. Complete microbial proteomes in TrEMBL are searched against the family profile collection, and confident matches are automatically annotated based on the corresponding family rule. Only in the case where no warning is produced may the entry enter UniProtKB/Swiss-Prot directly without manual verification. The remaining entries, which are still the vast majority of entries in any given proteome, are subject to classical manual annotation, but may also serve to ‘seed’ new family rules

protein structures into their constituent domains using various automatic and manual protocols. This includes applying the CATHEDRAL domain structure comparison method (Redfern et al. 2007) to recognise structural similarity between domains. CATH is an acronym for the four major levels of the classification hierarchy (C)lass, (A)rchitecture, (T)opology and (H)omology.

The top half of the hierarchy is essentially based on structural considerations, for example describing the alpha helical or beta strand composition of proteins and the arrangement of these secondary structure elements in 3D space (the ‘Class-Architecture-Topology’) whilst the lower half is determined through sequence similarity

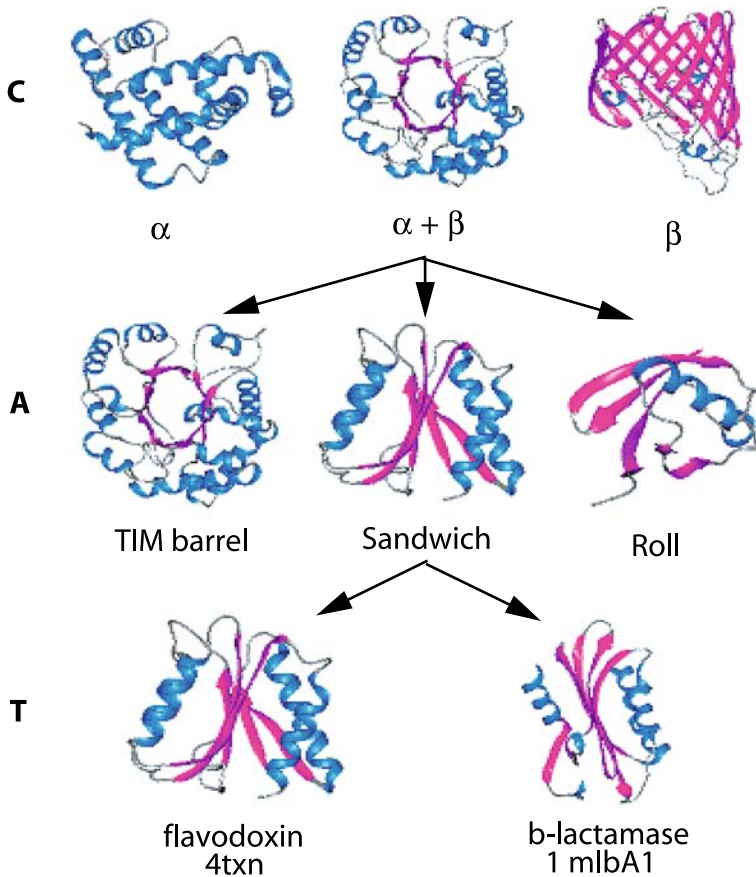


Fig. 5 The CATH structural hierarchy. Domain structures are extracted from the PDB and grouped according to structural features. The first three levels of the hierarchy are displayed: (C) – CLASS. The secondary structure composition of the domain (i.e. mostly helices). (A) – ARCHITECTURE. The three dimensional arrangement of the secondary structures irrespective of their order in the chain. (T) – TOPOLOGY. Assigned according to the overall shape and connectivity of the secondary structures. The final curated level is the H-level. When there is evidence that strongly supports two domains being evolutionarily related then they are put in the same homology group

amongst homologous relatives. See Fig. 5. for a detailed description. The intermediate H-level is the family level, since this is the layer in the classification that indicates that the members are evolutionarily related. Relationships at this level are usually recognised using both structural and sequence data. Each domain is assigned a position in the hierarchy, again using a combination of computational tools and expert analysis.

Since structural data is more reliable for identifying domain boundaries than purely sequence based data, the CATH domain sequences are used to seed profiles for detecting relatives in genome sequences. Representative sequences from each CATH family and subfamily are used to automatically create sets of representative multiple sequence alignments, for each Homology (H) family. From these, Hidden Markov Models (HMM) are generated, which can then be used to scan large public sequence libraries, or those of individual researchers, and assign regions of proteins to CATH domain families. The assignments of CATH domains to UniProt sequences are displayed in the Gene3D website and are available for download.

Gene3D (<http://gene3d.biochem.ucl.ac.uk/>) also integrates functional annotation from resources like the Gene Ontology (GO Consortium 2007), FunCat function descriptions (Ruepp et al. 2004), protein-protein interaction data from MPact (Guldener et al. 2005) and IntAct (Kerrien et al. 2007) and other domain and protein family resources, including Pfam (Finn et al. 2007). Pfam contains a collection of manually curated sequence alignments, which are used to generate profile HMMs, as with CATH. Since Pfam is not restricted to domains that have been structurally defined it is able to cover a larger proportion of proteins than structure-based approaches; however, structure-based resources are able to identify more distant homologues. By integrating the two resources it is possible to significantly increase overall domain family coverage in genome sequences (see Fig. 6).

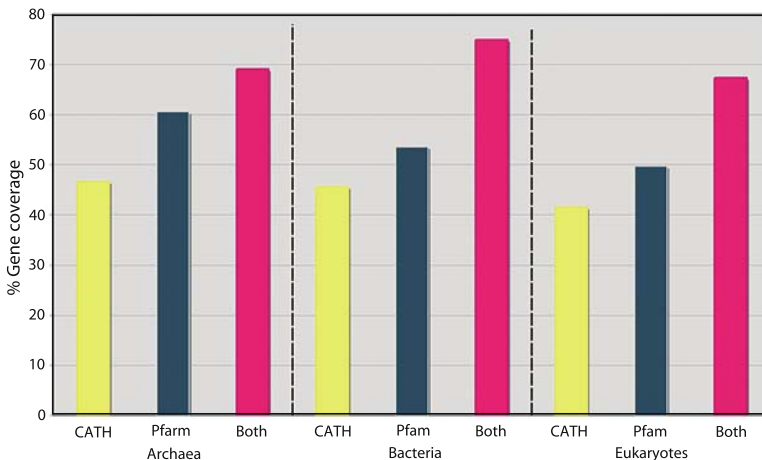


Fig. 6 Combining different resources to maximise coverage. Since it is possible to identify more distantly diverged homologues from structure data than sequence data (in general), the CATH database is able to identify a larger number of homologues per family than Pfam. However, since there is far more sequence data, Pfam is able to identify a greater number families. By combining the two resources, 10% extra sequences per genome are given at least one domain family assignment compared to Pfam alone, and 20% compared to CATH/Gene3D alone

The merging of functional resources in Gene3D with the mapping of structural domains allows the assignment of functions to particular combinations of domains. For newly identified proteins with that particular combination of domains the annotation can be directly transferred providing there is sufficient sequence similarity. Several groups have shown that pair-wise sequence identity thresholds of 50% (60%) can be used to inherit biochemical functions (e.g. enzyme annotations) between homologues with a reasonable error rate of 10% (5%). However, recent research using Gene3D has shown that within CATH enzyme families, 90% of domain pairs sharing the same multi-domain architecture have the same function (to 3 Enzyme Classification (EC) levels) at 30% sequence identity (see Fig. 7). Higher levels of sequence similarity improve the accuracy and specificity of functional annotation that is transferred. However, most analyses suggest that family specific thresholds are more reliable than generic thresholds for inheriting function and future releases of Gene3D will provide family specific thresholds for all well populated CATH-Gene3D enzyme domain families.

Although, as mentioned already, the safest method for transferring function between relatives is to identify orthologues (see Sect. 5.5.1 and 5.6.5) this can often prove problematic for eukaryotes and an alternative approach is being developed in the GEMMA protocol for Gene3D. GEMMA divides Gene3D sequences assigned to CATH

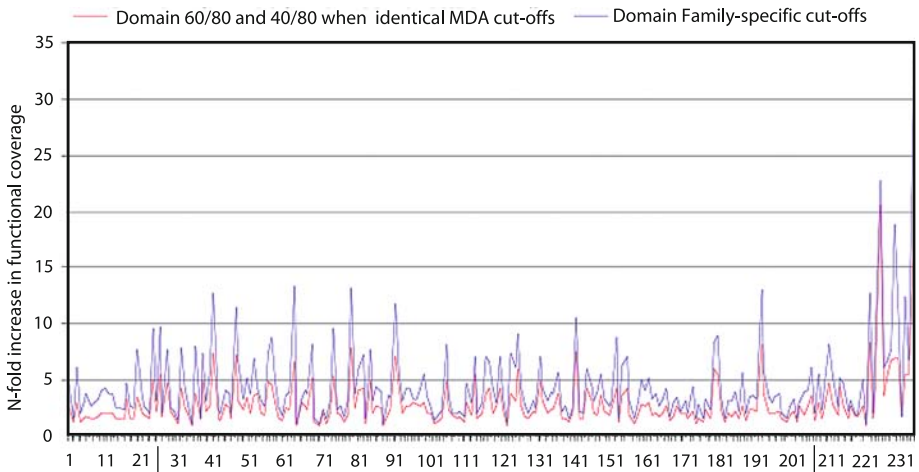


Fig. 7 Increase in enzyme annotation of 240 genomes in Gene3D. Increase in the number of uncharacterised sequences receiving an EC annotation using the optimum average sequence identity thresholds (i.e. above which a 95% conservation is observed across all enzyme families) (red) and family-specific derived sequence identity thresholds (blue) derived from the same comparison. The EC code is only transferred if the two sequences share the same domain composition. If this is the case then a lower sequence identity threshold can be used and more sequences annotated – although those with a unique domain architecture can't

H-level families into functionally coherent subfamilies. The method uses a simple iterative process to generate clusters of similar proteins. Individual sequences are used as seeds at the beginning of the protocol and searched against each other using the Compass profile-profile comparison tool. An optimal E-value threshold of sequence similarity was identified (normally $< 1 \times 10^{-50}$) to merge sequences into clusters. Once clustered, sequences are automatically realigned and the representative subfamily profiles rebuilt. The profiles are searched against each other again and merged using the inclusion threshold, and the process repeated until no more clusters merge. Benchmarking using a manually curated dataset has shown that improved levels of functional annotation of genomes are possible using the GEMMA functional families compared to the use of simple generic pair-wise sequence identity thresholds.

GeMMA will be used to provide more sophisticated function-based subfamilies for both the Gene3D protein families (described below) and the extended CATH domain families. The method was developed in collaboration with groups involved in the Protein Structure Initiative (PSI) structural genomics initiatives in the States and is being used to identify functional subfamilies which currently have no close structural relatives ($\geq 30\%$ sequence identity) to target for structure determination. Hopefully, as more structures are determined it will be possible to understand how structures and functions diverge within superfamilies families and use these insights to improve the function prediction methods.

In order to provide information on whole protein families as well as domain families and increase functional annotation through inheritance across protein families, Gene3D also carries out a large-scale protein clustering of the RefSeq and UniProt. Since these are very large resources, the data are first reduced by removing very similar sequences ($> 85\%$ identity) using a fast clustering method. Affinity propagation clustering is then used to group the reduced set of representative sequences into families of similar domain architecture. To complete the coverage of proteins the very similar sequences are then added back into the families.

This results in a final set of $\sim 200,000$ protein family clusters of more than 3 sequences each, with 1500 families having more than 100 members. Around 16% of proteins are left as 'singletons' – i.e. no clear, confident homology relationships can be defined by sequence comparison. Each family has been sub-clustered into subfamilies at 10 different sequence identity levels to provide a finer grain view of it and to enable function inheritance with varying degrees of confidence.

Therefore, functional assignment between relatives in Gene3D can be made by using information from whole protein families generated by the APC clustering or by using the domain based families identified by mapping CATH and Pfam domains onto genome sequences. A higher coverage of functional annotation can be obtained by applying both protein and domain based family inference. Figure 7 shows the increase in functional annotations that could be obtained by using superfamily specific thresholds for protein and domains sequences in Gene3D.

As well as predicting individual protein functions, family information in Gene3D can be exploited to predict functional associations between proteins. This is exemplified by the Gene3D based PhyloTuner phylogenetic profiling method (Ranea et al. 2007). For this approach, predicted CATH domain sequences in Gene3D were sub-clustered within each superfamily at 10 levels of sequence identity between 30 and 100%. At each level the number of representatives in a given family is counted in each genome to create a domain occurrence profile for that family. Occurrence profiles for all the families can then be searched against each other to identify profiles that show correlated patterns in domain number. From this, it is possible to identify protein families and subfamilies that are likely to be co-evolving and hence, likely to be functionally associated.

A suite of bioinformatics tools, based on Gene3D families, are being developed for predicting functional associations between families. The Gene3D-Biominer predictive system integrates a set of inferred protein-protein functional associations derived from multiple resources. The current components include: CODA, inferring interacting proteins by detecting homologues that have fused into a single protein in other species; hiPPI, inheriting protein-protein interactions through the Gene3D protein families; GEC, gene expression clustering for identifying co-expressed groups of genes; and PhyloTuner, already described above.

On their own all these methods have their own problems and biases. However, since two independent methods should make different errors (i.e. incorrect predictions) then merging the predictions will remove likely false positives whilst retaining correct predictions that are only weakly supported by each individual method. A statistical approach is being developed for Gene3D-Biominer to merge these results in a way to will maximise sensitivity and specificity.

1.5 From SMART to STRING and STITCH: diverse tools for deducing function from sequence

One of the oldest sequence profile-based domain identification tools is **SMART** (Simple Modular Architecture Research Tool) that started off with mobile signalling domains (Schultz et al. 1998) and has then been gradually extended to include a variety of domain signatures (see Letunic et al. 2006 and references therein). With more than 2500 citations distributed over six papers describing different aspects of the resource, it seems to be widely used, mostly by biologists who appreciate the intuitive interface. SMART not only reports the domain architecture, but also identifies other features in a sequence such as signal peptides, GPI anchors, coiled coil and transmembrane regions as well as compositionally biased segments and repeats.

Furthermore, SMART imports functional and structural signals from a number of other resources, most importantly Pfam, to increase coverage and annotation. One of the resources used to identify repeats is the **REP** program (Andrade et al., 2000) that can

also be used separately on the service site (<http://www.embl.de/~andrade/papers/rep/search.html>); it extrapolates from known repeats thus allowing more sensitivity in regions next to known repeats.

While domain annotation is very sensitive in the evolutionary sense as it reveals very distant homologies, one has to be cautious with function transfer and this often needs to be manually supervised. As discussed already above, much more fine-grained function annotation for a given gene can be carried out when an ortholog with known functional features can be assigned to a query gene. For this purpose we have developed **eggNOG** (an evolutionary genealogy of genes and Non-supervised Orthologous Groups; Jensen et al. 2008). It contains a hierarchy of orthologous groups that were derived by a similarity-based triangulation procedure, either based on seeds from the COG database (Tatusov 2003) or constructed de novo and amended by an automatic function assignment step using keyword mining (Jensen et al. 2008; <http://eggno.embl.de>).

Orthologous groups are required for a variety of different function transfer strategies. They are also at the core of **STRING** (Search Tool for the Retrieval of Interacting Genes and Proteins; von Mering et al. 2007) a resource that allows users to assign protein interactions and network context to a gene for which, amongst other inputs, a sequence is sufficient for a quick lookup (often via orthologous groups). There are several visualization modules around STRING, **MEDUSA** (Hooper and Bork 2005) being perhaps the most advanced one. STRING is explained in more detail in chapter X and allows users to annotate higher-order functions to genes in contrast to homology-based methods such as SMART that assign mostly molecular functions.

The interaction between proteins and chemicals is another type of functional annotation, which is not obviously obtained using current homology or network-based approaches. We have therefore recently launched a sister resource of STRING called **STITCH** (Search Tool for Interactions of Chemicals, <http://stitch.embl.de>), which offers a network perspective based on chemical protein interactions). It integrates data from several automatically and manually derived resources of protein-chemical interactions, including **MATADOR** (Manually Annnotated Target and Drug Onlne Resource, <http://matador.embl.de>), which contains a high quality reference set of human protein-drug relationships. Starting from one or more protein sequences, STITCH can construct a network of proteins, drugs and other small molecules that both enriches the functional annotation and gives the context of pharmacology and diseases.

There are also more direct routes to link sequences to related diseases. One is the popular tool **PolyPhen** (Ramensky 2002) that predicts functional consequences of point mutations in a protein sequence using reference structures and evolutionary constraints on each amino acid position (Sunyaev et al. 2001). This way, it can predict the likely involvement of mutations in disease. Another strategy to predict disease-related proteins based on sequence is to explore their context in the human genome. As numerous linkage analysis studies have been published, these can be integrated with other information resources as done by **G2D** (Genes to Diseases; Perez-Iratxeta et al.

Table 2 The SMART-STRING-STITCH list of function prediction resources

Program	Description
SMART	Database for domain analysis [Schultz et al. PNAS 1998; Letunic et al. NAR 2006; http://smart.embl.de]
REP	A resource for repeats analysis [Andrade et al. JMB, 2000; http://www.embl.de/~andrade/papers/rep/search.html]
eggNOG	Database of orthologous groups of genes [Jensen et al. NAR 2008; http://eggnog.embl.de]
STRING	Database of known and predicted protein-protein interactions in all known genomes [Snel et al. NAR, 2000; von Mering et al. NAR 2007; http://string.embl.de]
Medusa	A general graph visualization tool [Hooper and Bork Bioinformatics 2005; http://www.bork.embl.de/medusa]
STITCH	Database of known and predicted interactions of chemicals and proteins [Kuhn et al. NAR 2008; http://stitch.embl.de]
MATADOR	A resource for protein-chemical interactions [Günther et al. NAR 2008; http://matador.embl.de]
PolyPhen	Prediction of functional effect of human nsSNPs [Ramensky, NAR 2002; http://www.bork.embl.de/PolyPhen]
G2D	Analysis of candidate genes for mapped inherited human diseases [Perez-Iratxeta et al. Nat Genet 2002; Perez-Iratxeta et al. NAR 2007; http://www.ogic.ca/projects/g2d_2/]

2007), a server that specifically exploits literature information to associate phenotypic and genotypic information [http://www.ogic.ca/projects/g2d_2/].

Table 2 provides concise descriptions of these approaches. Although they all make use of sequence data, they utilize different strategies and combine the sequences with a wide range of other indicators to annotate biological function at different scales.

1.6 General approaches for inheriting functions between homologous proteins

The function of certain types of proteins is carried out by a small number of residues localized in a specific region of the protein structure. In enzymes, for instance, the catalytic reaction depends on a small number of residues located in the active site. That is the case, for example, of serine proteases where a Ser-His-Asp triad performs the key steps in catalysis. Specific functional sites are part of larger binding sites, e.g. substrate-effector binding sites. More generally, functional sites form part of protein-protein interacting regions, small ligand binding sites, nucleic acid binding sites, etc. Functional residues can be thought of as those required for the protein to carry out its molecular function or biological role. A change of the amino acid type in these positions have a potential effect on the protein function and the fitness of the organisms, as is the case of some diseases associated with point mutations. In the same way, this potential effect could also be directed to provide a modified protein function, which has a benefit for biotechnology. Therefore, the determination of the set of residues in a protein that is

responsible for its function is crucial to understand its molecular mechanism of action and, eventually, carry out a useful modification.

A major repository of functional residues is the database Catalytic Site Atlas (Porter et al. 2004; <http://www.ebi.ac.uk/thornton-srv/databases/CSA>) which provides catalytic residue annotation for enzymes in the Protein Data Bank (PDB). It contains two types of sites: a hand-annotated set of residues extracted from the original literature and an additional homologous set containing annotations inferred by sequence similarity searches and sequence alignment to the original set. The recently developed FireDB resource (López et al. 2007; <http://firedb.bioinfo.cnio.es/>), brings together both ligand binding and catalytic residues in one database for functionally important residues. FireDB integrates data from the close atomic contacts in PDB crystal structures and reliably annotated catalytic residues from the Catalytic Site Atlas. Clusters of PDB structures with 97% of sequence identity are then used to derive a sequence consensus where functional residues are filtered and mapped together. Additionally, FireDB is linked to FireStar (López et al. 2007b), a server for predicting ligand-binding residues in protein sequences, that uses the sequence templates provided by FireDB. Further information about FireDB and FireStar is given in Chap. 4.6.

A number of computational methods predict functional residues based on multiple sequence or structure alignments (MSAs) of homologous proteins (those sharing a common ancestor). These alignments are biologically relevant sources of information since they allow the comparison of equivalent residues between proteins within a family and hence the detection of amino acid changes allowed or precluded by evolution at each position due to structural or functional requirements. The first indicators of functionality extracted from sequence alignments were associated with fully conserved positions (Zuckerkindl and Pauling 1965): these positions are interpreted as important residues for the function of the protein, since they have been preserved during the evolutionary process. Fully conserved positions are associated with all types of functional sites: active/catalytic sites, protein–protein interacting regions, etc. However, conservation is not always related to function and can also be due to structural constraints, for example for positions constituting the structural core or driving the folding of the protein.

Interestingly, the concept of conservation can be extended to the subfamily level, that is: positions that are conserved within subfamilies, with the amino acid type being different between subfamilies. In a well-established manner, these subfamily-dependent conserved positions are related to functional specificity. That is, they are associated with the functional features that distinguish the subfamilies from each other, in contrast to fully conserved positions that are associated with the function common to the whole family. A general model for illustrating the relationship between fully conserved and subfamily-specific positions is shown in Fig. 8. In this case, subfamilies derived from the three main branches of the phylogenetic tree fit the three main functional specificities within the family.

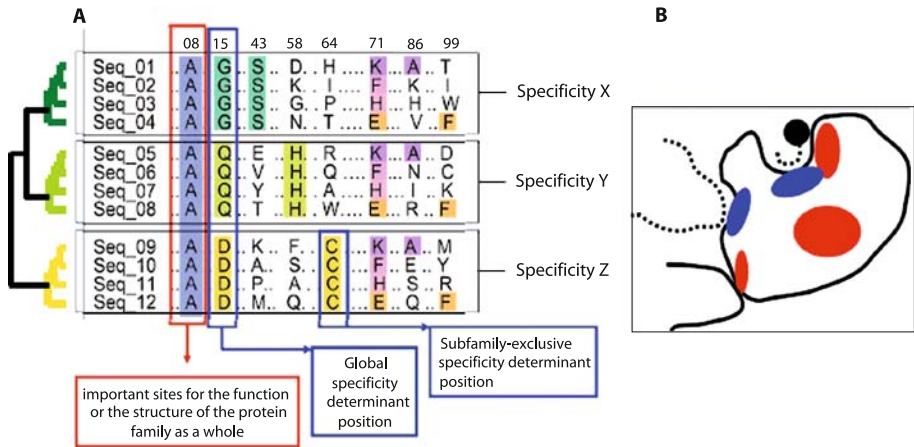


Fig. 8 Information extracted from multiple sequence alignments (MSAs) related with protein structure and function. **A**) Fully conserved and family-dependant conserved positions (taken from Pazos and Bang, 2006a) showing the relationships between these positions and functional and structural features. Conserved positions (red) are present in structural cores (due to structural reasons) and active sites. Subfamily specific positions (blue) are also present in sites close to conserved positions (e.g. determining specificity for substrates with slightly different characteristics) and in other parts of the protein related with specificity, like protein–protein interaction sites (reflecting the interaction with different partners)

A considerable number of methods that use evolutionary information to predict functional specificity residues have been developed. They have been tested independently for a number of different biological systems and the predictions have been experimentally validated in a number of cases, and these approaches are a growing area of interest in computational biology. An illustrative case is shown in Hernández-Falcón et al. (2004) and de Juan et al. (2005) where some of these methods were used to identify crucial residues in an important biological problem: the dimerization of chemokine receptors. Chemokines coordinate leukocyte trafficking by promoting oligomerization and signaling by G protein-coupled receptors. Using evolutionary based sequence analysis in combination with structural predictions, two residues were selected as important for dimerization of chemokine receptor CCR5 and further experimentally validated.

Following del Sol Mesa et al. (2003), these methods can be classified into three major categories: i) methods based on the comparison of sequences using phylogenetic trees as a guide (Lichtarge et al. 1996; del Sol Mesa et al. 2003; Reva et al. 2007); ii) methods that exploit the correlation between the variation at the residue level with the global family variability (del Sol Mesa et al. 2003; La et al. 2005); and iii) methods based on the decomposition of multiple sequence alignments using variants of principle component analysis (Casari et al. 1995; del Sol Mesa et al. 2003). Conservation and family-dependent conservation are sometimes combined with structural information to

restrict the predictions to the positions with the structural characteristics expected for a functional site (Aloy et al. 2001; Armon et al. 2001; Landgraf et al. 2001; Pupko et al. 2002; Kinoshita and Ota 2005; Yu et al. 2005; Glaser et al. 2006).

Interestingly, the TreeDet Server (A Carro et al. 2006; <http://treedet.bioinfo.cnio.es/>) has the strength of combining the results of three separate methods for the prediction of functional specificity residues, corresponding to each of the three former major categories, respectively: the Entropy based method (SS method), the Mutational Behavior method (MB), and the Fully Automated Sequence Space method (FASS). In addition, the server provides a tool (SQUARE; Tress et al. 2004) for evaluating the reliability of the multiple alignments that are used to extract the evolutionary information. These three methods, plus SQUARE, provide a unified view of the possible residues of functional interest and allow a systematic exploration of the sequence space.

It is interesting to think that it is also possible to define subgroups within a family according to criteria different from standard phylogenetic separation, for example functionally, phenotypically, by its cellular localization, etc. The interpretation of subfamily-dependent conserved positions will depend on the feature used to define them. A new generation of methods takes an external functional classification as input instead of the usual sequence-based divergent evolution to function scenario. In other interesting scenarios, certain specific situations can lead to a disagreement between the alignment-based classification and the functional classification of the proteins: i) Many functional and structural requirements drive the evolution of a protein family together, but only one phylogeny can be observed, which arises from a combination of all the different constraints. Hence, the specific divergence owing to a function of interest can be masked within this composite phylogeny. ii) When the alignment does not reflect the true phylogeny, e.g. in structural alignments linking distant proteins for which much of the sequence information relating the proteins has been lost (e.g. SH3 domains). iii) Finally, there may be convergent evolution in some specific parts of the protein.

To deal with these situations, Pazos et al. (2006b) presented two supervised methods for detecting functional sites from multiple protein alignments that can incorporate an external functional classification. One of the methods (Xdet) detects positions in the alignment for which the amino acid type composition better correlates with one predefined functional distance between proteins. The other method (MCdet) is based on a vectorial representation of the alignment for which Multiple Correspondence Analysis (MCA) is used to locate the residues that better fit the pattern of presence/absence of a given function. Both methods were successfully tested in different scenarios where the functional/phylogenetic disagreement arises from different causes.

Importantly, the sets of predicted residues provided by the existing unsupervised and supervised approaches can be used to assign proteins to functional classes, as done in e.g. Hannenhalli and Russell (2000) or Krishnamurthy et al. (2007). This possibility is particularly interesting in cases where function and phylogeny do not correlate, since the functional assignment cannot be done by the standard sequence similarity based

methods. Taken together, methods that predict functional residues can aid the functional classification of the ever increasing amount of new sequences and structures.

1.7 Non-homologous methods for predicting protein function from sequence

The biological function of a protein is ultimately defined by its three-dimensional structure and its native environment. Unfortunately, deducing function from structure alone is a non-trivial task and proteins with known 3D structure are still relatively rare. Sequence information, on the other hand, is widely available and the function of a protein can often be inferred from sequence homology to other proteins with known functions. Obviously, this approach relies on the availability of annotated homologues. An alternative strategy is based on the observation that proteins contain signals and properties determining their cellular processing and biological role. This means that proteins with the same function tend to exhibit similar feature patterns and functional similarity can be deduced from biochemical and biophysical properties. Examples of these are global properties such as average hydrophobicity, charge, and amino acid composition as well as local features like protein glycosylation, phosphorylation, and other post-translational modifications. Additionally, secondary structure content or the presence of transmembrane regions and targeting signals may constitute important features.

Non-homologous function prediction was first implemented in the ProtFun method for human proteins (Jensen et al. 2002, 2003; <http://www.cbs.dtu.dk/services/ProtFun>). A schematic ProtFun work-flow is shown in Fig. 9. The input features are

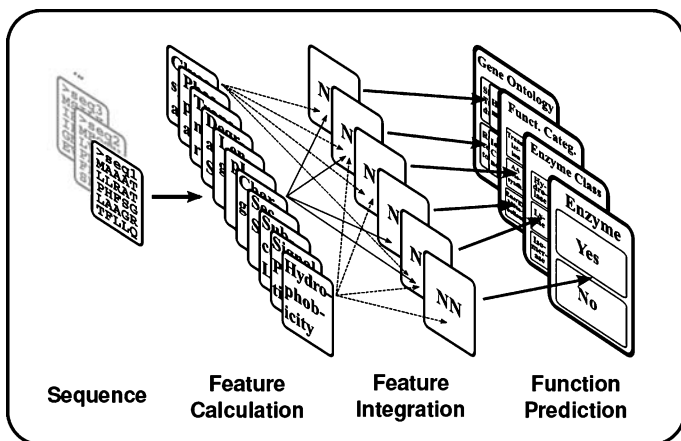


Figure 9

calculated from the amino acid sequence and used by ensembles of artificial neural networks (NN) trained to predict each of the individual function types. These predictions are then combined and ranked within four major groups: enzyme/non-enzyme, enzyme class, biological function, and a subset of Gene Ontology. The latter is represented by the categories signal transduction, receptor, hormone, structural protein, transporter, ion channel, voltage-gated ion channel, cation channel, transcription, transcription regulation, stress response, immune response, growth factor and metal ion transport. Prediction of each distinct function relies on a characteristic combination of protein features, which has been established by a feature selection process during training of the predictor.

By design, the strength of the ProtFun method lies in classification of unannotated and orphan proteins. More recent methods have adopted a ProtFun-like approach in combination with homology or structural input and report improved performances, particularly in prediction of the Gene Ontology categories (Pal and Eisenberg 2005; Lobley et al. 2007). Newer methods presumably benefit from the increasing quality and quantity of functional annotation of proteins. Furthermore, the combination of non-homologous prediction methods with homologous or structural methods is likely to overcome limitations inherent to each individual method.

Non-homologous approaches are not limited to prediction of human protein function and have been successfully tailored to tackle other problems such as prediction of archaeal protein function (<http://www.cbs.dtu.dk/services/ArchaeaFun>) and non-classical and leaderless secretion of proteins (<http://www.cbs.dtu.dk/services/SecretomeP>).

References

- Aloy P, Querol E, Aviles FX, Sternberg MJE (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol* 311: 395–408
- Altschul SF, Madden TL, Schäffer AA et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402
- Andrade MA, Ponting CP, Gibson TJ, Bork P (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *J Mol Biol* 298: 521–537
- Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* 307: 447–463
- Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10: 980
- Bisbee CA, Baker MA, Wilson AC, Haji-Azimi I, Fischberg M (1977) Albumin phylogeny for clawed frogs (*Xenopus*). *Science* 195: 785–787
- Boeckmann B, Blatter MC, Famiglietti L, Hinz U, Lane L, Roehert B, Bairoch A (2005) Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *C R Biol* 328: 882–899
- Brown SD, Gerlt JA, Seffernick JL, Babbitt PC (2006) A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol* 7: R8

- Carro A, Tress M, de Juan D et al. (2006) TreeDet: a web server to explore sequence space. *Nucleic Acids Res* 34: W110–W115
- Casari G, Sander C, Valencia A (1995) A method to predict functional residues in proteins. *Nat Struct Biol* 2: 171–178
- Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S, Rocchi M, Eichler EE (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437: 88–93
- Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69–87
- de Juan D, Mellado M, Rodríguez-Frade JM et al. (2005) A framework for computational and experimental methods: Identifying dimerization residues in CCR chemokine receptors. *Bioinformatics* 21: ii13–ii18
- del Sol Mesa A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues. *J Mol Biol* 326: 1289–1302
- Fleischmann RD, Adams MD, White O et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496–512
- Finn RD, Tate J, Mistry J et al. (2007) The Pfam protein families database. *Nucleic Acids Res* 36: D281–D288
- Fitch WM (1970) Distinguishing homologous from analogous proteins. *Sys Zool* 19: 99–113
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531–1545
- Furlong RF, Holland PW (2002) Were vertebrates octoploid? *Philos Trans R Soc Lond B Biol Sci* 357: 531–544
- Gattiker A, Michoud K, Rivoire C et al. (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput Biol Chem* 27: 49–58
- Gitelman I (2007) Evolution of the vertebrate twist family and synfunctionalization: a mechanism for differential gene loss through merging of expression domains. *Mol Biol Evol* 24: 1912–1925
- Glaser F, Morris RJ, Najmanovich RJ et al. (2006) A method for localizing ligand binding pockets in protein structures. *Proteins* 62: 479–488
- Greene LH, Lewis TE, Addou S et al. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 35: D291–D297
- Güldener U, Münsterkötter M, Oesterheld M et al. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res* 34: D436–D441
- Hannenhalli SS, Russell RB (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol* 303: 61–76
- Hernanz-Falcón P, Rodríguez-Frade JM, Serrano A et al. (2004) Identification of amino acid residues crucial for chemokine receptor dimerization. *Nat Immunol* 5: 216–223
- He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169: 1157–1164
- Hooper SD, Bork P (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics* 21: 4432–4433
- Jensen LJ, Gupta R, Blom N et al. (2002) Prediction of Human Protein Function from Post-translational Modifications and Localization Features. *J Mol Biol* 319: 1257–1265
- Jensen LJ, Gupta R, Stærfeldt HH, Brunak S (2003) Prediction of human protein function according to Gene Ontology categories. *Bioinformatics* 19: 635–642
- Jensen LJ, Julien P, Kuhn M et al. (2007) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 36: D250–D254
- Kaplan N, Sasson O, Inbar U et al. (2005) ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res* 33: D216–D218

- Kerrien S, Alam-Faruque Y, Aranda B et al. (2007) IntAct-open source resource for molecular interaction data. *Nucleic Acids Res* 35: D561–D565
- Kinoshita K, Ota M (2005) P-cats: prediction of catalytic residues in proteins from their tertiary structures. *Bioinformatics* 21: 3570–3571
- Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39: 309–338
- Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* 55: 709–742
- Krishnamurthy N, Brown D, Sjölander K (2007) FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evol Biol* 8: S12
- La D, Sutch B, Livesay DR (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins* 58: 309–320
- Landgraf R, Xenarios I, Eisenberg D (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol* 307: 1487–1502
- Letunic I, Copley RR, Pils B et al. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34: D257–D260
- Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257: 342–358
- Lobley A, Swindells MB, Orengo CA, Jones DT (2007) Inferring Function Using Patterns of Native Disorder in Proteins. *PLoS Comp Biol* 3: e162
- Lopez G, Valencia A, Tress ML (2007a) FireDB – a database of functionally important residues from proteins of known structure. *Nucleic Acids Res* 35: D219–D223
- Lopez G, Valencia A, Tress ML (2007b) Firestar – prediction of functionally important residues using structural alignments and alignment reliability. *Nucleic Acids Res* 35: W573–W577
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155
- Ouzounis C, Perez-Iratxeta C, Sander C, Valencia A (1998) Are binding residues conserved? Pacific Symposium on Biocomputing 3: 399–410
- Pal D, Eisenberg D (2005) Inference of Protein Function from Protein Structure. *Structure* 13: 121–30
- Pazos F, Bang JW (2006) Computacional prediction of functionally important regions in proteins. *Current Bioinformatics* 1: 15–23
- Pazos F, Rausell A, Valencia A (2006) Phylogeny-independent detection of functional residues. *Bioinformatics* 22: 1440–1448
- Perez-Iratxeta C, Bork P, Andrade-Navarro MA (2007) Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res* 35: W212–W216
- Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32: D129–D133
- Pruitt KD, Tatusova T, Maglott DR (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35: D61–D65
- Pupko T, Bell RE, Mayrose I et al. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18: S71–S77
- Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 30: 3894–3900
- Ranea JA, Yeats C, Grant A, Orengo CA (2007) Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes. *PLoS Comp Biol* 3(11): e237
- Reva BA, Antipin YA, Sander C (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* 8: R232
- Ruepp A, Zollner A, Maier D et al. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* 32: 5539–5545

- Schultz J, Copley RR, Doerks T et al. (2000) SMART: a Web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* 28: 231–234
- Sonnhammer EL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 18: 619–620
- Sunyaev S, Ramensky V, Koch I et al. (2001) Prediction of deleterious human alleles. *Hum Mol Genet* 10: 591–597
- Tatusov RL, Fedorova ND, Jackson JD et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41
- Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278: 631–637
- Tress ML, Graña O, Valencia A (2004) SQUARE—determining reliable regions in sequence alignments. *Bioinformatics* 20: 974–995
- The Gene Ontology Consortium (2007) The Gene Ontology project in 2008. *Nucleic Acids Res* 36: D440–D444
- The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res* 35: D193–D197
- Valdar WS, Thornton JM (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* 42: 108–124
- Valencia A (2005) Automatic annotation of protein function. *Curr Opin Struct Biol* 15: 267–274
- von Mering C, Jensen LJ, Kuhn M et al. (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 35: D358–D362
- Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387: 708–713
- Yeats C, Lees J, Reid A et al. (2007) Gene3D: comprehensive structural and functional annotation. *Nucleic Acids Res* 36: D414–D418
- Yu GX, Park BH, Chandramohan P et al. (2005) In silico discovery of enzyme-substrate specificity-determining residue clusters. *J Mol Biol* 352: 1105–1117
- Zuckermandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) *Evolving genes and proteins*. Academic Press, New York

CHAPTER 4.5

Structure to function

J. D. Watson¹, J. M. Thornton¹, M. L. Tress², G. Lopez², A. Valencia²,
O. Redfern³, C. A. Orengo³, I. Sommer⁴ and F. S. Domingues⁴

¹EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

²Structural Biology and Biocomputing Program, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

³Department of Biochemistry and Molecular Biology, University College London, London, UK

⁴Max-Planck-Institut für Informatik, Saarbrücken, Germany

1 Introduction to protein structure and function

Protein structural models are usually obtained by two experimental methods, X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). These models play a central role in the investigation of the molecular basis of protein structure and function. Based on these models it is possible to identify the secondary structure elements (secondary structure) and the spatial arrangement of the polypeptide chain (fold or tertiary structure). They also reveal which atoms or amino-acid residues are buried and which are at the protein surface. The local spatial arrangement of atoms and residues can be analysed in these models, and their chemical environments characterised. Many of these models are especially informative regarding the molecular function of the respective proteins. A typical example is when they include several interacting polypeptide chains, or when they include small molecules bound at the protein surface which act as natural ligand, substrates or as inhibitors. In particular, the determinants for interaction affinity and specificity can be investigated and molecular mechanisms of binding and catalysis can be inferred using these models. Structural models can also be predicted using computational approaches as described in the “Protein Structure Prediction” chapter. In general predicted models are less reliable than experimental models, but there has been considerable improvement in the quality of predicted models over the last years.

The current section describes computational methods for predicting function based on structural models. Obviously, each of these models is very informative, but there are additional factors encouraging the development of computational tools for structure

Corresponding author: Francisco S. Domingues, Max-Planck-Institut Informatik, Campus E1 4, 66123 Saarbrücken, Germany (e-mail: doming@mpi-sb.mpg.de)

based functional characterisation. Experimental structural models are becoming available for an increasing number of proteins. In this respect, structural genomics projects have played an important role in determining the structures of large numbers of proteins (Blundell and Mizuguchi 2000). These large scale projects have also generated many structural models for proteins with uncharacterised function, which are the main application targets for the methods described in this section. In addition, with the availability of large amounts of structural data it is now possible to train and test automated prediction methods. These methods are not only applicable to uncharacterised proteins, so when applied to the structures of annotated proteins they can provide additional insights on protein function, such as identifying new functions in multitasking proteins.

Traditionally, functional annotation has relied on the identification of evolutionary relationships (homology) based on sequence similarity (see “Sequence to Function” section). The identification of a homologous relationship between two proteins indicates a possible functional relationship. The increased availability of structural models provides new opportunities for functional annotation based on homology. As protein structure tends to be more conserved in evolution than sequence, backbone structure comparison methods have been applied in detecting homology between proteins when their sequence similarity is not significant. Backbone structure comparison methods are now routinely used in the comparison between newly determined structures and the models available in the Protein Data Bank (PDB), the archive of structural models determined experimentally (Berman et al. 2000).

Protein functional sites share certain characteristic properties. These common properties have been taken into account in the development of the different function prediction strategies. Proteins tend to bind other proteins, or other types of molecules when performing their function. They tend to bind other macromolecules (other proteins or nucleic acids), small organic compounds or metal ions at specific regions on the protein surface (binding sites). Enzyme active sites tend to locate in the neighbourhood of the substrate binding sites. Interacting binding sites tend to be complementary, both in terms of chemistry and shape, which results in specific binding. Small ligand binding sites are usually located at the largest surface clefts, which provide both high accessibility and complementarity. Residues in functional sites are usually less tolerant to mutations than other residues at the molecular surface, as the structural/chemical integrity of functional sites needs to be preserved in order for proteins to remain functional. Therefore functional sites usually correspond to clusters of highly conserved residues at the protein surface. Finally, it has been observed that in some cases of proteins with similar function there is considerable local structure conservation at their functional sites.

There are three main challenges to take into account in the investigation of structure/function relationships: localization, characterisation and classification. Given the structure of an uncharacterised protein, the first natural question is where the

functional site is. Several methods have been proposed to address this challenge, many of them taking into account that functional residues tend to be conserved and that they tend to cluster at the protein surface. A more difficult challenge is then to characterize the type of molecular function associated with the functional sites that have been identified. In this respect some approaches based on local structural comparison have been proposed. Finally, developing functional site classification databases is a considerable challenge, which requires reliable classification approaches that are able to process increasing amounts of structural information. These classification databases are becoming increasingly valuable to structural biologists, as functional sites are characterized for an increasing number of proteins. These classification databases play a fundamental role in the development and testing of new function prediction methods, and they are powerful tools for the investigation of the structural basis for protein function.

The current section describes four complementary resources for structure-based function prediction. First we describe FireDB (Lopez et al. 2007b), a functional site database, and firestar (Lopez et al. 2007c), a method that relies on FireDB for predicting functional residues. Two approaches for characterising the protein function based on statistical learning are then described. GOdot (Weinhold et al. 2008) is based on sequence and backbone structure similarity, while FLORA relies on local structure similarity. Finally, a combined approach for function prediction, ProFunc (Laskowski et al. 2005a) is described.

2 FireDB and firestar – the prediction of functionally important residues

2.1 Introduction

Genome sequencing projects are generating an almost unimaginable quantity of sequences. Very few of these sequences have been studied experimentally and as a result the vast majority of the sequences in the databases have poorly characterised function. One of the biggest challenges facing bioinformatics is to provide functional annotations for these proteins.

The predominant approach to function annotation is to search for homologous protein sequences for which function is known and simply transfer the functional annotation. However, the homology-based transfer of functional annotation based solely on the similarity of two sequences is not always reliable (Devos and Valencia 2000; Todd et al. 2001). While two homologous enzymes that have less than 30% sequence identity are almost certainly going to be structurally similar, there is likely to be significant functional differences. At higher identities small differences in residue composition can easily lead to changes in substrate specificity and there are even

recorded cases of single residue mutations leading to functional differences (Wilks et al. 1988). There is clearly a need for more sophisticated functional assignment techniques.

Protein function can be defined at many levels. A protein's function may be defined by its role in the cell, the metabolic pathway or regulatory network that it forms part of or by the physico-chemical effects that it brings about. It is perhaps at the level of individual residues where it is possible to be more specific about the function of a protein. Most function depend on physico-chemical processes that are mediated by amino acids, so amino acids that have catalytic activity or bind substrates are directly implicated in molecular function. Pinpointing residues of functional interest is especially important for studying function at biochemical and cellular levels and designing experiments.

Transferring residue-level functional information requires alignments between the target sequences and the functionally characterised sequences. While alignments are not critical in the transference of general protein function, they are important when residue level function is being transferred. It is important to assess the reliability of the alignment when transferring information from functional residues. We have addressed these problems by integrating databases of experimentally validated functional residues with sequence analysis tools. The goal is to establish an automatic system to annotate, validate and predict protein functional sites.

2.2 FireDB

FireDB (Lopez et al. 2007b) is a data bank of functional information relating to proteins of known structure. It contains the most comprehensive and detailed repository of known functionally important residues, bringing together both ligand binding and catalytic residues in one site. The sources of functional residues are the reliably annotated catalytic residues from the Catalytic Site Atlas (Porter et al. 2004) and biologically relevant protein-ligand atom contacts that are filtered from the close atomic contacts in PDB structures. Known solvents and other artefacts in the PDB are removed at this stage. The PDB is highly redundant, so proteins sharing at least 97% sequence identity have been clustered and each cluster in the database is represented by a unique "consensus" sequences. The functional information associated to each member of a cluster is collapsed onto the cluster consensus sequences.

Collapsing means that functional residues spanning equivalent positions are considered to belong to the same site. This process has the advantage of allowing functional sites to be compared within a cluster of sequences and gives an idea of the flexibility of binding sites and their capacity for binding different ligand analogs. The information that can be retrieved includes the type of site, a chemical description of the ligand, the list of chains that bind the ligand and the residues involved in binding. A sample of the FireDB output is shown in Fig. 1.

The database is updated monthly. As of October 4, 2007, FireDB contained a total of 93,559 chains in 18,711 clusters, of which 11,358 had associated bound ligands or catalytic sites.

2.3 Firestar

Firestar (Lopez et al. 2007c) is an expert system for predicting functionally important residues. The server provides a method for extrapolating from the large inventory of functionally important residues in FireDB and using them to predict likely functional residues. For *firestar* to work it must be possible to generate an alignment between the user-defined query sequence and a sequence of known structure that has functional information. Alignments in *firestar* are generated with PSI-BLAST (Altschul et al. 1997).

Firestar automatically generates alignments against the consensus sequences of the FireDB database and maps the FireDB functional residues onto the user's query sequence. The server uses a version of SQUARE (Tress et al. 2004), a method developed to predict regions of reliably aligned residues in sequence alignments, to determine which residues are reliably aligned. SQUARE is particularly effective at predicting the conservation of ligand binding residues (Tress et al. 2003). A multiple alignment interface allows the user to align the consensus sequences found by *firestar*, and to dynamically highlight functional residues. Binding site residue variations can also be viewed via LGA structural alignments (Zemla 2003) and the molecular visualisation tool, Jmol (<http://www.jmol.org/>).

The method is highly sensitive and we have been able to show that even if the only sequences with functional information found by PSI-BLAST are highly distant homologues, *firestar* can still locate functionally important residues. One disadvantage is that the results depend to a large extent on the quality of the alignment. For that reason *firestar* also allows the user to input their own alignments and structural alignments where possible.

The results are presented in a series of easy to read displays (Fig. 1b) that allow users to select the functionally important residues they are interested in and a multiple alignment option to allow comparison of functional residue conservation across homologous proteins.

One of the advantages of *firestar* is that it has allowed us to validate the biological relevance of small ligands in FireDB. A bound small ligand is considered to be biologically relevant (and not an artefact of the crystallisation conditions) if the residues that bind it can be found in a homologous protein in FireDB. The conservation of ligand binding residues is a strong sign of biological importance. A new interface has been added to browse this information and to compare with other information such as ligand nature and the size and residue composition of related sites. With the server it is easy to discern whether small molecule binding is conserved in homologous structures.

This facility was particularly useful during the function prediction assessment of the recent CASP7 experiment (Lopez et al. 2007a). This section of the CASP blind structure prediction experiment concentrated on the prediction of protein function and 22 groups from around the world predicted function for 104 “target” sequences. One of the tasks set by the organisers was the prediction of small ligand binding residues and as assessors we had to evaluate the ligand binding residues predicted by the participants. The assessors have access to the recently solved structures of the target sequences with bound ligands and one crucial part of the assessment was to determine whether the bound ligands in the target structures were biologically relevant or just a consequence of their crystallisation process. Here *firestar* was invaluable (Fig. 2). With *firestar* we were able to draw two conclusions from the CASP experiment – firstly that biological

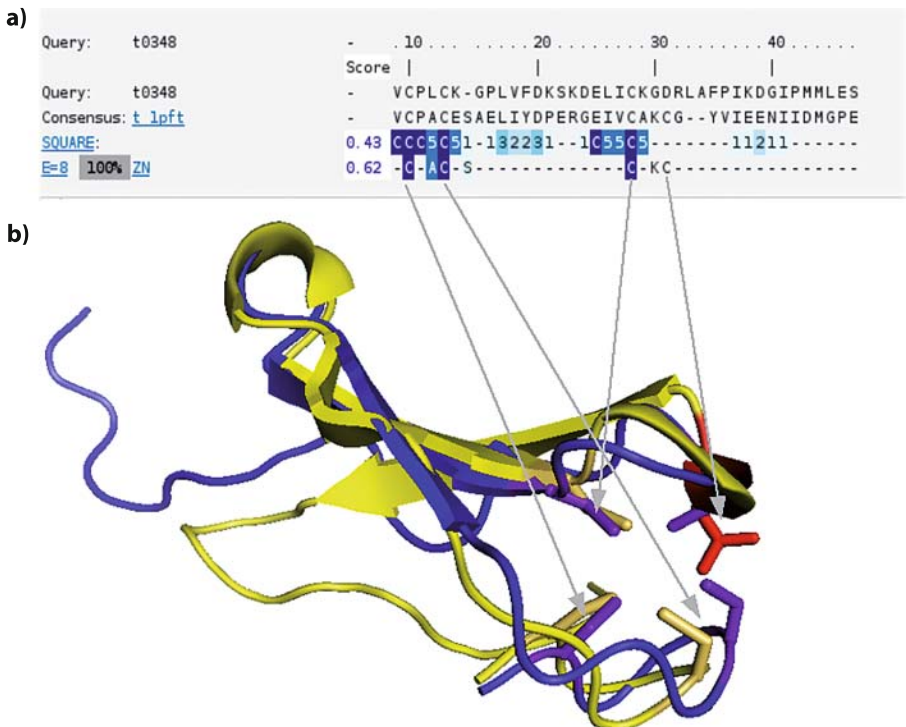


Fig. 2 *Firestar* prediction of a conserved zinc finger in CASP target T0348. A zinc atom was found in the structure of target T0348 in CASP7. As assessors we had to determine whether this zinc was biologically relevant or not. (a) Poor scoring BLAST alignment between the target and the template 1pft from the FireDB data bank. 1pft also binds zinc and the cysteines that bind zinc in 1pft are conserved in T0348 (the darker the blue, the more conserved the local environment of each residue). (b) The structural superposition of both the target and the template showing that the three cysteines conserved in sequence are also conserved in structure, although much of the rest of the two structures are not conserved

information was often important (and almost always overlooked) in the prediction of protein structure, and secondly that it was possible to predict ligand binding residues from 3D models of protein structure. There were few groups that were able to capture the correct ligand binding residues in CASP7, but the prediction of ligand binding residues will form an integral part of the CASP8 experiment to be held in 2008.

3 Modelling local function conservation in sequence and structure space for predicting molecular function

3.1 Introduction

Inference of the molecular function of a given protein according to the function of proteins related in sequence and structure is a powerful annotation approach but is also prone to errors. Function is not uniformly conserved in the space of protein sequence and structure relationships. The analysis of the extent of function conservation relative to the sequence and structure similarity between proteins provides valuable information for establishing whether an annotation can be transferred reliably. In addition, combining different sequence and structure similarity measures yields a potentially better coverage in function prediction.

GODot (Weinhold et al. 2008), is a method that assesses local function conservation in protein sequence and structure space in order to predict molecular function. GODot generates a list of functional terms ranked by function conservation scores, is based on different measures of sequence and structure similarity, and is extensible to other similarity measures.

The GODot method employs a two-stage protocol. In a training stage, sequence and structural similarities of known proteins are established and regions of local function conservation in protein space are determined. In the prediction stage this pre-calculated information is employed to derive estimates of the molecular function of unknown proteins. The training stage is compute-intensive and is performed only once, the prediction stage is repeated for every query protein to obtain a function prediction.

3.2 Method

In the training stage we establish a sequence and structure space by measuring sequence and structure similarity on a representative set of protein domains provided by the ASTRAL Compendium (Chandonia et al. 2004). The protein domains in this set are compared against each other with different measures for protein similarity. Two sequence-based methods are used to compute sequence similarity: local profile alignment and global profile alignment (von Öhsen et al. 2003). In addition, two structure-based programs are used to compute structure similarity, namely Combinatorial

Extension (CE) (Shindyalov and Bourne 1998) and TM-Align (Zhang and Skolnick 2005). The neighbours of each protein in protein sequence and structure space have high similarity scores and are then identified.

To describe function, we use the GO molecular function terms (Ashburner et al. 2000) annotated to every protein domain in the dataset according to Gene Ontology Annotation. More precisely, ASTRAL domains were assigned to their respective PDB structures. The PDB structures were then mapped to UniProt (Apweiler et al. 2004) sequences using PDBSWs (Martin 2005). The UniProt sequences were annotated with GO terms using the Gene Ontology Annotation (GOA) (Camon et al. 2004). We finally removed all domains having no GO annotation or those which are part of multidomain proteins according to SCOP (Murzin et al. 1995).

The extent of local function conservation is analysed for each protein domain and molecular function GO term. Given a protein domain, we determine for each GO term how often it is annotated to 200 other domains in the neighbourhood of the domain under investigation, according to a given similarity measure. This information is modelled with a logistic curve capturing how strongly the function is conserved in the local environment. The logistic curve is calculated separately for each similarity measure.

In the prediction phase, an uncharacterised query protein is compared to all known protein domains of the training set, thus determining the protein's location and neighbourhood in protein sequence and structure space. The GO terms on the nearest neighbours are candidates for the function predicted for the query protein. Given the similarities to the nearest neighbours according to each similarity measure, the logistic curves provide estimates for the probability of a certain GO term occurring within this similarity range. These estimates provide raw function conservation scores. The raw function conservation scores are cumulated for each GO term into a combined function conservation score. The score combination also ensures compliance with the GO true path rule, such that GO terms obtain scores that are at least as high as those of their GO descendants. The combined scores provide reliability estimates for the predicted GO terms.

3.3 Application

A web-server is available for the GODOt method (<http://godot.bioinf.mpi-inf.mpg.de>), where a typical query is an uncharacterised PDB structure. The method performs sequence and structure comparisons of the query protein to each of the protein domains used in the training stage. The extent of sequence and structure similarity between the query and characterised proteins is determined and used to compute reliability estimates for the functional GO terms. The output from the web-server provides a ranked list of predicted GO terms.

Figure 3 displays a screen-shot from the web-server with the results for a query protein with known structure and unknown function according to PDB. The results combine information on the neighbourhood of the protein and the predicted GO terms.

GOdot - Results | Help

Progress Status for Query: '1vc1'

TM align (TM) ✓

Combinatorial Extension (CE) ✓

Global Profile Alignments (GP) ✓

Local Profile Alignments (PL) ✓

Predicted GO terms (based on ranking besides)
(preliminary results)

Palette associated with Gdot scores

0.0	10^{-25}	10^{-15}	10^{-7}	10^{-3}	0.15	0.4	0.7
10^{-25}	10^{-15}	10^{-7}	10^{-3}	0.15	0.4	0.7	1.0

Predicted GO terms (ranked by score)
(preliminary results)

<< Showing Rank 1 to 12 (of 12) >>

#	GO term	Description	Score
1	GO:0003824	catalytic activity	0
2	GO:0016740	transferase activity	0
3	GO:0016741	transferase activity, transferring one-carbon groups	2.9034E-42
4	GO:0008168	methyltransferase activity	3.3965E-27
5	GO:0008757	S-adenosylmethionine-dependent methyltransferase activity	1.0028E-11
6	GO:0008173	RNA methyltransferase activity	0.551301
7	GO:0008643	RNA methyltransferase activity	0.796681
8	GO:0016433	RNA (adenine) methyltransferase activity	1
9	GO:0016435	RNA (guanine) methyltransferase activity	1
10	GO:0008989	RNA (guanine-N1-)-methyl transferase activity	1
11	GO:0008988	RNA (adenine-N6-)-methyl transferase activity	1
12	GO:0000179	RNA (adenine-N6,N6-)-dimethyltransferase activity	1

Global Profile Alignments		Local Profile Alignments	
#	SCOP Domain	Score	Score
1	d1qama_ (c:66.1.24)	136.297	188.035
1	d1vfa_ (c:66.1.41)		

Individual similarity results
(preliminary results)

TM align	Score	Combinatorial Extension	Score
# SCOP Domain	0.73682	# SCOP Domain	5.9
1 d1vma_ (c:66.1.41)		1 d1p3la_ (c:66.1.33)	

Global Profile Alignments		Local Profile Alignments	
#	SCOP Domain	Score	Score
1	d1qama_ (c:66.1.24)	136.297	188.035
1	d1vfa_ (c:66.1.41)		

4 Structural templates for functional characterization

4.1 Introduction

If two proteins exhibit high sequence and structural similarity, they are most likely to perform the same molecular function. However, there are notable exceptions where the substitution of a few key functional residues can inactivate an enzymatic site or interrupt protein interactions (Whisstock and Lesk 2003). Conversely, sequence profile (e.g. PSI-BLAST (Altschul et al. 1997), Hidden Markov models (Eddy 1996)) and structure comparison methods (e.g. DALI (Holm and Sander 1993), SSAP (Taylor and Orengo 1989), CE (Shindyalov and Bourne 1998), MSDfold (Krissinel and Henrick 2004) can detect proteins that are far more distantly related by evolution (homologues) yet retain the same function. But how do we distinguish between homologues where protein function has been conserved (for example, orthologous genes in different organisms) and where a gene has duplicated and been allowed to evolve a different function (e.g. paralogues)?

More distant homologous relationships are often far more evident at the level of protein structure than in the primary sequence (Chothia and Lesk 1986). However, although a large number of protein folds are associated with a specific molecular function, a small number of “superfolds” (e.g. the TIM barrel fold) have been extensively duplicated in the genomes and are associated with a vast number of different functions. As a consequence, global structural similarity cannot be applied universally to transfer function, especially between distant homologues.

4.2 Predicting protein function using structural templates

To address this problem, many groups have attempted to define local structural motifs (or templates) associated with specific functions, in the hope that it is the conformation of residues around functional sites that provide a clear relationship between structure and function. For example, the Catalytic Site Atlas (CSA) (Porter et al. 2004) concentrates on building 3D motifs of residues that are directly involved in ligand binding or the catalytic mechanism in an enzyme.

←

Fig. 3 GODO results for query protein PH0226 from *Pyrococcus horikoshii* (PDB entry: 1ve3). The four green checkmarks in the top line indicate that all four similarity measures have finished computation successfully. The GO subgraph on the left displays the relation of the GO terms predicted for this query; the greener the colour, the more likely the prediction. The colour scheme is the same in the list on the right, displaying the predicted GO sorted by combined function conservation score. The bottom line indicates according to each similarity measure the closest identified domain along with its SCOP classification. When holding the mouse over the domains, information is displayed for which GO terms they point to, similarly holding the mouse over the GO terms on the right displays information from which domains they were inferred

In contrast to exploiting information on known functional residues, the DRESPAT method (Wangikar et al. 2003) uses graph theory to extract recurring structural patterns across superfamilies in the SCOP database (Murzin et al. 1995). DRESPAT makes no assumptions about the location or nature of the motif positions, except by excluding hydrophobic residues. A statistical model is built to assess the significance of each recurring pattern and the authors were able to identify different metal binding sites in distantly related proteins. However, as with many methods which seek small structural motifs, distinguishing between genuine similarities and background is hampered by high false positive rates.

The PINTS methods (Stark and Russell 2003) also shows promise for automatically detecting structural motifs in protein families, although is not able to annotate novel proteins with high accuracy. Again recurring side chain patterns are identified through a pair-wise comparison of diverse members within a protein family. These motifs can then be used to scan against a novel structure.

4.3 FLORA method

The principle aim of FLORA is to discover structural motifs associated with specific molecular functions (e.g. a given enzymatic reaction) and distinguish between homologous proteins that have evolved different functions. Indeed, one of the specific motivations for developing this approach was to provide a means of separating large, divergent domain families within the CATH database (Orengo et al. 1997) into functional families (FunFams). See Fig. 4 for an overview of FLORA.

The first step in the FLORA algorithm is to construct a multiple structural alignment of protein domains with similar functions. This is achieved using the CORA algorithm (Orengo et al. 1999) which applies a double-dynamic programming approach to structural alignment to perform an iterative multiple alignment. As each domain is aligned, a consensus structure is calculated from the equivalent positions determined through dynamic programming and this consensus is then aligned to the next structure. In this way, CORA is able to gradually align more distant structures by focussing on the structurally conserved positions in the functional family. Once a multiple structural alignment of all domains in the functional family has been constructed, FLORA then selects residues to include in a template.

The next challenge for the algorithm is to decide which residues in all the enzyme structures must be present for the proteins to perform a given function. Where other template methods (e.g. CSA (Porter et al. 2004), DRESPAT (Wangikar et al. 2003)) focus on residues that are likely to play a direct role in catalysis or ligand binding, FLORA instead looks for patterns of structurally conserved residues. The rationale behind this approach is that structures in the PDB are solved with a variety of cognate and non-cognate ligands, in addition to enzymes with no ligands. Given that substantial structural changes are often observed on substrate binding (e.g. in fructose 1,6-

algorithm generates templates based on their predictive power. Residues are then further selected based on patterns of sequence conservation and solvent accessibility.

Once a template has been generated by the FLORA algorithm, the next step is to use it to predict the function of novel proteins and group functionally-similar domains in each superfamily. In general, function prediction methods encode selected residues into templates by storing the expected inter-atomic distances and/or residue types. The search algorithm (e.g. using graph theory, geometric hashing etc.) then seeks to match the residues in the expected conformation, within a given tolerance value. FLORA takes an alternative approach by producing an average co-ordinate set for template residues and then applying a double-dynamic programming algorithm to find the best alignment of the template to a given the query domain. Taking this approach allows the structural similarity between the template and a domain to be calculated on a continuous scale.

In order to develop and fairly assess the performance of any method of function prediction, it is essential to create a test data set of structures where the correct answer is known. One approach is to focus on enzymes which have been functionally classified by the Enzyme Commission (E.C.) (Webb 1992). This takes the form of a 4 digit code (e.g. 2.7.7.1) that describes the catalysed reaction at various levels of specificity. Where the first digit denotes the overall enzyme class (e.g. Transferase), proteins sharing the first three E.C. codes generally perform similar reactions, albeit on different substrates. To benchmark FLORA, a data set of diverse superfamilies in CATH was created, which contained more than 3 enzyme families (i.e. domains sharing at least their first 3 E.C. numbers). The final data set comprised: 36 enzyme families from 14 different CATH superfamilies, covering all 3 major protein classes.

Comparison to standard structural alignment using the SSAP algorithm and other structure-based function predictions methods showed that FLORA was able to detect more functionally similar proteins at low error rates.

5 An integrated pipeline for functional prediction

5.1 Introduction

Structural templates are but one of many protein function prediction tools. Due to the wide variety of proteins and their functions, at this moment in time no single method is 100% successful at function prediction. A more prudent approach therefore, is to utilise as many different methods as possible in the hope that at least one method will find the correct solution. Using a variety of methods also allows a consensus opinion to be drawn, the general assumption being that the greater the number of methods pointing to the same solution, the more likely it is for that solution to be true. One problem with such a wide ranging approach is that the manual submission, retrieval and interpretation of

results from the various servers across the globe is a time consuming process and the results are often returned in a number of formats. It would therefore be of enormous benefit to researchers for a single service to access as many of these services as possible. To this end a number of protein function prediction servers have been developed such as ProKnow (Pal and Eisenberg 2005) and ConFunc (Wass and Sternberg 2008), combining multiple methods in a single site, with an aim to attain the best possible predictions. Here we discuss the ProFunc server and look at some case studies to examine the effectiveness of such an approach.

5.2 The ProFunc server

ProFunc (Laskowski et al. 2005a) (<http://www.ebi.ac.uk/thornton-srv/databases/profunc/>) was developed to respond to the demands of the various structural genomics projects in operation across the globe, in particular to address the problem of functional prediction of hypothetical proteins of unknown function. It combines a number of sequence-based and structure-based methods (Fig. 5), utilising in-house software as well as external services (often through the use of webservice) to analyse an uploaded PDB structure. The results are presented to the user in an easy to navigate set of html pages with an additional summary of the most likely functions provided as

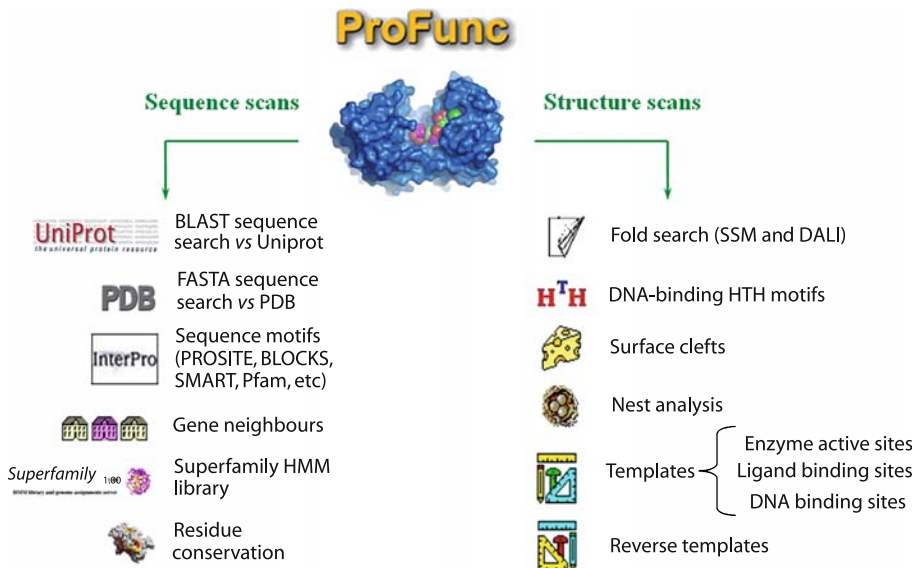


Fig. 5 ProFunc outline. Figure illustrating the types of analyses found in the ProFunc server. Sequence-based methods are found on the left hand side of the image, structure-based methods on the right

a list of gene ontology (GO) terms (Ashburner et al. 2000). It should be pointed out at this stage that this “executive summary” should only be taken as an approximate guide, since the primary aim of the server is to provide results from a number of analyses in an easily accessible format. This is to enable researchers to interpret the results themselves, bringing their own expertise to the problem, rather than a service designed to state “the function is. . .”.

5.2.1 Sequence-based searches

The first port of call when attempting to determine the function of a protein is to identify any sequence homologues with known function (see “Sequence to Function” section). Significant sequence similarity is often an indicator of similarity in function and it has been shown (Todd et al. 2001) that above 40% sequence identity, homologous proteins tend to have the same function, but conservation of function falls away rapidly below this threshold. There are always exceptions to this rule (Whisstock and Lesk 2003) and if the domain composition of a protein is retained, the function can be inferred at even lower sequence identities.

Simultaneously, two sequence searches are initiated: a FASTA (Pearson 1991) search against the Protein Data Bank (PDB) (Berman et al. 2000) and a standard BLAST (Altschul et al. 1997) search of the UniProt (Apweiler et al. 2004) sequence database. The search against the PDB is performed in order to quickly identify any obvious matches to proteins of known structure which could give structural insights into any putative function. The BLAST search against UniProt aims to identify closely related sequences, which are subsequently aligned using a simple pile-up procedure. This is used to generate a multiple sequence alignment from which residue conservation scores can be calculated using the method of Valdar and Thornton (Valdar and Thornton 2001). These scores are essential for many of the subsequent structural analyses and are mapped onto the structure for visualisation using Rasmol (Sayle and Milner-White 1995). This allows the user to identify patches of highly conserved residues on the protein surface which are often a strong indicator of important functional sites.

For every UniProt sequence match found by BLAST, the protein’s location on the source organism’s genome is identified and the 10 genes on either side are extracted, tabulated and illustrated in diagrammatic form. In many organisms neighbouring genes are often functionally related (e.g. bacterial operons) so the identification of any proteins of known function in close proximity can provide a clue to the function.

Finally, the query protein sequence is analysed for matches to any of the myriad of motifs, domains, patterns, fingerprints and Hidden Markov Models (HMMs) in InterPro (Mulder et al. 2007) using the InterProScan (Quevillon et al. 2005) webservice. InterPro contains a massive number of sequence patterns from the Pfam (Sonnhammer et al. 1998), PROSITE (Sigrist et al. 2002), SMART (Schultz et al. 1998), PRINTS (Attwood 2002), BLOCKS (Henikoff et al. 1999), TIGRFAMS (Haft et al. 2001),

ProDom (Servant et al. 2002) and Gene3D (Yeats et al. 2006) resources. Many of these patterns are specific to functional families and are excellent indicators of putative function. In addition to this, a separate scan is performed against the Superfamily (Gough and Chothia 2002) library of SCOP (Murzin et al. 1995) structural superfamily HMMs to identify possible matches to the PDB and individual domains.

5.2.2 Structure-based searches

When the sequence-based methods fail or provide few clues, investigation of the protein's three-dimensional structure can identify similarities and distant homologies that may help elude the biochemical function. These range from the large-scale global fold comparisons, down to highly specific template based approaches that pinpoint constellations of individual residues.

The first step in assigning function from structure is to compare the global fold of the protein with those already known. Proteins with a similar structure are usually evolutionarily related and therefore likely to share a similar function, but once again caution must be taken when interpreting these results as structural similarities might be the result of convergent evolution and therefore the function may be quite different. There are a number of different fold comparison methods available (for a review see Novotney et al. 2004) but within ProFunc two searches are made. The first uses a secondary structure matching algorithm developed at the EBI called MSDfold (Krissinel and Henrick 2004) (<http://www.ebi.ac.uk/msd-srv/ssm/>) and this is supplemented by a standard DALI (Holm and Sander 1993) search. Any significant matches are ranked and listed with the top hits illustrated in a figure of secondary structures aligned. A structural superposition of any number of the top hits can be viewed to allow the user to assess the level of similarity or to identify core regions of importance (e.g. when a particular structural domain is conserved across all homologues). In some cases a particular combination of secondary structure elements is strongly associated with a particular function. The most obvious example of this is the Helix-Turn-Helix (HTH) motif found in a number of DNA binding proteins (Aravind et al. 2005). As these are highly significant functional indicators, any structure submitted to ProFunc is scanned against a database of known HTH templates extracted from PDB structures known to bind DNA (Ferrer-Costa et al. 2005). To reduce the number of false positive matches the hits are filtered and scored by solvent accessibility and electrostatic potential. High ranking significant matches can be viewed on the structure using Rasmol.

Where the global fold or sub-domains are of limited use, the next stage is to investigate any pockets or clefts on the protein surface. It has been shown that a protein-ligand binding site (or "active site") is more often found to be the largest cleft in the protein and this cleft is significantly larger than any other one in the protein (Laskowski et al. 1996). In ProFunc the SURFNET (Laskowski 1995) algorithm is used to identify all the clefts in the protein structure ranked by size. The pockets can then be viewed using

Rasmol scripts and these can be coloured by residue type, nearest atom type, or by residue conservation. The most highly conserved pockets are those most likely to have some functional relevance and may even be ligand binding or cofactor binding sites. The identification of putative ligands or co-factors can help narrow down any putative functions and allow for experimental testing through ligand-binding assays. As always, care should be taken when attempting to infer function from binding sites as similarity in clefts is not always a strong indicator of ligand similarity, as seen in an analysis of human cytosolic sulfotransferases (Allali-Hassani et al. 2007). Another type of pocket searched for using ProFunc is much smaller than a cleft and is known as a “Nest” (Watson and Milner-White 2002). These are formed when the dihedral angles of the protein backbone alternate between right and left handed forms as defined in the Ramachandron plot. The consequence of this conformation is the creation of small slightly positive concavities which are able to bind anionic groups such as phosphates and sulphates. On their own they may not provide too many functional clues however, when multiple Nests are found in series they form larger functional motifs such as the ATP-binding P-loop or Iron-sulphur binding sites.

The final set of structure-based analyses involves the highly specific *n*-residue templates. As the three dimensional arrangement of enzyme active site residues is often more conserved than the overall fold, this can be used to identify functionally important local similarities in proteins with very different folds. As discussed in the previous subsection, there are a number of approaches available to identify geometric patterns of residues or atoms in protein structures. The ProFunc server runs four template-based scans using JESS (Barker and Thornton 2003) which uses a KD-tree data structure for rapid matching of templates and queries. Three of the searches scan the submitted protein structure for the presence of structural templates held in databases of known enzyme active sites, ligand binding sites and DNA-binding sites. The enzyme active sites are based on manually curated templates that were the foundation of the Catalytic Site Atlas (Porter et al. 2004), whereas the ligand-binding and DNA-binding templates are automatically generated from the PDB and updated on a weekly basis. The final run involves a “reverse template” search (Laskowski et al. 2005b) which turns the template idea on its head, generating templates from the submitted structure and scanning each generated template in turn against a representative sample of the PDB.

In each case, the probability of matching the orientation of any three residues by chance alone is very high; therefore identifying significant matches by root mean square deviation (rmsd) alone is not sufficient. To narrow down the hits to those most significant a 10 Å sphere is drawn around the centroid of the template and the match. Within this sphere any overlapping identical residues, chemically similar residues and empty space is identified and used to score the significance of the hit. The underlying idea here is that if a site is performing the same function, then the surrounding local environment is likely to be similar. Therefore the more similar the surrounding residues, the higher the significance of the template match.

5.3 Case studies

The ProFunc server was developed to respond to the problem of an increasing number of structures of unknown function being deposited by the various Structural Genomics initiatives across the globe. There are a number of goals for these high-throughput projects but one of the main aims was to determine as many structures as possible with novel folds in an attempt to cover the “fold space” of protein structure, whilst increasing automation to reduce the cost of structure determination. In collaboration with the NIH-funded Midwest Center for Structural Genomics (MCSG), a fully automated procedure has been set up to automatically submit each new structure to ProFunc as part of their high-throughput protocol. The results are then manually assessed and reports sent back to the depositors indicating any interesting findings. This process has provided a very useful dataset to test the ProFunc server and determine the most successful methods.

A large scale analysis (Watson et al. 2007) conducted at the end of the first stage of the NIH-funded Protein Structure Initiative (PSI), looked at the structure-based function predictions for MCSG proteins using the ProFunc server. The aim of the study was to determine, using structures subsequently functionally characterised, whether the ProFunc server could have identified this function at the time of release and which methods within the server show the greatest level of success. The results showed that, of the 92 proteins with known function taken from the 282 non-redundant deposits at that point, 70% would have had their functions assigned had the server been fully operational at the time of deposition. Of this 70% over three quarters had the correct function predicted by more than one method. The two most successful structure based methods were subsequently identified as the “reverse template” and the fold comparison (MSDfold) approaches, both of which showed between 50% and 60% success rates. In the majority of cases the two methods are complementary, finding the same protein for their top matches, but there are also examples where one method finds a correct match that the other does not. In general this relates to the differences in the focus of the methods, with the fold matches identifying common global features whereas the reverse templates identify more local similarities.

The usefulness of the ProFunc approach and the difficulties faced in function prediction are illustrated by two individual case studies below. The first case illustrates how the server can easily be used to identify function, whereas the second shows that some structures do not provide easy solutions.

5.3.1 Case study 1: published function identified

Tm0936 (PDB deposit 2plm) is annotated in the PDB as an uncharacterised protein from *Thermotoga maritima* and contains a Pfam amidohydrolase domain, yet the associated reference for this PDB entry indicates that the actual function

of the protein has been identified (Hermann et al. 2007). In this paper the authors describe a method involving the targeted docking of high-energy metabolic intermediates into an earlier structure of Tm0936 solved by the New York Structural Genomix Consortium (PDB entry 1p1m). Their results suggested that Tm0936 is an adenosine deaminase (E.C.3.5.4.4) and led them to determine the structure of Tm0936 with *S*-inosylhomocysteine bound (a product of *S*-adenosylhomocysteine deamination).

The structure of Tm0936 was submitted to the ProFunc server for analysis. All of the sequence-based methods indicated it to be a putative amidohydrolase, as would be expected for any member of the Pfam amidohydrolase domain, but could not provide specific details on the substrate. Moving to the structure-based methods, a number of very strong fold matches also confirmed a general amidohydrolase function but still no specificity. Only on examination of the template-based approaches was a more detailed function identified. A highly significant match was found to an enzyme active site template for adenine amidohydrolase (derived from a mouse enzyme, PDB entry 1a4l). Further examination of the match shows that the sequence similarity between Tm0936 and the template enzyme is only 17.5% thereby making it difficult to identify with sequence-based searches. The overall structural similarity is however 95% and the local sequence similarity (i.e. the residues surrounding the template match) is almost 30%. The importance of this local similarity is seen in the superposition of the two active sites (Fig. 6), both of which contain bound cofactors in almost identical conformations. This suggests that had the structure been submitted to the ProFunc server, the specific adenosine deaminase function would have been identified. It is also interesting to note

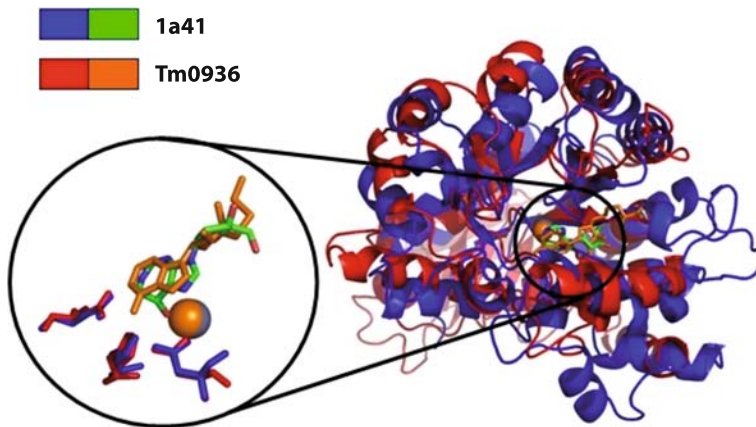


Fig. 6 Adenosine deaminase enzyme active site template superposition on Tm0936. Overall folds of Tm0936 (blue cartoon, green cofactor) and 1a4l – adenosine deaminase (red cartoon, orange cofactor) show remarkable similarity for only 17.5% sequence identity. The magnified region shows the almost identical active site residue conformations and the superposed cofactors in both structures

that a ProFunc run of the protein used in the docking experiment also has a significant match to the adenine amidohydrolase template, and that the bound methionine molecule is located exactly where the methionine-like substructure of S-adenosylhomocysteine would be expected.

5.3.2 Case study 2: function unclear

Yjcs is a hypothetical protein from *Bacillus subtilis* solved as part of the MCSG and is deposited as PDB entry 1q8b. Sequence analysis indicates similarity to a number of proteins of unknown function and a few putative antibiotic biosynthesis monooxygenases. There are no Pfam domains or functional motifs present but it is a member of the CATH (Orengo et al. 1997) homologous superfamily 3.30.70.900 (annotated as oxidoreductases). There are only a few conserved residues in the core of the protein and no genome locations can be found for any homologues. Looking at the structure does not provide any more detail, with a number of fold and reverse-template matches to various proteins including putative monooxygenases, sugar-binding proteins as well as other hypothetical proteins of unknown function. In this case the most likely function is some kind of putative monooxygenase but little more can be said about possible substrates or pathways it could be involved in. Additional analyses with new computational techniques and further experimental study will be required to provide detailed functional predictions in this case.

5.4 Conclusion

There are a wide variety of methods utilising protein sequence and structure with the aim to predict protein function. No one method provides 100% success, so only by combining as many methods as possible can one maximise the probability of identifying the correct function. There are several meta-servers currently available which integrate a wide variety of resources, such as the ProFunc server. We have seen that, although the ProFunc server has shown a great deal of success, cases abound where every method currently available provides no hits or functionally uninformative hits to hypothetical proteins of unknown function. It is therefore vital that new methods continue to be developed and are integrated into pipelines like the ProFunc server, providing researchers with “one-stop-shop” sites to submit their data for analysis. Ultimately though, the predictions are meaningless without any experimental validation, and therefore further improvements to high-throughput functional assays (such as those described in Kuznetsova et al. 2005) will greatly assist biochemical validation of predictions in the future.

Acknowledgments

This work was performed with funding from the National Institutes of Health, grant number GM62414 and the US DoE under contract W-31-109-Eng-38.

References

- Allali-Hassani A, Pan PW, Dombrowski L, et al. (2007) Structural and Chemical Profiling of the Human Cytosolic Sulfotransferases. *PLoS Biol* 5: e97
- Altschul SF, Madden TL, Schaffer AA, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res* 25: 3389–3402
- Apweiler R, Bairoch A, Wu CH, et al. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 32: D115–D119
- Aravind L, Anantharaman V, Balaji S, et al. (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* 29: 231–262
- Ashburner M, Ball CA, Blake JA, et al. (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25: 25–29
- Attwood TK (2002) The PRINTS database: a resource for identification of protein families. *Brief Bioinform* 3: 252–263
- Barker JA, Thornton JM (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* 19: 1644–1649
- Berman HM, Westbrook J, Feng Z, et al. (2000) The protein data bank. *Nucleic Acids Res* 28: 235–242
- Blundell TL, Mizuguchi K (2000) Structural genomics: an overview. *Prog Biophys Mol Biol* 73: 289–295
- Camon E, Magrane M, Barrell D, et al. (2004) The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res* 32: D262–D266
- Chandonia J, Hon G, Walker NS, et al. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res* 32: D189–D192
- Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5: 823–826
- Devos D, Valencia A (2000) Practical limits of function prediction. *Proteins* 41: 98–107
- Eddy SR (1996) Hidden Markov models. *Curr Opin Struct Biol* 6: 361–365
- Ferrer-Costa C, Shanahan HP, Jones S, et al. (2005) HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. *Bioinformatics* 21: 3679–3680
- Gough J, Chothia C (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* 30: 268–272
- Haft DH, Loftus BJ, Richardson DL et al. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res* 29: 41–43
- Henikoff S, Henikoff JG, Pietrokovski S (1999) Blocks +: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15: 471–479
- Hermann JC, Marti-Arbona R, Fedorov AA, et al. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature* 448: 775–779
- Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol* 233: 123–138
- Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D60*: 2256–2268
- Kuznetsova E, Proudfoot M, Sanders SA, et al. (2005) Enzyme genomics: Application of general enzymatic screens to discover new enzymes. *FEMS Microbiol Rev* 29: 263–279
- Lopez G, Rojas, AM, Tress ML, et al. (2007a) Assessment of predictions submitted for the CASP7 function prediction category. *Proteins* 69(S8): 165–174
- Lopez G, Valencia A, Tress ML (2007b) FireDB – a database of functionally important residues from proteins of known structure. *Nucleic Acids Res* 35: D219–D223
- Lopez G, Valencia A, Tress ML (2007c) firestar – prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res* 35: W573–W575

- Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 13: 323–328
- Laskowski RA, Luscombe N, Swindells M, et al. (1996) Protein clefts in molecular recognition and function. *Protein Sci* 5: 2438–2452
- Laskowski RA, Watson JD, Thornton JM (2005a) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* 33: W89–W93
- Laskowski RA, Watson JD, Thornton JM (2005b) Protein function prediction using local 3D templates. *J Mol Biol* 351: 614–626
- Martin ACR (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics* 21: 4297–4301
- Mulder NJ, Apweiler R, Attwood TK, et al. (2007) New developments in the InterPro database. *Nucleic Acids Res* 35: D224–D228
- Murzin AG, Brenner SE, Hubbard T, et al. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247: 536–540
- Novotny M, Madsen D, Kleywegt GJ (2004) Evaluation of protein fold comparison servers. *Proteins* 54: 260–270
- Orengo CA (1999) CORA—topological fingerprints for protein structural families. *Protein Sci* 8: 699–715
- Orengo CA, Michie AD, Jones S, et al. (1997) CATH – a hierarchic classification of protein domain structures. *Structure* 5: 1093–1108
- Pal D, Eisenberg D (2005) Inference of protein function from protein structure. *Structure* 13: 121–130
- Pearson WR (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 11: 635–650
- Polacco BJ, Babbitt PC (2006) Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*. 22: 723–730
- Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32: D129–D133
- Quevillon E, Silventoinen V, Pillai S, et al. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res* 33: W116–W120
- Sayle RA, Milner-White EJ (1995) RASMOL: biomolecular graphics for all. *Trends Biochem Sci* 20: 374–376
- Schultz J, Milpetz F, Bork P, et al. (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc Natl Acad Sci USA* 95: 5857–5864
- Servant F, Bru C, Carrere S, et al. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform* 3: 246–251
- Shindyalov I, Bourne P (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11: 739–747
- Sigrist CJA, Cerutti L, Hulo N, et al. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 3: 265–274
- Sonnhammer EL, Eddy SR, Birney E, et al. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 26: 320–322
- Stark A, Russell RB (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res* 31: 3341–3344
- Taylor WR, Orengo CA (1989) Protein structure alignment. *J Mol Biol* 208: 1–22
- Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* 307: 1113–1143
- Torrance JW, Bartlett GJ, Porter CT, et al. (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J Mol Biol* 347: 565–581
- Tress ML, Jones DT, Valencia A (2003) Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol* 330: 705–718

- Tress ML, Graña O, Valencia A (2004) SQUARE-determining reliable regions in sequence alignments. *Bioinformatics* 20: 974–975
- Waldar WSJ, Thornton JM (2001) Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol* 313: 399–416
- von Öhsen N, Sommer I, Zimmer R (2003) Profile-profile alignment: a powerful tool for protein structure prediction. *Pac Symp Biocomput* 8: 252–263
- Wangikar PP, Tendulkar AV, Ramya S, et al. (2003) Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol* 326: 955–978
- Wass MN, Sternberg MJ (2008) ConFunc-functional annotation in the twilight zone. *Bioinformatics* 24: 798–806
- Watson JD, Milner-White EJ (2002) A novel main-chain anion-binding site in proteins: the nest. A particular combination of phi, psi values in successive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions. *J Mol Biol* 315: 171–182
- Watson JD, Sanderson S, Ezersky A, et al. (2007) Towards fully automated structure-based function prediction in structural genomics: a case study. *J Mol Biol* 367: 1511–1522
- Webb EC (1992) *Enzyme nomenclature*. Academic Press, San Diego
- Weinhold N, Sander O, Domingues FS, Lengauer T, Sommer I (2008) Local function conservation in sequence and structure space. *PLoS Comput Biol* 4: e1000105
- Whisstock JC, Lesk AM (2003). Prediction of protein function from protein sequence and structure. *Q Rev Biophys* 36: 307–340
- Wilks HM, Hart KW, Feeney R, et al. (1988) A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. *Science* 242: 1541–1544
- Yeats C, Maibaum M, Marsden R, et al. (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res* 34: D281–D284
- Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acid Res* 31: 3370–3374
- Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33: 2302–2309

CHAPTER 4.6

Harvesting the information from a family of proteins

B. Vroling and G. Vriend

CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

1 Introduction

The need for information about the functional roles of the elements comprising the human genome became larger than ever with the completion of its sequencing in 2003 (Lander et al. 2001; Venter et al. 2001; Consortium 2004b). BioSapiens (Excellence 2005) is contributing to the ENCODE (Consortium 2004a) program, which is providing a biologically informative representation of the human genome using high-throughput methods to identify and catalogue all functional elements. The ENCODE pilot project consisted of annotating 1% of the genome (Consortium et al. 2007). The presently ongoing functional annotation of the other 99% will be a crucial next step for many, diverse fields of science.

Previous chapters have illustrated many good databases and good tools for the prediction of function from sequence. The best, and most often used, tool for functional annotation is a direct sequence comparison between the homo sapiens sequence with unknown function and homologous sequences in other species for which the function is known. Most often this is done with BLAST or PSI-BLAST, but threading methods and profile-profile comparison methods are also often used. This transfer of information from one protein to the other does not stop with the protein's function. Types of information that can be transferred when a sequence alignment with a well-studied protein can be made include; dimer interface residues, active site residues, post-translational modification sites, metal-binding sites, etc. The latter is important for research areas like drug design or quantitative systems biology for which it is not only important what proteins do, but also how they do it. Answering that question requires that knowledge is obtained about the roles of individual amino acids in the various aspects of protein function.

Corresponding author: G. Vriend, CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Geert Grooteplein 26-28, 6525 GA Nijmegen, The Netherlands (e-mail: vriend@cmbi.ru.nl)

The use of multiple sequence alignments is not limited to just the transfer of experimental information. The footprints that evolution left behind in these sequences can be reconstructed from conservation and variability patterns, and a number of computational techniques exists that can harvest this information. Some of those will be discussed in this chapter.

The final goal of most bioinformatics studies is answering biological questions that can widely vary in detail and complexity. Often it is needed to use all available information (3D structures, multiple sequence alignments, expression and distribution patterns, interaction information, etc.) to answer such questions. The G protein-coupled receptor (GPCR) family will be used to illustrate this because the GPCRDB (Horn et al. 1998b, 2001, 2003) provides a molecular class specific information system (MCSIS) that holds much, heterogeneous data in a well organized and easily accessible form. With their 16,000 sequences, the GPCRs are one of the largest sequence families known to date. The additional availability of 10,000 well annotated mutations (Beukers et al. 1999; Edvardsen et al. 2002), more than 12,000 binding constants for ligands (Lutje Hulsik 2002), a large number of well described disease phenotypes (Hamosh et al. 2005), genome locations and organizations, and an up-to-date annotated SNP database (Kazius et al. 2007) make the GPCR family a great subject for inferring new information using a wide spectrum of bioinformatics techniques.

1.1 Information transfer

Chothia and Lesk (1986) studied the relation between sequence similarity and structure similarity, and they concluded that structures stay conserved in evolution much longer than sequences. Sander and Schneider (Sander and Schneider 1991) later quantified this relation (see Fig. 1).

A basic assumption in bioinformatics is that residues at equivalent positions in homologous proteins have similar functions. That implies that the function of a residue is determined by its location, while the residue type determines how that function is performed. A residue known to be in contact with the endogenous agonist in the mouse muscarinic type-3 receptor, for example, is most likely also involved in contacting the endogenous ligand in the human muscarinic type-1 receptor. Figure 5 shows an important aspartic acid that is conserved in all amine binding GPCRs in which it is always the counter ion for the positive amine group in the ligand. In the hormone receptor family, on the other hand, we always find a serine or threonine at this position that interacts with the ligand, while in the chemokine receptor family we find nearly invariably a tyrosine that interacts with the ligand at this same position.

Carrying over information is the most elementary use of homology, and many chapters in this book use this concept one way or another. In the remainder of this chapter we will concentrate on a different use of this principle. Namely, what can we learn from the patterns of variability and conservation that evolution has left in a

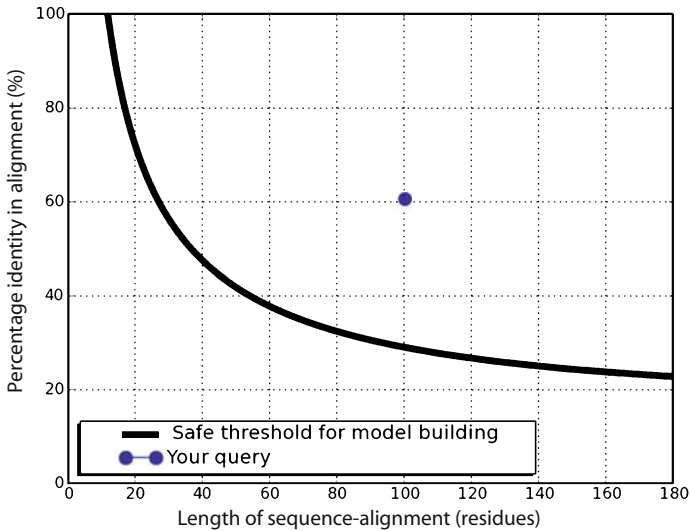


Fig. 1 The Sander and Schneider plot. If the percentage sequence identity at a given alignment length is above the curve, it is safe to transfer information and, for example, to build a 3D model by homology. In this example a user requested information about a protein that can be aligned with 60% sequence identity (over 100 amino acids) to a well-studied protein with known structure. The blue dot is above the ‘safe threshold curve’ so modelling is possible, and thus, all kinds of information transfer regarding the roles of individual amino acid is also possible.

multiple sequence alignment. For example, if we hadn’t known yet that the aspartic acid in the amine receptors, the serine/threonine in the hormone receptors, and the tyrosines in the chemokine receptors all were involved in ligand binding, could we then have derived that knowledge from the multiple sequence alignment?

But first we will discuss a system to generate multiple sequence alignments in a wider context, the so-called molecular class specific information system. Such a system is not strictly needed to produce and analyze multiple sequence alignments, but as you will read in many chapters in this book, it often is very beneficial to have all available information easily available.

2 Molecular class-specific information systems

Studies that involve carrying over information from one protein to the other seem simple at a first glance. However, the amount of data that needs to be collected from heterogeneous sources, converted to syntactic and semantic homogeneity, validated, stored, and indexed, is enormous. And the data sources from which relevant data can be collected grow continuously, both in volume and in number. The enormous amount of data that is entered each day into databases and the literature is outstripping the ability of experi-

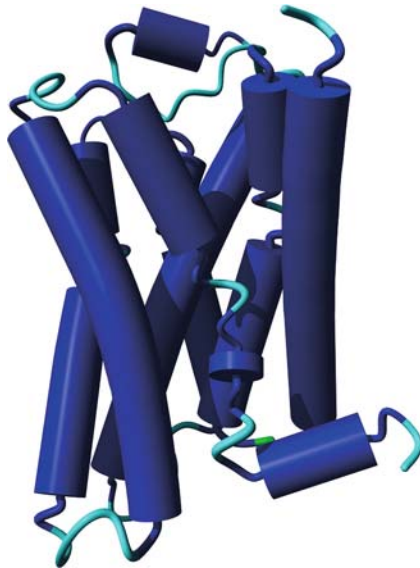


Fig. 2 Cartoon representation of the recently resolved human β_2 -adrenoceptor structure.

mental scientists to keep pace. Large monolithic databases like SwissProt, EMBL, PDB, etc., are invaluable for biomedical scientists. Most scientists, though, tend to use many databases while concentrating on one molecule, or one family of molecules. The main aim of molecular class-specific information systems is to gather heterogeneous data from across a variety of electronic sources in order to draw new inferences about the target protein families. The number of experimental data types (primary data), is limited, but there are hardly any limits to the number of data types that can be derived computationally (secondary data). It is therefore important to consider the questions that the system should help answering when adding more and more computational data.

2.1 G-protein-coupled receptors

Over 1000 human genes encode G protein-coupled receptors. The ligands that bind and activate these receptors are heterogeneous and include photons, odours, pheromones, hormones, ions, neurotransmitters, and proteases. GPCRs transmit signals from outside the cell to amplification cascades controlling sight, taste, smell, slow neurotransmission, cell division, etc. All GPCRs form a seven transmembrane (TM) α -helical bundle, connected by three intracellular and three extra-cellular loops. Within some GPCR families the overall sequence identity between family members can be lower than 25%.

GPCRs are a major target for the pharmaceutical industry as is reflected by the fact that nearly 50% of all known drugs act on a GPCR (Howard et al. 2001). Some of the

major questions relevant to GPCR pharmacology include the following: What residues are critical for ligand binding and for the activation of G-proteins or other proteins? What do different receptor families have in common with regard to their activation mechanism? And which residues are responsible for the differences and should thus especially, or especially not be influenced by potential drug molecules? The GPCRDB is designed to be a data storage medium, as well as a tool to aid biomedical scientists with answer questions like these by offering a single point of access to many types of data that are integrated and visualized in a user-friendly way.

The primary data in the GPCRDB are sequences, structures, and mutation and ligand binding data. Sequences are automatically imported from the SWISS-PROT and UniProt databases (Bairoch and Apweiler 2000). cDNA sequences are imported from the EMBL (Kanz et al. 2005) databank. Structure data is retrieved from the PDB (Berman et al. 2000). Mutation data is obtained from the manually curated tinyGRAP (Beukers et al. 1999; Edvardsen et al. 2002) database, as well as mutation data extracted from online literature using an automated procedure (Horn et al. 2004). SNP data comes from the NAVA system (Kazius et al. 2007).

The data organization in the GPCRDB is based on the pharmacological classification of GPCRs and the main way to access the data is *via* a hierarchical list of known families in agreement with this classification. For a specific family, users can access individual sequences, multiple sequence alignments, the profiles used to perform the latter, snake-like diagrams and phylogenetic trees. Two-dimensional snake-like diagrams are used to represent and combine GPCR sequence, 2D structure and mutation information (Campagne et al. 1999). Furthermore, entries can be retrieved using a query system, and data can be saved either using the WWW-pages or *via* ftp access. One important dissemination and inference facility in the GPCRDB is a large series of Multiple Sequence Alignment (MSA) analysis tools.

3 Extracting information from sequences

During evolution certain residues can mutate almost without restrictions while other residues are so important for the function of the protein that they have never mutated. There are also residue positions that tend to be conserved in subfamilies but different between subfamilies. Different parts of the molecule are under different types of evolutionary pressure, leading to different patterns of conservation and variation that can be analyzed to learn more about the role of the individual residue positions.

Residue conservation has been evaluated in multiple sequence alignments by means of variability (number of different amino acids found), Shannon entropy, and variance-based and score-matrix indices (Mirny and Shakhnovich 2001; Pei and Grishin 2001; Shenkin et al. 1991). The patterns of conservation in proteins have been described as the

fingerprints left by evolution in the structure (Zuckermandl and Pauling 1965), and they have been used for quality assessment and refinement of multiple-sequence alignments (Pei and Grishin 2001). Conserved residues are often clustered in certain regions of protein structures, sometimes at universally conserved positions (Mirny and Shakhnovich 1999), so called because they can form a motif characteristic of the fold. Sometimes, these positions are also found in the corresponding sequence segments of analogs, and their location often coincides with that of supersites (Russell et al. 1998). Many groups have used the identification of conservation patterns in proteins as a method to search for function. Some of these methods are based on energy calculations on proteins of known structure looking for charge and shape complementarities in protein and ligand surfaces that are thought to interact (Kuntz et al. 1982; Desjarlais et al. 1988; Miranker and Karplus 1991; Honig and Nicholls 1995; Lamb and Jorgensen 1997; Wang et al. 2001; Glaser et al. 2006). Kufareva et al. (2007) predicted binding interfaces from single proteins with a known 3D structure. Other groups have predicted functional motifs using principal component analysis (Casari et al. 1995), analysis of physicochemical descriptors to score protein-protein interactions (Jones and Thornton 1997; Fernandez-Recio et al. 2005), search for motifs in Blocks databases (Petrokovski et al. 1996), or alignment of hinge regions (Shatsky et al. 2002). Some methods combine evolutionary information extracted from multiple sequence alignments with three dimensional structure information (Lichtarge et al. 1996a; Aloy et al. 2001; Armon et al. 2001; Glaser et al. 2003).

3.1 Correlated mutation analysis

It seems obvious that a residue conserved in a sequence family must be involved in a function common to the family, while a residue conserved only in subfamilies is likely to have a functional role in only those subfamilies. This concept can work in two directions: deductive and inductive. In both directions correlations of conservation and variability patterns in a MSA are being analyzed. Most studies where these patterns are correlated with known facts tend to be deductive, while most studies that correlate these patterns against each other tend to be inductive, but scientific creativity can easily blur this division.

The term correlated mutation refers to the tendency of pairs of residue positions to either mutate in tandem or stay conserved together, and correlated mutation analysis (CMA) is a technique for the identification of these patterns in a MSA. It has been originally described (Göbel et al. 1994; Singer et al. 2002) to predict physical contacts within proteins that could be used for structure prediction (see Fig. 3), but that is beyond the scope of this chapter. CMA was proven to be a powerful technique for predicting amino acid contacts at protein-protein interfaces, since correlated mutations tend to accumulate at the protein surface (Oliveira et al. 1993; Pazos et al. 1997; Horn et al. 1998a). Kuipers et al. (1997a, b) and Oliveira et al. (2003b) determined correlations

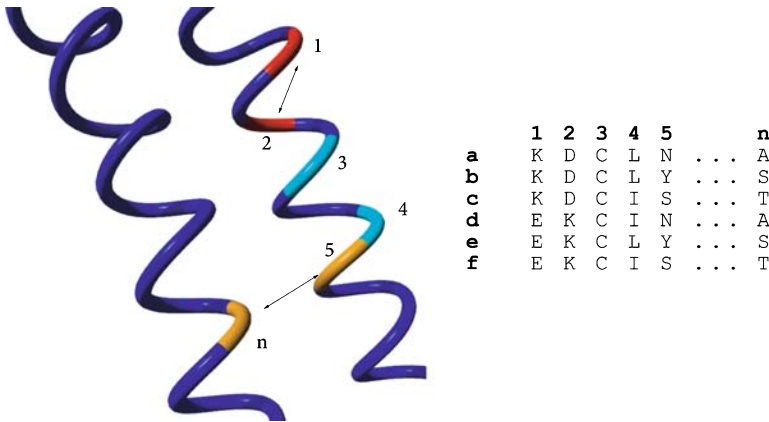


Fig. 3 Illustration of correlated mutation analysis. Several residues are shown in their structure context, in this example, two nearby α -helices. For these residues, six sequences (a–f) are shown as a multiple alignment. Positions 1 and 2 show correlated substitutions (connected by arrows), as do positions 5 and n.

between residue positions and ligand binding characteristics of receptors to determine which residues were involved in that ligand binding.

4 Correlation studies on GPCRs

Kuipers et al. (1997a, b) and Oliveira et al. (2003b) determined the correlation between the absence and presence of residues and the binding or not binding of pindolol to GPCRs from the Class-A amine sub-family. In this example of deductive use of CMA they observed one asparagine in helix VII that correlated perfectly. This asparagine was mutated in the serotonin 1a receptor to a valine which resulted in

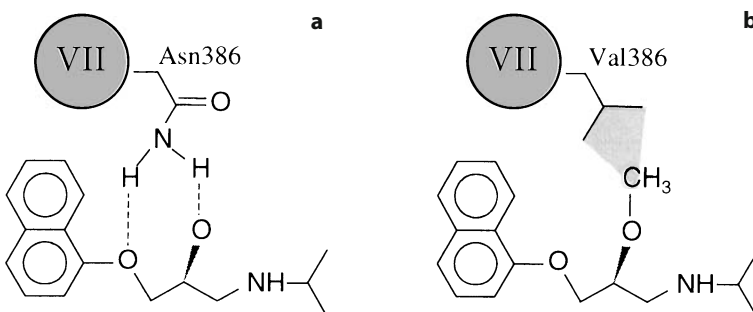


Fig. 4 (a) The original binding mode of the ligand. (b) The mutation Asn386Val resulted in a loss of binding affinity, but when a methyl group was added to the ligand, binding affinity was restored.

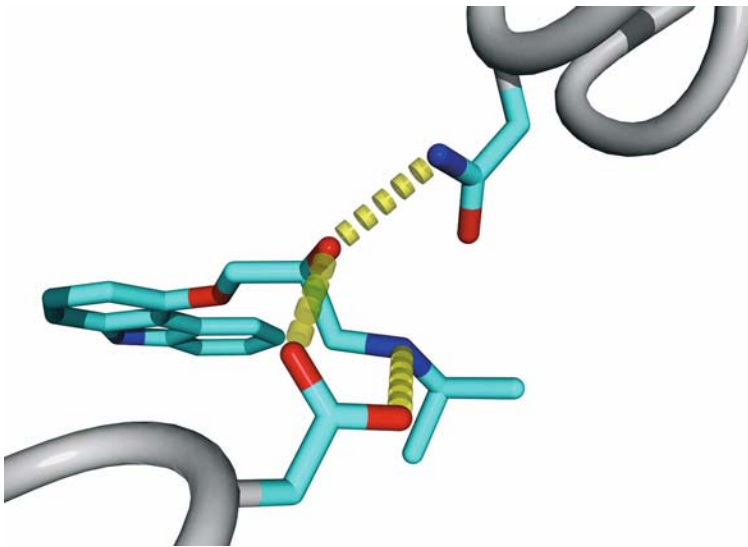


Fig. 5 The binding mode of the pindolol analogue in the β_2 -adrenergic receptor. The asparagine indicated by Kuipers et al. as responsible for pindolol selectivity is shown, as well as an aspartate that is absolutely conserved and essential for ligand binding

abolishing the binding. When an extra methyl group was introduced in the pindolol molecule, at the position where the interaction with the asparagine was predicted to take place, the binding was regained. This experiment is illustrated in Fig. 4.

The recently solved β_2 -adrenoceptor structure contains an inverse agonist that is highly similar to pindolol, and that has exactly the same active group between the ring system and the nitrogen as pindolol. The β_2 -adrenoceptor sequence is very similar to the sequence of the serotonin 1a receptor, and the β_2 -adrenoceptor also binds pindolol well, and it has an asparagine at the same position as the one mutated in the serotonin 1a receptor. Figure 5 shows the location of the inverse agonist and this asparagine in the β_2 -adrenoceptor structure. This prediction was made more than 13 years before its first experimental confirmation, and it was based on a deductive correlation study of heterogeneous data: an MSA and ligand binding studies. This example beautifully illustrates the deductive power of the CMA method.

Oliveira et al. (1993) used CMA in an inductive way to predict residue relations in GPRCs. Their reasoning was simple: if the mutation patterns of residue positions are strongly correlated, i.e. give a strong signal in a CMA analysis, then those residue positions must be involved in a common function. Other information like the position in the structure, ligand binding information, or mutation studies can then reveal which function was detected. When it was observed that correlated mutations accumulate at protein surfaces, the technique was used as a new approach to

protein–protein docking and the prediction of protein–protein interfaces (Pazos et al. 1997). The rationale behind this is that mutations at one of the interfaces must be compensated by a mutation in its counterpart. Gouldson et al. (2001) used such intermolecular CMA studies to predict that many GPCRs form dimers. This idea was also used to identify homo- and hetero-dimerisation interfaces for GPCRs (Gouldson et al. 1998, 2000, 2001; Filizola and Weinstein 2002; Filizola et al. 2002; Valencia and Pazos 2002; Hernanz-Falcón et al. 2004) as well as for G-protein – GPCR interfaces (Oliveira et al. 1999, 2002, 2003b; Horn et al. 2000; Gouldson et al. 2001; Möller et al. 2001).

4.1 Evolutionary trace method

The evolutionary trace method (ET) by Lichtarge et al. (1996a, b, 1997) is a special case of CMA. It can be used deductively and inductively. It predicts functionally important residues in a protein family given a three dimensional protein structure and a MSA. The starting point of the ET method is a MSA, which must contain sequences of a protein family with divergently related members. The tree is then partitioned into sub-groups corresponding to functional classes. A consensus sequence is then generated for each sub-group, and these sequences are compared. Residue positions that are conserved within sub-groups, but vary among them are called class-specific- or trace-residues. The rank of a specific residue is determined as the number of tree divisions needed for a residue to become a trace-residue. The trace residues are mapped on a 3D structure, and functional sites are indicated by the clustering of these trace-residues. The ET method is illustrated in Fig. 6. Despite rather different approaches, CMA and ET are highly similar in what they can detect from an MSA. The ET method is restricted by the need to use a phylogenetic tree, and consequently cannot detect correlations that do not perfectly follow the tree branching pattern.

Evolutionary trace analyses have been successfully applied in a number of studies (Lichtarge et al. 1996a, b, 1997; Dean et al. 2001; Madabushi et al. 2004). Dean et al. (2001), for example, used the ET method to confirm that GPCRs might actually be domain swapped dimers. Sowa et al. (2001) performed an ET analysis of 42 members of the RGS (regulators of G-protein signalling) family that revealed a novel functional surface, located next to the interface between the RGS and $G\alpha$. It was predicted that the G-protein effector subunit PDE γ would bind the RGS- $G\alpha$ complex by straddling both $G\alpha$ and the newly discovered functional site. Indeed, mutagenesis of RGS based on the ET prediction revealed that three residues out of the six selected for mutagenesis had profound effects on the regulation of activity by PDE γ . After the ET-based mutagenesis was completed, the crystal structure of RGS9Gi/ α PDE γ was solved by Slep et al. (2001) and confirmed the predicted position of the PDE γ interaction site on the RGS domain.

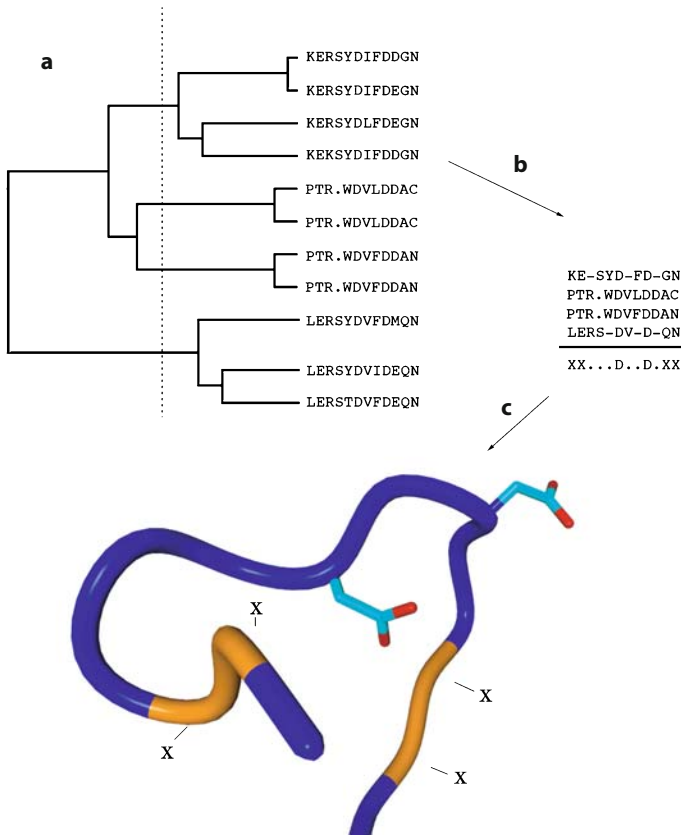


Fig. 6 The ET method. (a) The sequences in a protein family are aligned and a tree is generated. (b) For each class, a consensus sequence is created. The consensus sequences are compared, and trace residues are selected. (c) The three dimensional protein structure is used to map the trace residues. A functional site is indicated by the clustering of these trace residues.

These studies suggest that ET can be used for understanding protein functions if it can be applied at a large scale. Madabushi et al. (2002, 2004) streamlined the input preparation for an automated ET implementation. They also developed formal statistics to assess the significance of trace clusters, and tested its performance on proteins with diverse folds, structures, and evolutionary history.

The ET method is by no means the only method available that performs this kind of analyses; we simply took ET as it is representative for a large class of methods. A number of other good tools and methods have been published that use positional comparison of amino acid types (e.g. TreeDet (Carro et al. 2006); SDPpred (Kalinina et al. 2004); etc.).

4.2 Entropy-variability analysis

Conservation and variability patterns can be correlated with external information or with each other. Oliveira et al. (2003a) have developed a sequence analysis technique that harvests the information that is implicitly present in the variability and conservation at one single position in a MSA. This method is based on the combination of two sequence variability measures. The first is a Shannon-type entropy, while the second, variability, is simply the number of different amino acid types observed at one position in a multiple sequence alignment. They showed that there is a relation between the function of a residue and its location in a plot of entropy versus variability.

The method was tested on four protein families for which very many sequences are available and for which the function of nearly all residues have been well-established experimentally: globin chains, GPCRs, ras-like proteins, and serine-proteases (Oliveira et al. 2003a). Positions related to the main function, related to co-factor or regulator binding, positions in the core of the protein, and positions not associated with any known function all tend to cluster in separate areas in the entropy-variability plots

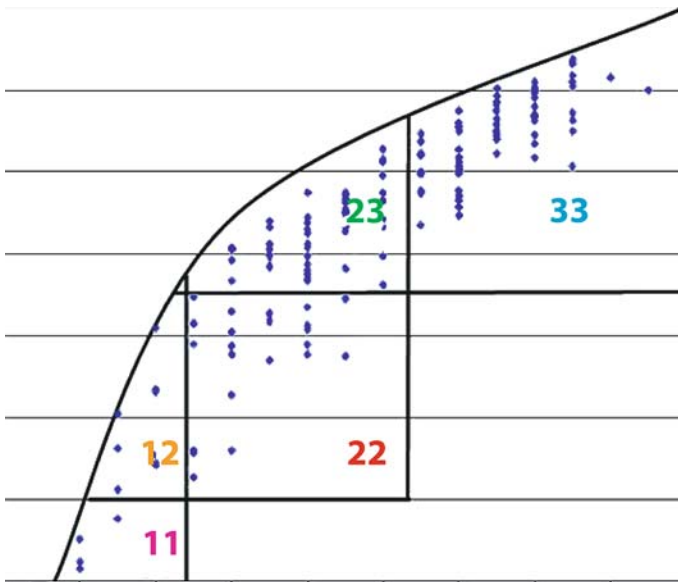


Fig. 7 Entropy-variability plot indicating the relation between variability patterns and residue function. Residues in area 11 perform the main function of the proteins (G-protein binding in GPCRs). Residues in area 12 provide support to the residues in area 11. Residues in area 23 are involved in modulator binding (ligand binding in GPCRs). Residues in area 22 tend to be located between residues in the areas 23 and 12 and tend to be responsible for communication between these two sites. Finally, residues in box 33 are seldom found involved in any function.

(see Fig. 7). This entropy-variability method is inductive; it can, for example, predict that a residue is involved in signal transduction, but additional information is needed to determine which signal that is, and how the signalling is done; or it can determine that a residue is involved in ligand/modulator binding but not what ligand or modulator is bound or how this binding takes place.

Folkertsma et al. (2004) used entropy-variability analysis together with structure and mutation data to find the functions of residues in the ligand-binding domain of nuclear receptors. Shulman et al. (2004) obtained similar results using statistical coupling analysis, a powerful but much more complicated method that produces results that are highly similar to the entropy-variability analysis method.

4.3 Sequence harmony

Members of the GPCR family can be activated by a wide variety of ligands. These ligands range from a few atoms to large molecules. Certain residue positions in the ligand binding area therefore tend to be involved in ligand binding in just a few GPCR families. To find such positions, a method is needed that searches for residue positions that are conserved in certain families, but not in others. The sequence harmony (SH) method was developed by Pirovano and Feenstra (2006, 2007). In contrast to CMA methods mentioned above, which focus on sites that are conserved in one or both groups and subsequently select those sites that are different between these groups, the sequence harmony method can also detect sites that are not highly conserved within each of the groups. The input of the SH method is a multiple sequence alignment, which is split into two groups. For each group, as well as for the combined groups, entropies are calculated. The SH score is calculated using these entropies. Ye et al. (2008) used the sequence harmony method to find GPCR residues that seem important for the function of just a few of the many GPCR families while being functionally unimportant in all other classes.

5 Discussion

The main topic of this book is determining the function of proteins, and clearly, if we want to fully harvest the wealth of information available in the human genome, that is the first and most important thing to do. After all, how can we make progress in system biology if we don't know what to do with large numbers of important proteins? But after we have determined what a protein does, the natural next question is how it does it. Several other chapters also allude to this topic. We listed a small, but representative, series of commonly used techniques that have in common that a multiple sequence alignment is at their hearth, and we showed what these techniques could do in the field of GPCR research. Correlation studies and entropy/variability measures often provide a

lot of information. This might not always be the exact and correct answer to the questions asked, but these methods most certainly do steer experimental work very well. The strength of these methods was perhaps best illustrated by a study that revealed individual atomic contacts between a ligand and a GPCR many years before the first structural information became available.

References

- Aloy P, Querol E, Aviles FX, Sternberg MJ (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol* 311: 395–408
- Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* 307: 447–463
- Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28: 45–48
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235–242
- Beukers MW, Kristiansen I, AP IJ, Edvardsen I (1999) TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data. *Trends Pharmacol Sci* 20: 475–477
- Campagne F, Jestin R, Reversat JL, Bernassau JM, Maigret B (1999) Visualisation and integration of G protein-coupled receptor related information help the modelling: description and applications of the Viseur program. *J Comput Aided Mol Des* 13: 625–643
- Carro A, Tress M, de Juan D, Pazos F, Lopez-Romero P, del Sol A, Valencia A, Rojas AM (2006) TreeDet: a web server to explore sequence space. *Nucleic Acids Res* 34: W110–W115
- Casari G, Sander C, Valencia A (1995) A method to predict functional residues in proteins. *Nat Struct Biol* 2: 171–178
- Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5: 823–826
- Consortium EP (2004a) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306: 636–640
- Consortium EP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korb J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816
- Consortium IHGS (2004b) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945
- Dean MK, Higgs C, Smith RE, Bywater RP, Snell CR, Scott PD, Upton GJ, Howe TJ, Reynolds CA (2001) Dimerization of G-protein-coupled receptors. *J Med Chem* 44: 4595–4614

- DesJarlais RL, Sheridan RP, Seibel GL, Dixon JS, Kuntz ID, Venkataraghavan R (1988) Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J Med Chem* 31: 722–729
- Edvardsen O, Reiersen AL, Beukers MW, Kristiansen K (2002) tGRAP, the G-protein coupled receptors mutant database. *Nucleic Acids Res* 30: 361–363
- Excellence TBNo (2005) Research networks: BioSapiens: a European network for integrated genome annotation. *Eur J Hum Genet* 13: 994–997
- Feenstra KA, Pirovano W, Krab K, Heringa J (2007) Sequence harmony: detecting functional specificity from alignments. *Nucleic Acids Res* 35: W495–W498
- Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R (2005) Optimal docking area: a new method for predicting protein–protein interaction sites. *Proteins* 58: 134–143
- Filizola M, Olmea O, Weinstein H (2002) Prediction of heterodimerization interfaces of G-protein coupled receptors with a new subtractive correlated mutation method. *Protein Eng* 15: 881–885
- Filizola M, Weinstein H (2002) Structural models for dimerization of G-protein coupled receptors: the opioid receptor homodimers. *Biopolymers* 66: 317–325
- Folkertsma S, Van Noort P, Van Durme J, Joosten HJ, Bettler E, Fleuren W, Oliveira L, Horn F, de Vlieg J, Vriend G (2004) A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. *J Mol Biol* 341: 321–335
- Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM (2006) A method for localizing ligand binding pockets in protein structures. *Proteins* 62: 479–488
- Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19: 163–164
- Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18: 309–317
- Gouldson PR, Dean MK, Snell CR, Bywater RP, Gkoutos G, Reynolds CA (2001) Lipid-facing correlated mutations and dimerization in G-protein coupled receptors. *Protein Eng* 14: 759–767
- Gouldson PR, Higgs C, Smith RE, Dean MK, Gkoutos GV, Reynolds CA (2000) Dimerization and domain swapping in G-protein-coupled receptors: a computational study. *Neuropsychopharmacology* 23: S60–S77
- Gouldson PR, Snell CR, Bywater RP, Higgs C, Reynolds CA (1998) Domain swapping in G-protein coupled receptor dimers. *Protein Eng* 11: 1181–1193
- Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33: D514–D517
- Hernanz-Falcón P, Rodríguez-Frade JM, Serrano A, Juan D, del Sol A, Soriano SF, Roncal F, Gómez L, Valencia A, Martínez-A C, Mellado M (2004) Identification of amino acid residues crucial for chemokine receptor dimerization. *Nat Immunol* 5: 216–223
- Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. *Science* 268: 1144–1149
- Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G (2003) GPCRDB information system for G-protein-coupled receptors. *Nucleic Acids Res* 31: 294–297
- Horn F, Bywater R, Krause G, Kuipers W, Oliveira L, Paiva AC, Sander C, Vriend G (1998a) The interaction of class B G-protein-coupled receptors with their hormones. *Recept Channels* 5: 305–314
- Horn F, Lau AL, Cohen FE (2004) Automated extraction of mutation data from the literature: application of MuteXt to G-protein-coupled receptors and nuclear hormone receptors. *Bioinformatics* 20: 557–568
- Horn F, Van der Wenden EM, Oliveira L, IJzerman AP, Vriend G (2000) Receptors coupling to G proteins: is there a signal behind the sequence? *Proteins* 41: 448–459
- Horn F, Vriend G, Cohen FE (2001) Collecting and harvesting biological data: the GPCRDB and NuclearRDB information systems. *Nucleic Acids Res* 29: 346–349

- Horn F, Weare J, Beukers MW, Hörsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G (1998b) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res* 26: 275–279
- Howard AD, McAllister G, Feighner SD, Liu Q, Nargund RP, Van der Ploeg LH, Patchett AA (2001) Orphan G-protein-coupled receptors and natural ligand discovery. *Trends Pharmacol Sci* 22: 132–140
- Lutje Hulsik D (2002) Public-domain database of GPCR interaction parameters. *Trends Pharmacol Sci* 23: 258–259
- Jones S, Thornton JM (1997) Prediction of protein–protein interaction sites using patch analysis. *J Mol Biol* 272: 133–143
- Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res* 32: W424–W428
- Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, Browne P, van den Broek A, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Gamble J, Diez FG, Harte N, Kulikova T, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Sobhany S, Stoehr P, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* 33: D29–D33
- Kazius J, Wurdinger K, van Iterson M, Kok J, Bäck T, IJzerman AP (2007) GPCR NaVa database: natural variants in human G-protein-coupled receptors. *Hum Mutat*
- Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R (2007) PIER: protein interface recognition for structural proteomics. *Proteins* 67: 400–417
- Kuipers W, Link R, Standaar PJ, Stoit AR, Van Wijngaarden I, Leurs R, IJzerman AP (1997a) Study of the interaction between aryloxypropanolamines and Asn386 in helix VII of the human 5-hydroxytryptamine1A receptor. *Mol Pharmacol* 51: 889–896
- Kuipers W, Oliveira L, Vriend G, IJzerman AP (1997b) Identification of class-determining residues in G protein-coupled receptors by sequence analysis. *Receptors Channels* 5: 159–174
- Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule–ligand interactions. *J Mol Biol* 161: 269–288
- Lamb ML, Jorgensen WL (1997) Computational approaches to molecular recognition. *Curr Opin Chem Biol* 1: 449–457
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860–921
- Lichtarge O, Bourne HR, Cohen FE (1996a) Evolutionarily conserved Galphabeta gamma binding surfaces support a model of the G protein–receptor complex. *Proc Natl Acad Sci USA* 93: 7507–7511
- Lichtarge O, Bourne HR, Cohen FE (1996b) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257: 342–358
- Lichtarge O, Yamamoto KR, Cohen FE (1997) Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J Mol Biol* 274: 325–337

- Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J Biol Chem* 279: 8126–8132
- Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J Mol Biol* 316: 139–154
- Miranker A, Karplus M (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* 11: 29–34
- Mirny L, Shakhnovich E (2001) Evolutionary conservation of the folding nucleus. *J Mol Biol* 308: 123–129
- Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* 291: 177–196
- Möller S, Vilo J, Croning MD (2001) Prediction of the coupling specificity of G protein coupled receptors to their G proteins. *Bioinformatics* 17 (Suppl 1): S174–S181
- Oliveira L, Paiva ACM, Vriend G (1993) A common motif in G protein-coupled seven transmembrane helix receptors. *J Comp Aided Mol Des* 7: 649–658
- Oliveira L, Paiva ACM, Vriend G (1999) A low resolution model for the interaction of G-proteins with G-protein-coupled receptors. *Protein Eng* 12: 1087–1095
- Oliveira L, Paiva ACM, Vriend G (2002) Correlated mutation analyses on very large sequence families. *Chembiochem* 3: 1010–1017
- Oliveira L, Paiva PB, Paiva ACM, Vriend G (2003a) Identification of functionally conserved residues with the use of entropy-variability plots. *Proteins* 52: 544–552
- Oliveira L, Paiva PB, Paiva ACM, Vriend G (2003b) Sequence analysis reveals how G-protein-coupled receptors transduce the signal to the G-protein. *Proteins* 52: 553–560
- Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein–protein interaction. *J Mol Biol* 271: 511–523
- Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 17: 700–712
- Petrokovski S, Henikoff JG, Henikoff S (1996) The Blocks database – a system for protein classification. *Nucleic Acids Res* 24: 197–200
- Pirovano W, Feenstra KA, Heringa J (2006) Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res* 34: 6540–6548
- Russell RB, Sasieni PD, Sternberg MJ (1998) Supersites within superfolds. Binding site similarity in the absence of homology. *J Mol Biol* 282: 903–918
- Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9: 56–68
- Shatsky M, Nussinov R, Wolfson HJ (2002) Flexible protein alignment and hinge detection. *Proteins* 48: 242–256
- Shenkin PS, Erman B, Mastrandrea LD (1991) Information-theoretical entropy as a measure of sequence variability. *Proteins* 11: 297–313
- Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R (2004) Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* 116: 417–429
- Singer MS, Vriend G, Bywater RP (2002) Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Protein Eng* 15: 721–725
- Slep KC, Kercher MA, He W, Cowan CW, Wensel TG, Sigler PB (2001) Structural determinants for regulation of phosphodiesterase by a G-protein at 2.0 Å. *Nature* 409: 1071–1077
- Sowa ME, He W, Slep KC, Kercher MA, Lichtarge O, Wensel TG (2001) Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat Struct Biol* 8: 234–237
- Valencia A, Pazos F (2002) Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol* 12: 368–373

- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, et al. (2001) The sequence of the human genome. *Science* 291: 1304–1351
- Wang W, Donini O, Reyes CM, Kollman PA (2001) Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Bioph Biom* 30: 211–243
- Ye K, Anton Feenstra K, Heringa J, Ijzerman AP, Marchiori E (2008) Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a machine-learning approach for feature weighting. *Bioinformatics* 24: 18–25
- Zuckermandl E (1965) Evolutionary divergence and convergence in proteins. In: Zuckermandl E, Pauling L (eds) *Evolving genes and proteins*. Academic Press

SECTION 5

Protein structure prediction

CHAPTER 5.1

Structure prediction of globular proteins

A. Tramontano¹, D. Jones², L. Rychlewski³, R. Casadio⁴, P. Martelli⁴, D. Raimondo¹
and A. Giorgetti¹

¹Department of Biochemical Sciences, Sapienza University, Rome, Italy

²Department of Computer Science, University College London, London, UK

³BioInfoBank Institute, Poznan, Poland

⁴Department of Biology, Alma Mater University, Bologna, Italy

1 The folding problem

The previous chapters should have convinced the reader that understanding protein function is an essential problem in biomedical and biotechnological sciences while, at the same time, a rather elusive one. Protein function is, by and large, determined by the protein three-dimensional structure with some exceptions that we will not discuss here.

One possible route to annotate a genome is to try and assign a structure to the protein products of the genes. In principle one could follow two routes: a physico-chemical approach whereby one tries to calculate the protein structure, or a heuristic approach where rules relating sequence to structure are derived from the analysis of known protein structures that have been experimentally determined.

The first route is clearly much more intellectually appealing. After all, given a protein sequence we know exactly its chemical composition, if we do not consider post-translational modifications, and all we need to know are the forces acting on each of the atoms so that we can compute their optimal relative position.

In order to follow this route we need to make sure that the functional protein structure is the conformation corresponding to the free energy minimum and, if this is the case, that we are able to calculate the energy of all possible protein conformation accurately enough to distinguish between the correct structure and all the others.

If one takes a folded protein, i.e. a protein in its functional conformation, places it in chemical conditions where all the forces are weakened and therefore where the protein unfolds, it is sufficient to remove the chemical agents used for denaturing the protein to recover the folded functional protein. This is the result of a very elegant experiment

Corresponding author. Anna Tramontano, Department of Biochemical Sciences, Sapienza University, Rome, Italy (e-mail: anna.tramontano@gmail.com)

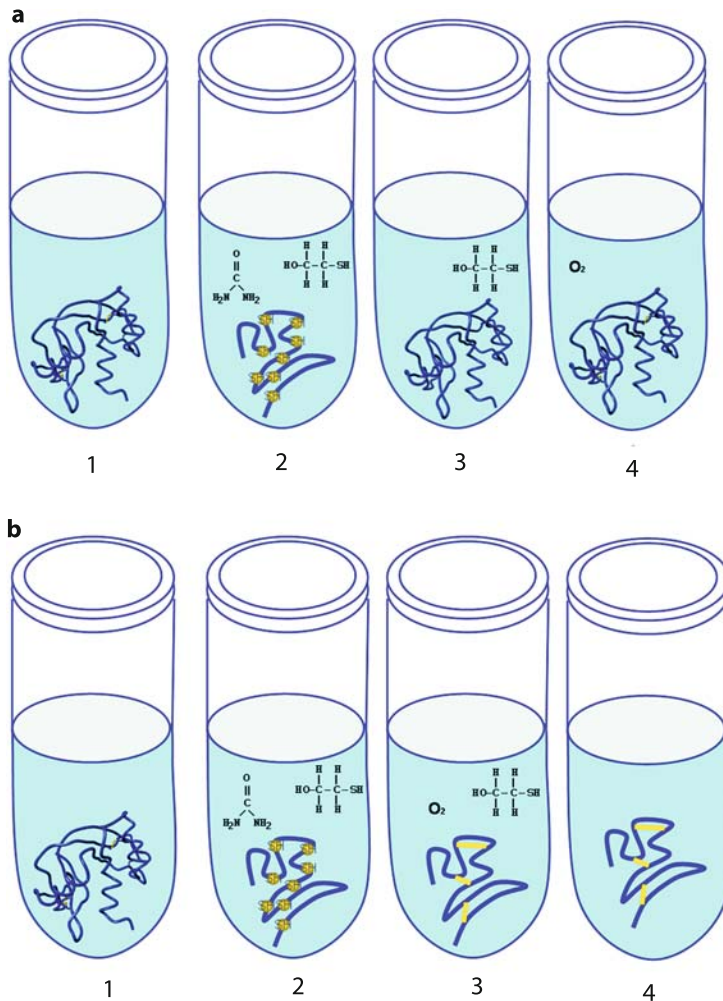


Fig. 1 Scheme of the Anfinsen experiment. The protein used for the experiment is Ribonuclease A, an extracellular enzyme of 124 residues with four disulfide bonds. In the experiment schematized in panel a, the S-S bonds were first reduced to -SH groups using mercaptoethanol (HS-CH₂-CH₂-OH). Next, the protein was denatured with 8 M urea. Anfinsen verified that, in these conditions, the protein is inactive. Next he removed urea by dialysis and oxidized the -SH groups back to S-S bonds. The protein regained its activity. However, the experiment would not be complete without its associated control, because we would have no proof that the protein was really completely unfolded in step 2. In the control experiment, shown in panel b, the protein was first reduced and denatured as in a, but in step 3, the enzyme was first oxidized to form S-S bonds, and then the urea was removed. The final protein was only 1-2% active. This implies that, in step 2 of the experiment the protein behaved as a flexible random polymer, so that disulfide bonds would form between different pairs of cysteines in each molecule at random. The number of possible pairs that can be formed is 105, therefore only one protein out of 105 will, on average, have the correct pairing and will therefore be able, once the urea has been removed, to form the native functional structure

performed by Christian Anfinsen in 1973 (Anfinsen 1973) (Fig. 1). The obvious interpretation of the experiment is that a protein sequence contains all the information needed to achieve its functional structure (the experiment is carried out in a test tube where there is nothing else but the protein) and that the functional or native structure is the one corresponding to the free energy minimum among those that the protein can explore (no matter how many times you repeat the experiment you always end up with the same final structure). Therefore we can assume that the native protein structure is the one corresponding to the free energy minimum (the limits of validity of this assumption are discussed later in this chapter).

All we need to do is to compute the energy of all possible conformations of a protein and select the one with minimum free energy. However there are at least two hurdles in this strategy, the first is that proteins are only marginally stable, i.e. the energy needed to unfold them is of the order of a few kcal/mol and is brought about by a very large number of weak interactions, and therefore we would need to compute the energy of each interaction very accurately to distinguish between the native protein structure and all the others. The second is that the number of possible conformations of proteins is simply enormous. There are many interesting attempt to try and simulate the folding of a protein in a computer using various tricks, approximations and strategies, as it will be discussed later in this chapter, but in practice we do not have at the moment any method that can fold any protein only on the basis of the physico-chemical properties of its sequence and we have to recur to heuristic methods by exploiting the fact that we have access to several solved instances of our problem: all proteins of known sequence whose structure has been solved experimentally.

The enormous number of conformations available to a protein not only makes the task of computing them impossible, but implies that the protein itself cannot be randomly searching its conformational space.

The case against proteins searching conformational space for the global minimum of free energy was argued by Cyrus Levinthal in 1968 (Levinthal 1968). The Levinthal paradox, as it is commonly known, can be demonstrated fairly easily. If we consider a protein chain of N residues, we can estimate the size of its conformational space as roughly 10^N states. This assumes that the main chain conformation of a protein may be adequately represented by a suitable choice from just 10 different local conformations per residue. More technically, the assumption is that there are just 10 different common combinations of phi, psi and omega torsion angles for each residue type. This of course neglects the additional conformational space provided by the side chain torsion angles, but is a reasonable rough estimate, albeit an underestimate. The so-called paradox comes from estimating the time required for a protein chain to search its conformational space for the global energy minimum. Let's think about a typical protein chain of length 100 residues and let's assume that the atoms can move very fast – the speed of light even. Even at these physically impossible atom velocities, it would take the chain around 10^{82} seconds to search the entire conformational space, which compares rather

unfavourably to the estimated age of the Universe (10^{17} seconds). Clearly proteins do not fold by searching their entire conformational space.

2 The evolution of protein structures and its implications for protein structure prediction

By and large, proteins evolve by accumulating mutations (amino acid replacements, insertions and deletions) which can be transmitted to the progeny and fixed in the population if they do not alter the functionality of the protein. At some stage of the evolution of a species, some individuals might diverge sufficiently to give rise to a different species, i.e. become unable to interbreed in the wild producing fertile offspring with the other members of the originating species.

As we mentioned, proteins have limited stability brought about by a multitude of rather weak interactions among their atoms. This suggests that the delicate balance between destabilizing and stabilizing forces might be easily destroyed by a mutation and the mutated protein might not be able to fold. However, during evolution, function has to be preserved, therefore all the proteins that we observe can only contain non destabilizing mutations with respect to their immediate ancestor sequence. Can a small change destabilize the original protein structure and stabilize a completely different one, preserving stability, function, folding ability, etc.? This is rather unlikely, and indeed never observed. It follows that evolutionarily related proteins, that is proteins derived by a common ancestor via the accumulation of small changes, cannot but have similar structure, where mutations have been accommodated only causing small local rearrangements. If the number of changes, that is the evolutionary distance, is high these local rearrangements can cumulatively affect the protein structure and produce relevant distortions, but the general architecture, that is the fold, of the protein has to be conserved. On the other hand, if two proteins have evolved from a common ancestor it is likely that a sufficient proportion of their sequences has remained unchanged so that an evolutionary relationship can be deduced by their comparative analysis. Therefore if we can infer that two proteins are homologous, that is evolutionary related, the structure of one can be used as a first approximation of the structure of the other. This forms the basis of the technique known as comparative or homology modelling (Tramontano 2006).

How well is a protein structure preserved during evolution? In a seminal work, Chothia and Lesk (Chothia and Lesk 1986) analyzed 32 pairs of homologous proteins of known structure and asked the question of how much the *core* of the structures diverged as a function of the sequence identity (a rough measure of the evolutionary distance). There are several definitions of the core of a protein structure. In their work, Chothia and Lesk used an almost tautological definition of core as the part of the protein structures that is more conserved between the two homologous proteins under study. Regardless the specific definition, we can intuitively understand what the core of a

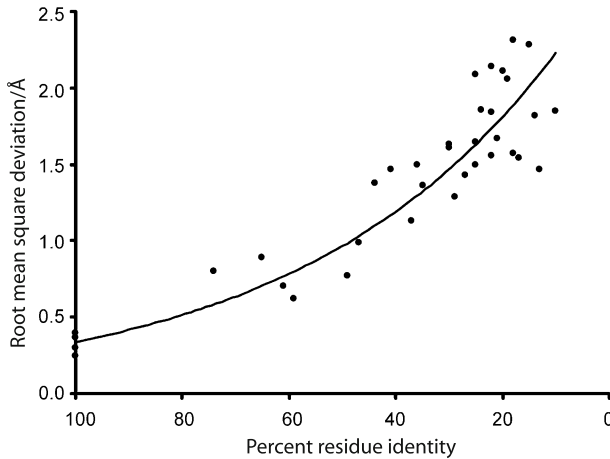


Fig. 2 A plot showing the result of the Chothia and Lesk experiment. They selected 32 pairs of homologous proteins plotted their structural similarity, measured as rmsd deviation in the core as a function of their sequence identity

protein is: the part of the structure that is not peripheral to the folded nucleus of the protein, i.e. the protein without external “decorations” such as loops and small domains that are usually not very well conserved in evolution. In the same paper, Chothia and Lesk also analyzed the extent to which the core is conserved as a function of sequence identity. Their conclusion, supported by many subsequent analysis, is that there is a clear relationship between the divergence of the structures of homologous proteins and that it can be expressed as a function of their sequence identity (Fig. 2).

3 Template based modelling

Given a protein of unknown structure, the target, the first step of a comparative modelling experiment is the detection of proteins evolutionarily related to it whose structure is known (templates). The next question we need to ask ourselves is: which amino acid of the target protein corresponds to which amino acid of the templates? In other words we need a sequence alignment between the target sequence and the sequence(s) of the template protein(s). This is, without doubts, the most crucial aspect of a modelling procedure and one of the most difficult ones. There are several methods for aligning protein sequences, but here is the catch. All these methods try to reconstruct the evolutionary history of the protein. In other words, they tell us which amino acids are likely to be derived from the same amino acid of the ancestral protein that gave origin to the present sequences. However, this is not necessarily the alignment we need for homology modelling. Let us try and explain this with an example. Suppose that there is

an insertion of one amino acid in a given position in our target sequence with respect to its template(s). Not only the inserted amino acid of the target does not have any equivalent amino acid in the template(s), but also the amino acids surrounding it are likely to have changed their position relative to the rest of the structure in order to accommodate the insertion and using their evolutionary counterparts as structural templates for their position is incorrect.

3.1 Homology-based selection of the template

The other question that we have to discuss is the selection of the best protein(s) to be used as template(s). If more than one protein of known structure evolutionarily related to the target is available we have several possible choices. We can:

- use the one evolutionarily closer to the target, i.e. the one with the highest sequence similarity,
- “average” the coordinates of the templates and build a “theoretical template”,
- take the structure of different regions from the different proteins selecting the regions where the local similarity is higher,
- build a model on the basis of each of the available templates and select the best one according to some criteria,
- derive constraints from the templates and subsequently build a structure that satisfies as many of them as possible.

Essentially, all these strategies are used in practice by different tools available to users (Sali and Blundell 1993; Kopp and Schwede 2006). It is difficult to say which is the best in general, although it is becoming clearer that using multiple templates has to be preferred and probably the constraint based strategy is more effective in many cases.

3.2 Fold recognition

Evolutionary related proteins share similarities in sequence and structure. It is however possible, and observed, that the two proteins have diverged so much that the sequence signal falls below the detection level. It is equally possible, and observed, that some topologies or folds are used by proteins presumably not sharing any evolutionary relationship.

Both observations allow us to reformulate the protein prediction problem in a different way. Rather than asking which is the structure of our protein, we can ask whether any of the known structures can represent a reasonable model for it, independently on our ability to detect an evolutionary relationship. This is equivalent to ask whether the sequence of our target protein fits any of the structures present in our database. Even if we detect such a fit, the structures might not be very similar, but still the

template protein can represent a sufficient approximation of our protein structure useful for many applications.

In this case, we have to face the problem of evaluating sequence structure fitness.

The available methods can be roughly divided into two categories, although several hybrid combinations are possible.

In one approach (Eisenberg et al. 1992), we can analyse the sequence and replace each amino acid with its propensity to be observed in a given structural environment, usually we take into account the propensity of an amino acid for being in one of the secondary structure types, of being in a hydrophobic or an hydrophilic environment and of being more or less exposed to a polar solvent. This will recast our sequence in a new sequence in a different alphabet. Whenever possible, we use a multiple alignment of all available proteins homologous to our target sequence, as we know that they will share the same (albeit unknown) structure. Next, we can analyse each of the proteins of known structure. For every position, we will not take into account which amino acid happens to be present, but rather examine the property of the position, i.e. its secondary structure, its environment and its exposition to the solvent and will assign a symbol describing the observed combination of properties. In this way our database of protein structures will be represented by a set of strings.

The string representing the query sequence or its multiple alignment can now be compared, with techniques similar to those described for the detection of evolutionary relationships, with each of the strings representing the structures. Once again we will need a background distribution with which we compare the obtained score. This can be obtained by reshuffling our sequence, or by creating reasonable “decoy” structures.

The other approach, known under the name of threading (Sippl 1995; Jones and Thornton 1996), builds as many models of the target proteins as there are structures in the database (or some reasonably selected subset of it) using each structure as a template, optimising a fitness parameter depending upon the interactions between the amino acids in the model. The optimisation can be achieved by a technique called double dynamic programming, or using some approximations that will not be discussed here.

The fitness function is usually a pair-wise potential between amino acid side chains reflecting the likelihood of the observed set of interactions. This is discussed later in this chapter.

The fold recognition methods expand significantly the set of proteins that we can model and often allow unexpected evolutionary relationships to be high-lighted, but it is much more difficult to evaluate their reliability a priori and they cannot guarantee that functional regions are more reliably predicted than the rest of the structure.

3.3 Using sequence based tools for selecting the template

The selection of the appropriate template in a template based prediction can be supported to some extent by tools able to predict local structural features starting

from the amino acid sequence. Secondary structure and presence of disulfide bonds are among the features that can be successfully predicted. Generally speaking, prediction tools rely on the fact that, even if the overall structure is determined by the whole sequence, specific structure features can be strongly influenced by local features of the sequence. For example, alpha-helices and beta-sheets have different amino acid composition, and the same is true for the neighboring residues of disulfide-bonded and free cysteines. If we are able to understand these differences, we can use them to evaluate the probability for a residue to be in a secondary structure or for a cysteine to be disulfide bonded.

The basic idea is to analyze the set of proteins known at atomic resolution and adopt methods suited to extract correlations between structural features and local sequence features. Simplest methods are based on classical statistics and evaluate, for example, the propensity of alanine residues to be in a alpha-helical structure simply by computing the ratio between the alanine composition of alpha-helices and the overall alanine composition in proteins. Statistical methods can take into consideration more elaborate sequence features, but they often fail in extracting useful correlations when the complexity of the problem increases. For that reason, more versatile and flexible methods have been designed and implemented on the basis of the so called “machine-learning” theory. Among them, Neural Networks, Support Vector Machines and Hidden Markov Models are the most widely adopted. With different strategies, they are able to extract information from a set of known examples in an automatic way, on the basis of a rigorous mathematical framework. Owing to their architectures, they are able to deduce more complex rules of association between input (sequence) and output (structural feature) than classical statistic methods do. These rules are encoded in a set of numerical parameters whose values are fixed during the training phase and then used for predicting new sequences.

Versatility of machine learning methods allows different input encodings, more informative than the sole sequence, to be considered. In particular a general improvement of the performance can be obtained using sequence profiles upon multiple sequence alignments. In practice, given a sequence, similar sequences are searched in the data base and then aligned so as to obtain a representation of a whole family instead of a simple sequence. This representation highlights, for example, the conserved and mutated residues and this supplements the predictor input with evolutionary information.

The classical application of predictive methods to protein structure is the determination of secondary structure starting from sequence. Best methods for this task are based on Neural Network and Support Vector Machines and take as input the sequence profile of a 15/25-residue long window, centered around the residue to be predicted. When validated on proteins with known structure not used during the training phase, these methods predict the correct secondary structure for about 78% of residues (Jacoboni et al. 2000; Ward et al. 2003). Better results can be obtained implementing a consensus of different methods (Cuff et al. 1998; Ward et al. 2003).

Another important structural feature that can be predicted is the presence of disulfide bonds, that is the bond between the sulfur atoms of two cysteine side chains. This is the only covalent bond that non adjacent residue can form in the native state and a correct prediction of the topology of disulfide connection strongly constrains the prediction of the overall structure. This task can be easily split into two steps. First of all, since only about 1/3 of cysteine residues are involved in disulfide bonding, it is necessary to discriminate them. Then the topology of the connections can be predicted. Concerning the first step, very efficient methods have been implemented that are able to predict the correct bonding state for 88% of cysteine residues and to give an overall correct prediction for 84% of proteins (Martelli et al. 2002). They are currently based on systems that integrate a Neural Network and a Hidden Markov Model. The former analyzes the composition of the profile in windows centered around each cysteine residue while the latter correlates the outputs that the neural networks computes for all the cysteine residue in the sequence (Martelli et al. 2004).

The prediction of the topology of the disulfide bridges, i.e. of which cysteine pairs with which, is more difficult due to the combinatorial number of possible connection patterns for a given number of bonded cysteine residues. Important achievements have been reached, although a reliable prediction of the disulfide connectivity pattern can be performed only when two or three disulfide bonds are present in the protein (Fariselli and Casadio 2001; Tsai et al. 2007).

In conclusion, the prediction of structural features starting from the sequence is not able to completely reconstruct the protein conformation. Nevertheless these procedures can greatly help this task since the predict constraints limit the number of possible conformations. Moreover the output of these tools can supply information useful in the implementation of fold recognition methods.

3.4 Completing and refining the model

Whatever the strategy for selecting the template and predicting the part of the target protein that is conserved with respect to the template, the next steps are the modelling of insertions and deletions and of the conformation of the side chains. For insertion and deletions, methods are usually based on either an energy driven search for the possible conformations of the region of interest or on a database search of regions of protein of known structure that can provide a local template (Tramontano et al. 1989; Brucoleri and Karplus 1990; Holm and Sander 1992; China et al. 1995; Tramontano 1995; Fiser et al. 2000). The latter are usually selected on the basis of either a good fit of the regions flanking the region between target and local template, or on local sequence similarity. Side chain modelling often takes advantage of the preference of side chains for specific conformations, as deduced by the analysis of known protein structures (Fig. 3). These preferences, tabulated in so called rotamer libraries (Godzik and Skolnick 1992;

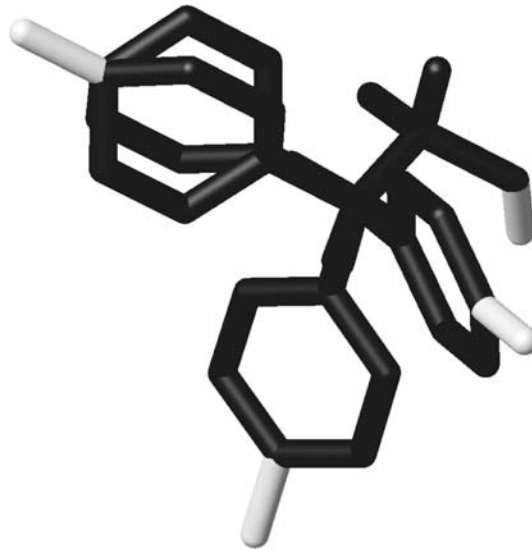


Fig. 3 Each amino acid has different preferred conformations of its side chain, called rotamers. In the figure the most frequently observed rotamers of the amino acid tyrosine are shown

Holm and Sander 1992; Chinae et al. 1995; Keller et al. 1995), are usually used as a starting point for subsequent refinement of the overall structure.

Once we have built our initial model, we need to “refine” it. What this simply means is that we now need to model the effect of the specific sequence changes that have occurred in our protein with respect to its template(s).

3.5 Current state of the art in template based methods

How well current modeling methods, combined with sequence based methods such as those described above, are able to predict the overall structure of a protein, the conformation of regions where insertions and deletions have occurred and to refine the model are some of the questions that the international community tries to answer by carrying out large scale blind tests, the CASP (Critical Assessment of Methods for Protein Structure Prediction) experiments (Moult 1996).

Every two years crystallographers and NMR spectroscopists who are about to solve a protein structure are asked to make the sequence of the protein available together with a tentative date for the release of the final coordinates. Predictors produce and deposit models for these proteins before the structures are made available and, finally, a panel of assessors compares the models with the structures as soon as they are available and tries to evaluate the quality of the models and to draw some conclusions about the state of the art of the different methods. The results are discussed in a meeting where assessors and

predictors convene and the conclusions are made available to the whole scientific community via the World Wide Web and the publication of a special issue of the journal *Proteins: Structure, Function, and Genetics*. The collected data, amounting to tens of thousands of models for hundreds of targets is an invaluable resource for assessing the quality of protein models.

Although embarrassing, we have to admit that, so far, no available method, is able to consistently produce the correct structure for regions where insertions and deletions are located or to improve the initial model and make it “better”, i.e. closer to the real structure, while the accuracy of side chain modeling methods seems to be only limited by the quality of the prediction of the rest of the structure.

Notwithstanding the limitations of comparative modeling, this method remains the method of choice whenever possible for at least two reasons. First of all, the relative quality of a comparative model depends on the evolutionary distance between two proteins. In fact, both the probability of inferring the correct alignment between two proteins and the structural divergence between their structures are correlated with their evolutionary distance which can be estimated *a priori*. This implies both that it is possible to estimate the expected quality of a comparative model and its possible range of application beforehand and hence decide whether it is reasonable to embark in the task and also, perhaps most importantly, that one can attach an approximate reliability to any of the conclusions derived from the analysis of the model. The second, equally important aspect, is that the methodology will be especially effective in modeling regions of a protein that are more conserved during evolution. This implies that functionally important regions will be more correctly modeled than other, often of lower interest, regions.

4 Template-free protein structure prediction

As we have seen, where a template structure can be found, it is relatively easy to build a realistic protein model that is reasonably close to the native structure. The closer related the template is to the target protein, generally the more accurate the final model will be. But what can be done if no template structure can be found? This can happen either because there is a suitable template available but from comparing the sequences it is impossible to find the correct template, or because the template simply is not present in the database i.e. the target protein in fact has a novel fold.

In the absence of a template structure it is possible to build a model, but this is a far more difficult process, with a much greater chance of producing a completely incorrect model.

Perhaps the most ambitious approach to predicting protein structure is to simulate the protein folding process itself using basic physics. Much research effort is expended in developing such simulation techniques, but as yet simulation is only useful for short peptides and small molecules. Simulation techniques are, however, very useful for

predicting unknown loop conformations as part of comparative modeling. The basic principle of simulation-based protein structure prediction is that the native fold of a protein can be found by finding the conformation of the protein which has the lowest energy as defined by a suitable potential energy function.

Unfortunately, the exact form of this energy function is as yet unknown, but it is reasonable to assume that it would incorporate terms pertaining to the types of interactions observed in protein structures, such as hydrogen bonding and van der Waals effects. The conceptual simplicity of this model for protein folding stimulated much early research into *ab initio* tertiary structure prediction. A successful *ab initio* approach necessitates the solution of two problems. The first problem to solve is to find a potential function for which the energy of the native conformation of the protein is the conformation of lowest energy. The second problem is to construct an algorithm capable of finding the global minimum of this function.

We briefly mentioned Levinthal's paradox before. There are many ways of explaining it away, but there is one quite simple explanation and that is that a protein folds by following a *folding pathway*. Imagine a very poor golfer playing golf on a flat golf course. The golfer hits the ball and it lands somewhere on the course. He keeps hitting the ball randomly until eventually he gets a hole in one. Clearly our unfortunate golfer will take a long time to reach his target of getting the ball into the hole. Now imagine a similarly unskilled golfer, but one who has the good fortune of playing on a hilly golf course. He hits the ball and it rolls down a slope until it reaches the bottom of the slope. He hits the ball again and it rolls down another slope. Now imagine that on this golf course, the hole has been placed at the very lowest point on the course (in energy terms we would call this the global energy minimum). It doesn't require much thinking to realize that the second golfer will find the hole long before the first one. Every time the second golfer hits the ball it has a good chance of getting closer to the target – because most paths the ball will take will be downhill and therefore most likely closer to the target. This is a good mental model of how a protein can find its global energy minimum without testing all possible conformations. Despite the fact that a long protein chain (like our unlucky first golfer) cannot find its global energy minimum by a blind global search, small *pieces* of the chain (say 5–10 residues) can quite easily locate their own global energy minimum within the average lifetime of a protein. These folding fragments can be thought of as the equivalents of the downhill moves made by our lucky second golfer. It is generally thought that the location of the native fold for a protein is located by the folding of such short fragments (Moult and Unger 1991). Thus, Levinthal's paradox is only a paradox if the free energy function forms a highly convoluted energy surface, with no obvious downhill paths leading to the global minimum. The folding of short fragment can be envisaged as the traversal of a small downhill segment of the free energy surface (the golf ball rolling down a small slope), and if these paths eventually converge on the global energy minimum, then the protein is provided with a simple means of rapidly locating its native fold.

One subtle point to make in passing about the relationship between the minimization of a protein's free energy and protein folding is that the native conformation need not necessarily correspond to the global minimum of free energy. One possibility is that the folding pathway initially locates a local minimum, but a local minimum which provides stability for the average lifetime of the protein. In this case, the protein in question would always be observed with a free energy slightly higher than the global minimum *in vivo*, until it is degraded by cellular proteases, but would eventually locate its global minimum if extracted from the cell, purified, and left long enough *in vitro* - though the location of the global minimum could take many years. Thus, a biologically active protein could in fact be in a metastable state, rather than a stable one.

For template free protein modeling, the most successful approaches take some (but not all) of the ideas from experimental protein folding studies and apply them to making a computationally efficient search for the most likely protein fold for any given target sequence. There is no guarantee that methods which allow computers to predict correct folds are related to the actual protein folding.

A typical *de novo* prediction method can be described along the following lines:

1. Define a (possibly simplified) representation of a polypeptide chain and the surrounding solvent. In fact, usually the surrounding solvent is left out of the calculation entirely to save computational effort.
2. Define some kind of scoring function which attempts to identify native-like structures for the given protein. This can be an energy function modeled on what we believe to be the physicochemical forces found to stabilize real proteins or it can be a completely arbitrary scoring function based on our knowledge of protein structures.
3. Search for the protein chain conformation which has the lowest energy - generally by some kind of restricted random search, or by perhaps assembling pieced of existing protein structures in order to build plausible new structures.

This general approach to *de novo* protein structure prediction has a long history, starting with the pioneering work by Scheraga and co-workers in the 1970s (Burgess and Scheraga 1975), Warshel and Levitt's work in 1975 (Levitt and Warshel 1975), Kolinski and Skolnick's work with simplified lattice models in the 1990s (Kolinski and Skolnick 1994a; Kolinski and Skolnick 1994b) and many others. The reason for there remaining such an interest in these methods for folding proteins is that there is a general feeling that this kind of approach may help to answer some of the questions pertaining to how proteins fold *in vivo*. The basic idea is that by keeping as close to real physics as possible, even simplified models of protein folding will provide useful insights into protein folding processes. In recent years, the term *ab initio* has been reserved for methods which attempt to mimic real protein folding as much as possible. However, despite a few recent successes on small proteins, such *ab initio* methods are still not able to predict protein structure with much success.

All of the recent success in template-free modeling has really come from knowledge-based methods. Knowledge-based prediction methods attempt to predict structure by applying rules. These rules are based on observations made on known protein structures. A trivial example of such a rule might be, for example, that proline residues are uncommon in alpha-helices or that hydrophobic residues are commonly found buried in the core of a stable protein.

4.1 Energy functions for protein structure prediction

There are many different energy functions described in the literature, all of which have been applied at some time or another to protein modeling in some form or other. So-called *classical potentials* attempt to model the basic physicochemical effects that are known to stabilize protein folds e.g. electrostatic effects, covalent bonding, van der Waals effects and so on. These potentials are not very useful for *de novo* protein modeling as they generally do not take into consideration the entropic effects of solvent that surrounds the protein. Although potentials based on such physical principles could be used for protein structure prediction if the solvent is explicitly built into the simulation in some way, and are frequently used for refining protein models, it has generally been the case that the best results in *de novo* modeling have been obtained by using *knowledge-based potentials* i.e. potentials that are based on the observed statistics of atoms in real protein structures. Such potentials are called *potentials of mean force*.

One very basic principle in statistical physics is that the difference in energy between two states of a system is related to the *transition probability* between the states. In the case of protein folding, the two states we typically consider are the folded (i.e. native state) and the unfolded states. Suppose we wish to estimate how much a close contact between two alanine residues helps to stabilize the native folded state of proteins. Let us suppose that we count the number of times we see this particular “event” across many different folded proteins (i.e. all the proteins found in the Protein Data Bank). Let us take this count of alanine-alanine contacts and calculate the fraction of all alanine-alanine pairs that are in fact in close contact. Let’s call this value $P_{\text{folded}}(\text{Ala}-\text{Ala})$. Let us also suppose that we have access to the structures of many different unfolded proteins and we calculate a similar value $P_{\text{unfolded}}(\text{Ala}-\text{Ala})$. By comparing these two probabilities we can make an estimate of the free energy change contributed by two contacting alanine residues when going from the unfolded to the folded state. This free energy change upon folding for an Ala-Ala interaction is obtained using the inverse Boltzmann equation as follows:

$$\Delta E(\text{Ala} - \text{Ala}) = -kT \ln \left(\frac{P_{\text{folded}}(\text{Ala} - \text{Ala})}{P_{\text{unfolded}}(\text{Ala} - \text{Ala})} \right)$$

where k is the Boltzmann constant and T is the temperature (typically room temperature). It should be noted that for protein structure prediction we can ignore kT as it is

simply a multiplicative constant that gives the energy in particular units (kcal per mole usually).

Of course, in practice we have no knowledge of the structures of unfolded proteins, and so we have to estimate this probability based on the assumption that unfolded proteins adopt completely random conformations.

For most practical protein structure prediction potentials, the interactions between residue pairs are usually sub-classified according to their sequence separation (i.e. how far apart they are in the amino acid sequence) and their spatial separation (i.e. how far apart they are in 3-D space). The reason for this is that the folding properties of a short chain are different from those of a long chain, because of the differing number of bonds (the ends of a long chain can obviously move further apart than the ends of a short chain). As a result of this, the interactions between two amino acids close together in the sequence need to be treated differently from those which are far apart in the sequence.

Although many groups use such knowledge-based potentials, there is, not surprisingly, little agreement on exactly how they should be calculated or even which atoms should be included in the calculations. Some groups (e.g. some of the current authors) restrict their calculations to just the beta-carbons of each amino acid. Other groups calculate their potentials for all of the atoms in the protein side chains. Some groups attempt to model the effects of solvent on each amino acid by calculating their solvation free energy i.e. the free energy change when a single amino acid is moved from being in solvent to being buried in the core of a protein. How these different terms should best be combined is another source of debate. Nevertheless, despite all these variations the aim is the same each time i.e. to compute a number which produces a lower value for the native structure than for any alternative structures. No energy functions manage to achieve this in the majority of cases, but even quite inaccurate energy functions can still help identify useful protein models from a small set of alternatives.

4.2 Lattice methods

Before describing the most widely used approach to template-free modeling (fragment assembly) it is worth mentioning an earlier approach to the problem, namely the use of a lattice approach to protein folding. The idea of lattice-based methods is to limit the number of protein conformations searched by restricting the coordinates of each atom to lie on a set of regular points in space (the lattice). The simplest type of lattice is a cubic lattice where for example, atoms can only be placed at the vertices of cubes with sides of length 3.8 Å (the normal distance between alpha carbons in a polypeptide chain), say. However, other lattice arrangements have been proposed e.g. tetrahedral lattices. Because the positions of the atoms are restricted, the total number of possible conformations is restricted and so a deep search of available conformations can be carried out reasonably efficiently.

The main problem with lattice approaches is that it is difficult to accurately model the geometry of a protein chain when the atomic coordinates are restricted to lattice coordinates. To get around this problem, modern lattice methods such as those developed by Skolnick and colleagues (2001) combine lattice simulations with off-lattice procedures. Once a low-energy conformation of the protein has been identified by a lattice search, the lattice is removed and the protein refined by normal energy minimization techniques to produce proper bonds lengths and angles.

4.3 Fragment assembly methods

By far the most successful methods for predicting protein structure in the absence of a template structure are those which attempt to assemble plausible protein folds from fragments of other proteins. The idea of *fragment assembly* was first used in comparative modeling (e.g. (Greer 1991)) where fragments of known structures were used to build missing loops after a framework of the protein had been built according to a template structure. The extension of fragment assembly from loop building to the accurate prediction of a complete novel protein fold without a template first demonstrated by David Jones in the 2nd CASP experiment held in 1996 (Jones 1997). At around the same time, David Baker and colleagues independently developed Rosetta (Simons et al. 1997), which employed a similar approach to protein structure prediction and went on to demonstrate great success in the 3rd CASP experiment and has performed consistently well in all subsequent CASP experiments. Now there are many published fragment assembly methods for predicting protein structure, though Rosetta still arguably remains the most successful.

A generic fragment assembly method (see Fig. 5) comprises 4 steps:

1. *Identification of fragments.* The target protein sequence is matched against fragments of known protein structures and *at each position in the sequence* a list of possible fragments is compiled.
2. *Random recombination of fragments.* Fragments are picked at random from the lists made in step 1 and joined together. Some combinations of fragments can be rejected immediately as they might produce a chain conformation that overlaps itself.
3. *Evaluation.* Assuming the chain does not have any serious clashes, it can be evaluated using an energy function to decide how well the generated structure matches the target protein.
4. *Return to Step 2* until no lower energy structures can be found. Generally speaking this is carried out with a *simulated annealing* strategy. In simulated annealing, conformational changes which produce an increase in energy are not immediately rejected. Such changes are only rejected according to a rule based on a notional system temperature, which is reduced as the algorithm proceeds. So, in the early stages of the simulation, every combination of fragments is considered acceptable, but towards the

end of the simulation only fragment swaps which reduce the energy are permitted. In between, fragment swaps are accepted according to a probability derived from the current temperature. A sequence of random numbers is used both to select the fragments and to decide whether or not to accept the generated conformation.

5. *Repeat from Step 1* and collect a large number of different final models. This is not really part of the algorithm itself, but it has proven to be very useful to rerun fragment assembly methods (with different random numbers of course) many times. In Rosetta (Simons et al. 1997), for example, the algorithm is run thousands of times and the thousands of different final structures are clustered to identify folds which frequently recur across the independent simulations. This clustering step probably serves to reduce the sampling error that comes from having an inaccurate energy function, but Baker has suggested that this might in fact serve as a crude model of the *molten globule state* in natural protein folding.

Different labs of course implement these steps in different ways. The main difference between programs is in the fragment selection step. In the FRAGFOLD method of Jones and colleagues (2001), the fragment selection focuses on fragments which correspond to super-secondary structural motifs e.g. beta-alpha-beta units or beta-hairpins. In Rosetta, Baker and colleagues mostly use short fragments (9 residues typically) of proteins. There are also differences in how each fragment is actually matched against the target protein. In FRAGFOLD, the selection is mostly by means of a threading approach i.e. an energy function is used to assess the compatibility of the fragment structure with that part of the target sequence (essentially fold recognition on a small scale). In Rosetta, the selection is mostly made on the basis of sequence similarity between the fragment and the corresponding part of the target sequence.

4.4 Practical considerations

As we have said, fragment assembly methods are now commonly used to model proteins of unknown structure. Less commonly, lattice-based methods are also employed, but fragment-based approaches represent the current state-of-the-art. Popular as they may be, what are the practical considerations of using fragment assembly? What are their limitations? Is the protein folding problem finally solved?

The answers to all these questions can be obtained by looking at recent results from CASP experiments and evaluating both the quality of the best models obtained and the fraction of cases where a good answer is obtained. Since CASP2, there has been a lot of progress. The fragment assembly prediction of the structure of NK-Lysin by David Jones in CASP2 (Jones 1997) was a watershed event because it was the first time a novel protein structure was correctly predicted in CASP. However, this was just a single correct prediction of a relatively small protein (78 residues). Since CASP2 in 1996, there have been many other examples of excellent novel fold predictions, particularly from the

Baker group's Rosetta method, but nonetheless the failures still outnumber the successes. If we look at the successes and compare them to the failures we can start to see what it takes for a template-free prediction to have a good chance of success:

1. *Size.* Successful predictions have typically been small proteins (<120 residues). The best prediction ever seen was for target T0281 in CASP6, where the Rosetta model was very close to the native structure (1.6 Å RMSD), but this protein was only 70 residues long (Bradley et al. 2005). There has been much less success in predicting larger proteins.
2. *Simple fold.* Looking at the most successful predictions, the predicted folds have been mostly simple in terms of topology. NK-Lysin in CASP2 was only a simple helix bundle, and even though T0281 in CASP6 did include a small beta sheet, it was still of very simple topology.
3. *Clear secondary structure.* Generally speaking, the most successful predictions have been made where the secondary structure of the target protein is very accurately predicted or even known from, say, NMR chemical shifts. Secondary structure prediction methods, for example the PSIPRED method (Jones 1999) are now very accurate (77–80% accurate in terms of predicting helix/strand/coil) if there are any many homologous sequences available for the target protein. Targets which are sequence orphans (have no known homologues) or very few homologues are often predicted very poorly by secondary structure prediction methods and consequently badly by fragment assembly methods. Evidently the best fragment assembly methods are highly dependent on good secondary structure predictions.

5 Automated structure prediction

Objective community-wide structure prediction benchmarking experiments such as CASP are the main driving force for the latest development of prediction protocols. These experiments taught us that experts utilizing diverse sources of information are more successful than groups relying on a single structure prediction method. Hints influencing the selection of templates for modeling may come from biological insights (recognition of active site residues, or characteristic secondary structure patterns) or from literature searches, which point to particular fold families hypothesized for homologs of the target. Unfortunately, such procedures are difficult to implement in an automated and reproducible fashion and remain restricted to a group of highly skilled structural biologists. As an alternative, scientists can increase the diversity of putative structure predictions by utilizing the growing number of prediction algorithms. In contrast to tailored literature scanning this approach can be easily automated. A framework to profit from the diversity of prediction methods was created by meta-servers, which collect and analyze models from many prediction services spread around the globe and present the results in a standardized, comparable fashion.

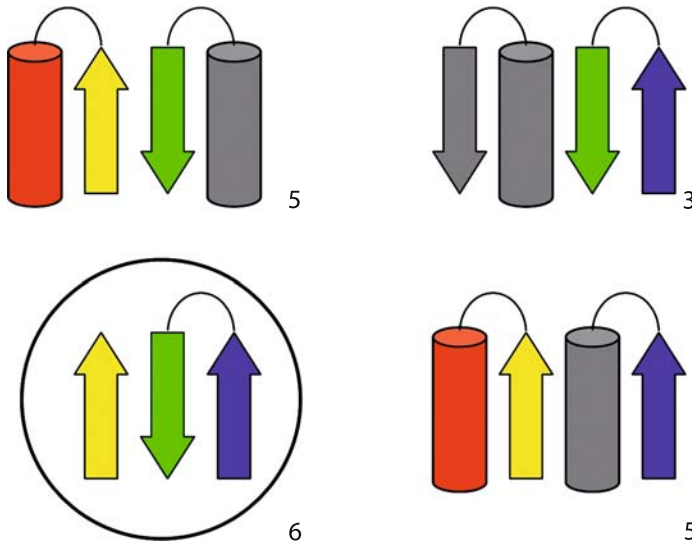


Fig. 4 A hypothetical set of four schematic protein structure models is shown. Structural elements, which our found in more than 1 model, are colored in red, yellow and blue. The model in the circle represents the consensus model because it consists of most abundant elements, i.e. each of its three structural elements is found in two other models (6 similarities). No other model has such a large number of similarities (5, 5 and 3 respectively)

A naïve selection of the most reliable model, based on the highest confidence score reported by employed methods, is hampered by the uniqueness of the different scoring functions. Such attempts did not result in more accurate prediction protocols than the best employed algorithm. The first successful attempt to benefit from the diversity of models was based on the selection of the most abundant fold in the set of high-scoring models, a procedure reminiscent of the clustering of models obtained in fragment based methods. First automated meta predictors selected from the presented set the model with the highest similarity to other models (see Fig. 4). Initial benchmarking results obtained in the first years confirmed that meta predictors are more accurate than independent methods (Bujnicki et al. 2001). Their strength is mainly attributed to the structural clustering of collected models. Even if many of them are wrong, it can be expected that structures of incorrectly predicted fragments of the models have random conformations. Only structures of fragments corresponding to preferred, native conformations occur with higher than expected frequency. The presence of clusters of similar models built using evolutionary distant templates provides additional support for the reliability of the fold assignment based on these clusters.

The promising evaluation results boosted further development of meta predictors. Currently available versions differ in several aspects: the way the models are compared,

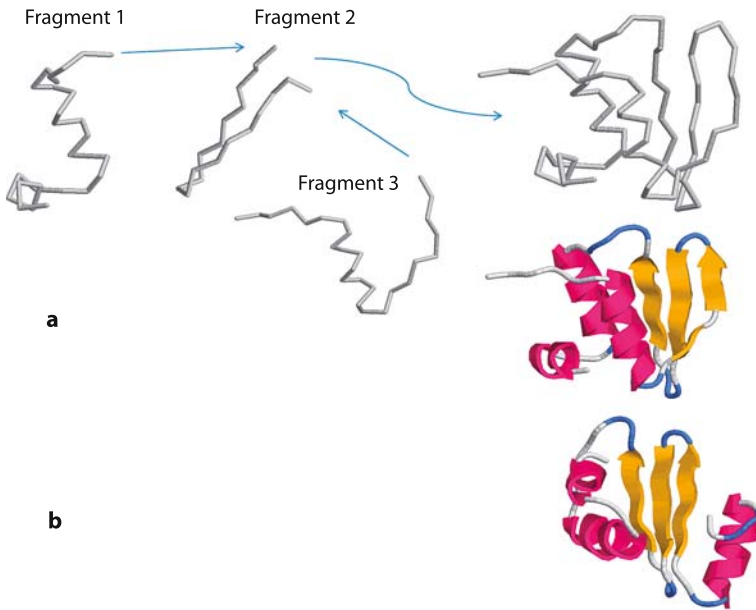


Fig. 5 A simple example of fragment assembly is shown. (a) Three small protein fragments are combined to form a complete small protein fold. (b) The same middle fragment is combined with two different fragments to form a different chain fold

the way the final model is generated and in the use of the initial scores assigned to the models by individual servers. When models do not exhibit high structural similarity, the initial scores assigned to each model by the original prediction method can be consulted to improve the selection procedure. However, this is not that simple, because different fold libraries and scoring schemes are employed by different prediction servers. Some meta predictors develop server-specific neural networks to translate the initial values into uniform scores. Others ignore the scores altogether and base their consensus evaluation only on the abundance of folds or structural motifs. The final consensus model is either identical to one of the original models or additional modifications are performed using one or multiple models as templates. Some meta predictors compile the final model by concatenating fragments taken from several initial models. Others use the selected consensus model as a starting point for a more complex and time consuming *ab initio* simulation.

5.1 Practical lessons from benchmarking experiments

Years of experience with benchmarking prediction methods (Fischer et al. 1999,2001,2003; Fischer and Rychlewski 2003; Koh et al. 2003; Eyrich et al. 2005;

Rychlewski and Fischer 2005) taught the community to treat the results with appropriate caution. The tests are affected by many technical problems, such as missing predictions due to server downtime or outdated template libraries. Despite these difficulties, practical conclusions for automated and manual structure prediction projects can be drawn from the experiments and include:

1. The differences in accuracy between popular prediction services are relatively small. The completeness of the template database remains an important aspect of the quality of the service. Some sophisticated algorithms require substantial effort to update the database thus simple but frequently updated services should always be consulted.
2. Individual threading methods and hybrid methods, utilizing structural information in the scoring function, are probably not more accurate than well-tuned sequence profile comparison methods that ignore structural information. The combination of profile alignment methods with meta prediction approaches that build consensus models (for example as implemented in hhpred) have proven very powerful lately.
3. Meta predictors are clearly superior to simple individual methods. For quite some time, meta predictors are heading the ranking in sensitivity (number of correct prediction) and specificity (reliability of the confidence score).
4. In contrast to popular sequence alignment methods (Blast (Altschul et al. 1990)), structure prediction servers are generally not prepared to deal with multi-domain targets. Division of the sequence into domains and iterative submission of corresponding domain sequences to prediction servers is strongly advised. This is especially important for eukaryotic proteins, since many of them contain multiple structural domains and disordered regions. The disordered regions in proteins can be predicted with dedicated methods and should be removed before submitting the sequence to fold prediction services.
5. The score for hits reported by some meta predictors (for example 3D-Jury (Ginalski et al. 2003)) is sometimes artificially increased if one of the component servers generates many almost identical models. A high score is only significant if several independent servers confirm the fold assignment.
6. For the majority of difficult prediction cases, the confidence scores reported by the servers are below the reliability threshold and the correct models are not always the top ranking ones. Expert users are sometimes able to select the correct predictions using additional knowledge, such as the similarity of function between target and template or the conservation of essential amino acids, short sequence patterns or typical secondary structure motifs. In many cases, the experts have to conduct extensive literature analysis and consider predictions for homologs to guess the right fold. This time-consuming but frequently very fruitful exercise is advised to all predictors, provided that there is enough time to analyze the target of interest.

7. In very difficult cases results of *ab initio* methods or servers using *ab initio* components can be consulted. Difficult targets can be distinguished from others based on low scores of meta predictions. In general, such models have very low quality independent of the applied algorithm, but some biological hints can be gained from fold assignment and subsequent comparison.
8. Models can be improved manually by experts. Detailed structural analysis of the target and the template families to identify core regions and key residues is mandatory in such cases. A model can be improved significantly, if the expert detects a substantial error in the alignment resulting, for instance, from the insertion of an entire domain. In many cases, however, expert improvement remains marginal.
9. It is possible to estimate which parts of the models are likely to be correct and which parts are most likely wrong. The most reliable (consensus) approach is the analysis of alignments obtained from different servers and selection for regions with structurally consistent predictions. Regions where different methods report different alignment to similar templates are likely to be misaligned, or structurally diverged. The quality of models can also be evaluated with quality assessment programs, such as Verify3D (Eisenberg et al. 1997). Unfortunately, the quality of difficult fold recognition models is below the standards of the training sets used to tune the majority of quality assessment methods. Such methods perform better in simple homology modeling experiments.
10. Most on-line protein structure meta predictors are too slow to be used in high throughput annotation projects. For such purpose, it is better to construct in-house meta predictors using several simple, but diverse, independent components. A simple function, which tells by how many components the prediction was confirmed, can be used as reliability score. This is a general suggestion not only for structural annotation, but also for functional annotation, which is routinely conducted with only one method.

6 Conclusions and future outlook

Hopefully, this short survey of the methods currently used for protein structure prediction has left the reader with at least two take-home messages. First, the possibility of obtaining reasonably accurate models of proteins, although not deriving from an understanding of the protein folding process, is within our reach. Second, accurate benchmarking and testing the methods is a key step in proceeding further.

There is one aspect of the problem that has been touched only in passing in this chapter, and this is the ability to evaluate a priori the expected quality of a model. This is not an irrelevant aspect of the field, in fact it might be the most relevant for the end users of the models. A three-dimensional model of a protein structure should come with an attached reliability score, so that a user can extract the information at the correct level of

detail without over interpreting the features of the model. We have seen that some methods are likely to be more reliable than others, for example comparative modeling in the presence of a closely related template, and that some regions, essentially the “core”, are expected to be predicted better than the rest of the structure, but these are still rules of thumb, and, as of today, there is no tool that, given a single model of a protein, can evaluate its accuracy reliably. As we saw in the discussion of the meta server technologies, consensus methods can be rather successful in ranking a set of models for the same protein according to their expected quality, but this does not completely solve our problem. In principle we would like to estimate whether a given model is accurate enough to be used as a guide, say, in drug design or docking experiments, or to assess whether a given residue is involved in substrate recognition or which are the precise boundaries of the protein’s structural domains. Unfortunately, only in the case of comparative model we can try and provide accuracy estimates, essentially relying on the evolutionary distance between the target and template protein(s) (Cozzetto et al. 2007).

It is almost trivial to predict that the ability of linking a model to its expected accuracy and therefore to its possible applications will be the decisive factor in making the methods described in this chapter a useful component of the toolset that computational biology has made available to the life scientists.

The other aspect representing a serious and important challenge in the field is the ability of using the models to predict the molecular function of the target protein. Once again, many efforts for developing methods are ongoing as discussed later in this book. Although the results of the Function Prediction category in CASP have not been very promising (Soro and Tramontano 2005; Pellegrini-Calace et al. 2006; Lopez et al. 2007), we like to think that this is mostly due to the fact that the protein structure and protein function prediction communities do not overlap sufficiently and to the obvious difficulty of evaluating a function prediction in a context such as CASP. Nevertheless, the fact stays that protein structure prediction cannot just provide structural models anymore: the genomic and post-genomic challenges are much more demanding and the field needs to take the task of providing added values to the three-dimensional models very seriously and this will, undoubtedly, be the case in the near future.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410
- Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science*: 223–230
- Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D (2005) Free modeling with Rosetta in CASP6. *Proteins* 61(Suppl 7): 128–134
- Bruccoleri RE, Karplus M (1990) Conformational sampling using high-temperature molecular dynamics. *Biopolymers* 29: 1847–1862
- Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001) Structure prediction meta server. *Bioinformatics* 17: 750–751

- Burgess AW, Scheraga HA (1975) Assessment of some problems associated with prediction of the three-dimensional structure of a protein from its amino-acid sequence. *Proc Natl Acad Sci USA* 72: 1221–1225
- Chinae G, Padron G, Hooft R, Sander C, Vriend G (1995) The use of position-specific rotamers in model building by homology. *Proteins* 23: 415–421
- Chothia C, Lesk A (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5: 823–826
- Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A (2007) Assessment of predictions in the model quality assessment category. *Proteins* 69(Suppl 8): 175–183
- Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics* 14: 892–893
- Eisenberg D, Bowie JU, Luthy R, Choe S (1992) Three-dimensional profiles for analysing protein sequence-structure relationships. *Faraday Discuss* 93: 25–34
- Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol* 277: 396–404
- Eyrich VA, Kryshtafovych A, Milostan M, Fidelis K (2005) System for accepting server predictions in CASP6. *Proteins* 61(Suppl 7): 24–26
- Fariselli P, Casadio R (2001) Prediction of disulfide connectivity in proteins. *Bioinformatics* 17: 957–964
- Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg M (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins* 37(Suppl 3): 209–217
- Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL Jr (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins* 45(Suppl 5): 171–183
- Fischer D, Rychlewski L (2003) The 2002 Olympic Games of protein structure prediction. *Protein Eng* 16: 157–160
- Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR, Elofsson A (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins* 53(Suppl 6): 503–516
- Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. *Protein Sci* 9: 1753–1773
- Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19: 1015–1018
- Godzik A, Skolnick J (1992) Sequence-structure matching in globular proteins: application to super-secondary and tertiary structure determination. *Proc Nat Acad Sci USA* 89: 12098–12102
- Greer J (1991) Comparative modeling of homologous proteins. *Methods Enzymol* 202: 239–252
- Holm L, Sander C (1992) Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins* 14: 213–223
- Jacoboni I, Martelli PL, Fariselli P, Compiani M, Casadio R (2000) Predictions of protein segments with the same amino acid sequence and different secondary structure: a benchmark for predictive methods. *Proteins* 41: 535–544
- Jones DT (1997) Successful *ab initio* prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins* 29(Suppl 1): 185–191
- Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292: 195–202
- Jones DT (2001) Predicting novel protein folds by using FRAGFOLD. *Proteins Suppl* 5: 127–132
- Jones DT, Thornton JM (1996) Potential energy functions for threading. *Curr Opin Struct Biol* 6: 210–216
- Keller D, Shibata M, Marcus E, Ornstein R, Rein R (1995) Finding the global minimum: a fuzzy end elimination implementation. *Protein Eng* 8: 893–904

- Koh IY, Eylich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Grana O, Pazos F, Valencia A, Sali A, Rost B (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res* 31: 3311–3315
- Kolinski A, Skolnick J (1994a) Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* 18: 338–352
- Kolinski A, Skolnick J (1994b) Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins* 18: 353–366
- Kopp J, Schwede T (2006) The SWISS-MODEL repository: new features and functionalities. *Nucleic Acids Res* 34: D315–D318
- Levinthal C (1968) Are there pathways for protein folding? *J de Chimie Phys PCB* 65: 44–45
- Levitt M, Warshel A (1975) Computer simulation of protein folding. *Nature* 253: 694–698
- Lopez G, Rojas A, Tress M, Valencia A (2007) Assessment of predictions submitted for the CASP7 function prediction category. *Proteins* 69(Suppl 8): 165–174
- Martelli PL, Fariselli P, Casadio R (2004) Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. *Proteomics* 4: 1665–1671
- Martelli PL, Fariselli P, Malaguti L, Casadio R (2002) Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. *Protein Eng* 15: 951–953
- Moult J (1996) The current state of the art in protein structure prediction. *Curr Opin Biotech* 7: 422–427
- Moult J, Unger R (1991) An analysis of protein folding pathways. *Biochemistry* 30: 3816–3824
- Pellegrini-Calace M, Soro S, Tramontano A (2006) Revisiting the prediction of protein function at CASP6. *FEBS J* 273: 2977–2983
- Rychlewski L, Fischer D (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci* 14: 240–245
- Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234: 779–815
- Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268: 209–225
- Sippl MJ (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 5: 229–235
- Skolnick J, Kolinski A, Kihara D, Betancourt M, Rotkiewicz P, Boniecki M (2001) *Ab initio* protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins* 45(Suppl 5): 149–156
- Soro S, Tramontano A (2005) The prediction of protein function at CASP6. *Proteins* 61(Suppl 7): 201–213
- Tramontano A (1995) The architecture of loops in proteins. In: Villar HO (ed) *Advances in computational biology*. JAI Press, Greenwich, pp 239–259
- Tramontano A (2006) *Protein Structure Prediction*. Wiley Inc., Weinheim, German
- Tramontano A, Chothia C, Lesk AM (1989) Structural determinants of the conformations of medium-sized loops in proteins. *Proteins* 6: 382–394
- Tsai CH, Chan CH, Chen BJ, Kao CY, Liu HL, Hsu JP (2007) Bioinformatics approaches for disulfide connectivity prediction. *Curr Protein Pept Sci* 8: 243–260
- Ward JJ, McGuffin LJ, Buxton BF, Jones DT (2003) Secondary structure prediction with support vector machines. *Bioinformatics* 19: 1650–1655

CHAPTER 5.2

The state of the art of membrane protein structure prediction: from sequence to 3D structure

R. Casadio¹, P. Fariselli¹, P. L. Martelli¹, A. Pierleoni¹, I. Rossi¹ and G. von Heijne²

¹ Biocomputing Group, Department of Biology, University of Bologna, Bologna, Italy

² Center for Biomembrane Research, Dept of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

1 Why membrane proteins?

Membrane proteins constitute a very large set of yet-to-be characterized proteins mediating all the relevant life-related functions both in prokaryotes and eukaryotes. Estimates are suggesting that in whole genomes the content of this protein type may vary from 10 to 40% of the whole proteome, depending on the organism.

As to present (and this may change on time) we may browse data bases and find out that the number of protein sequences is $\sim 6,000,000$ (in the Non Redundant data base [<http://www.ncbi.nlm.nih.gov/>]); that sequences annotated as “membrane protein” are 45,281 in Swiss-Prot (<http://expasy.org/sprot/> where the annotation is manually curated), and that atomic structures of membrane proteins are 281 in the Protein Data Bank (<http://www.rcsb.org/pdb/>, PDB). And then the question is how many membrane proteins are out there, to be still annotated and characterized among the many millions of putative protein sequences in the genomes that are in the process of being/will be sequenced? Indeed we do not yet know how many species are presently present in our planet.

We may consider a rough average of 30% of membrane proteins per genome (as derived from sequence similarity search) and end up with an approximate number of about 2,000,000 membrane proteins in the data bases. We can then easily evaluate that less than $\sim 0.6\%$ of membrane proteins are annotated and that $\sim 0.001\%$ of all the membrane protein sequences are known with atomic resolution.

Why is this the case, when so many globular proteins are known with atomic resolution? Membrane proteins are difficult to study. They are inserted into lipid bilayers surrounding the cell and its sub-compartments, and expose to the polar outer

Corresponding author: Rita Casadio, Biocomputing Group, Department of Biology, University of Bologna, Bologna, Italy (e-mail: casadio@biocomp.unibo.it)

and inner environments portions of different sizes. When isolated from membranes, membrane proteins are generally less stable than globular ones. It is therefore difficult to purify them in the native, functional form, and more difficult to crystallize them. Historically it is worth mentioning that Deisenhofer, Huber and Michel were awarded with the Nobel Prize in Chemistry in 1988 for having solved the X-ray structure of the photosynthetic reaction center from *Rhodospseudomonas viridis* (www.nobel.se), the first membrane protein to be solved with atomic resolution (Deisenhofer et al. 1984). Crystallization of this type of proteins is yet a very difficult process, given the fact that they expose two different chemico-physical surfaces to the environment: water- and lipid-like. However the lipid environment constraints the membrane protein stable folding: it is indeed evident from the structures that have been deposited so far that only two structural organizations are present in nature: all-alpha and beta-barrel membrane proteins (Fig. 1). Indeed most of membrane proteins in the PDB (67%) consist of bundles of transmembrane helices with different tilting with respect to the membrane plane, when known with enough details to be realistic, and to each other. The relative structural organization of the transmembrane helices is also very much dependent on the protein function. The all-alpha membrane proteins can be classified in relation to the number of membrane spanning helices, and a major grouping discriminates monotopic versus polytopic membrane proteins (Fig. 1).

Some of the membrane proteins are located in the outer membrane of mitochondria and chloroplasts, or in the outer membrane of gram-negative bacteria; in this case they

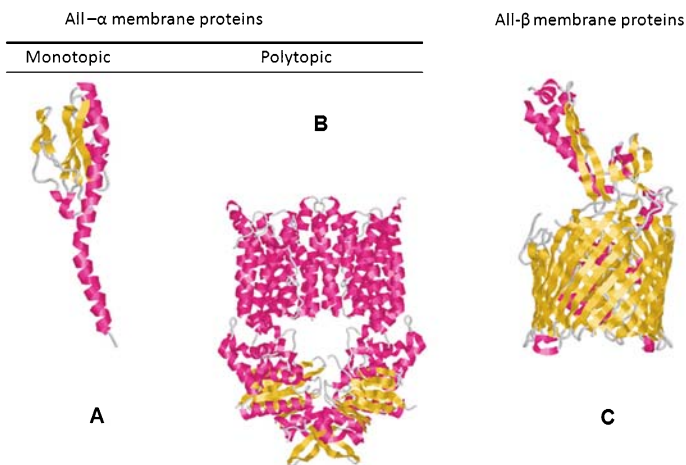


Fig. 1 Structural types of membrane proteins. **A:** Phosphorylated Pilin from *Neisseria meningitidis* (PDB code: 2pil); **B:** Putative metal-chelate type ABC transporter from *Haemophilus influenzae* (PDB code:2nq2); **C:** Colicin I receptor Cir from *Escherichia coli* in complex with receptor binding domain of Colicin Ia (PDB code:2hdi)

are endowed with a very well conserved architecture known as transmembrane beta barrel, where the only variables are the even number of beta strands in the barrel (from 8 to 22) and its plane of shear (when known) with respect to the membrane plane (Fig. 1).

Thanks to several experimental efforts, we also know that in all-beta outer membrane proteins in Gram-negative bacteria a signal peptide is present in the protein precursor, and that this signal peptide may or may not be present in membrane proteins of other organisms. The location of the N- and C-termini as well as of the internal loops of a membrane protein relative to the lipid bilayer (cytoplasmic or extra-cytoplasmic) can also be experimentally determined, either by using so-called reporter fusions or by various kinds of covalent modifications targeting residues introduced into the protein by site-directed mutagenesis. Detailed topology models have been produced for many membrane proteins by such techniques (<http://expasy.org/sprot/>). Recently, the cytoplasmic/extra-cytoplasmic location of the C termini of the entire all-alpha membrane proteomes in *Saccharomyces cerevisiae* and *Escherichia coli* were mapped experimentally using reporter fusions (Daley et al. 2004; Kim et al. 2005); this data has been used in the benchmarking study reported below.

2 Many functions

Membrane proteins constitute an important part of the cell proteome. They perform basic functions, fundamental for the cell life, including cell signaling, energy conservation and transformation, ion exchange and many others. It is well established that membrane proteins take part into many different cell functions, in that they mediate/regulate most of the cell-environment interactions as well as trafficking among different subcellular compartments. Their functions are related to their structural architecture. All-alpha membrane proteins are typically: G Protein-Coupled Receptors, Channels, Proteases, Transporters, Antiporters, ATP-ases of different types, Respiratory proteins, and Oxidases. In turn all-beta membrane proteins may act as enzymes and/or porins, kind of molecular sieves that are involved in transmembrane transport (<http://blanco.biomol.uci.edu>). Interestingly enough, in pathogenic bacteria all-beta membrane proteins are involved in pathogenicity, and can therefore be quite useful drug targets (Casadio et al. 2003a).

3 Bioinformatics and membrane proteins: is it feasible to predict the 3D structure of a membrane protein?

In the “omic” era hundreds of genomes are available for protein sequence analysis, and we may estimate that some 30% of all the sequences are of membrane proteins. Differently from globular proteins, a three-dimensional model for membrane proteins

can hardly be computed starting from the sequence. Why is this? What can we really compute and with what reliability? Can we build models of membrane proteins based on threading techniques?

These issues are addressed with approaches that may be different by those generally adopted when globular proteins are predicted, and their solution often requires an expert-driven methodology. The question is then how many methods do we have that may be integrated for getting a successful prediction of a membrane protein?

Another major problem in large-sequence projects is the annotation of those genes which have no counterpart in the database of presently known sequences with a given function. Can we then contribute to the annotation process with predictive methods? And again: how many membrane proteins are present in the human genome?

These and other questions may be answered in the post-genomic era by taking advantage of all the theoretical efforts aiming at developing tools based on our present knowledge that are capable of extracting selected structural/functional features from known sequences/structures and of computing the likelihood of their presence in never-seen before sequences/structures. In the following we will highlight when and how it is possible to recognize a membrane protein sequence and when it is possible from the sequence to compute its folded 3D structure (Casadio et al. 2003b; Elofsson and von Heijne 2007; Punta et al. 2007).

4 Predicting the topology of membrane proteins

Most of the computational methods presently available allow predicting two basic features of membrane proteins: topography (the location of transmembrane domains along the protein chain) and topology (the location of the N- and C-termini with respect to lipid membrane). Topological models are sufficient in many instances to design experiments in order to prove to a certain extent (or correct) the location of the inner and outer loops with respect to the membrane, and concomitantly the number of transmembrane segments.

Methods presently available are mainly based on machine learning (Fig. 2). All the machine learning methods described so far for the prediction of transmembrane topology are based on neural networks (NNs), hidden Markov models (HMMs), support vector machines (SVM) or their ensemble. Their implementation requires: i) the selection of a training set with little homology to the testing set; ii) a training phase where the variable parameters of the algorithms are adjusted to fixed values, according to a learning procedure; iii) a testing phase, when the system is scored according to statistical indexes (and for comparison among the different methods, when they are described in the literature). Either a jackknife or a cross validation procedure are adopted in order to perform the training and testing simultaneously; also, in some cases, a blind test is adopted to further validate the method. It is very important to neatly

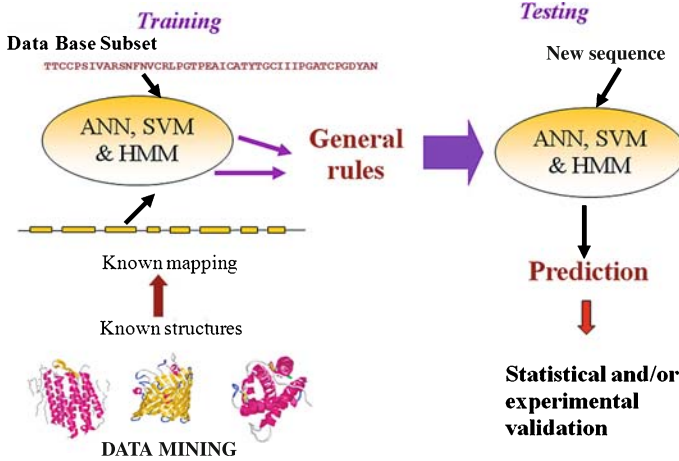


Fig. 2 Tools out of machine learning approaches: Artificial Neural Networks (ANNs), Hidden Markov Models (HMMs), Support Vector Machines (SVMs). During the training phase known examples are presented and the general rules of the association between the input and the output are stored in the training parameters. During the testing phase, new sequences are predicted. The statistical evaluation consists in the comparison between the predicted and expected features

separate the training and the testing sets in order to evaluate the prediction performances on examples that are not similar to data used for setting the trainable parameters. It should be emphasized that feed-forward neural networks are able to capture the information contained in local contexts of the sequence, whereas hidden Markov models are able to cast the global features of proteins. With a neural network-based approach, each residue in the sequence is predicted to be or not be in transmembrane state considering a window (usually 17–25 residues long) centered on the residue to be predicted. The contribution of each residue in the window is weighted by a specific trainable parameter and the contributions are then combined in a non linear way. HMMs are able to describe the basic features of membrane proteins. These features can be stated as: i) there are transmembrane segments that are connected by loops; ii) the loops face alternatively the inner and the outer side of the membrane; iii) transmembrane segments and loops are endowed with minimum and/or maximum lengths. These features are described by a set of trainable parameters known as transition probabilities. Moreover HMMs, by means of the trainable emission probabilities are able to cast the different residue compositions in the different portions of the proteins. SVMs are known to behave very similar to NNs; however their discrimination capabilities are routinely superior (Jones 2007).

Topological models can be computed after predicting the protein membrane topography, with specific rules. The most popular for the all-alpha membrane proteins is the positive-inside rule (Nilsson et al. 2004). For beta-barrel outer membrane

proteins, in Gram negative bacteria, longer loops are generally exposed to the external phase (Martelli et al. 2005). Neural network based methods can predict the topography; the topology is then obtained with specifically implemented rules and dynamic programming (Fariselli et al. 2003a). In the case of HMM the topological model is derived from the prediction, according to an optimization algorithm or again after dynamic programming (Fariselli et al. 2005). All the machine learning based methods improve the predictive performance when evolutionary information in the form of sequence profile, computed from multiple sequence alignment, is used as input. Also a dynamic programming filter for locating the transmembrane segments can increase the scoring of all the methods (Punta et al. 2007).

5 How many methods to predict membrane protein topology?

Recently we revisited the problem of membrane protein prediction by implementing, testing and comparing results from the top scoring predictors of membrane protein topology. These methods are essentially based on statistical propensity scales, neural networks and hidden Markov models. How well do they perform? This question is particularly relevant when we consider genome filtering. For finding a solution to this problem the rate of false positives and false negative is an important issue. Furthermore the long-standing-around propensity scale of Kyte and Doolittle (KD) was revisited by us including a version taking as input sequence profile and it was adopted as a baseline predictor in our experiments (with these methods only topography can be predicted; see Table 1). For beta-barrel membrane proteins we have shown that HMM based predictors are superior to neural network-based ones, being endowed with a more stringent selectivity (Bagos et al. 2005).

5.1 From theory to practice

A new integrated and browseable server that contains also precomputed predictions of sequences contained in the UNIPROT dataset (22 Sept. 2004) is now available. The new environment/web-server/DAS-server is called PONGO (Amico et al. 2006). It is based on a relational database and it can be queried through the web interface available at <http://pongo.biocomp.unibo.it> (or following a link from www.biocomp.unibo.it), or through the DAS client at EBI. Due to this implementation, any protein chain can be filtered by different predictors. The predicted features highlight whether the protein is or is not endowed with a signal peptide, whether the sequence is or is not a membrane protein, and in this case its putative topology as computed by six different predictors. The predictive methods that have been selected and implemented have been previously described in literature and are considered top scoring for their performance. This

Table 1 Performance on 121 high-resolved membrane protein chains from PDB

	$Q_{\text{topography}}$	Q_{topology}
<i>Based on single-sequence</i>		
Kyte-Doolittle	82/121 (68%)	–
TMHMM	88/121 (73%)	67/121 (55%)
TMHMMdomfix	87/121 (72%)	74/121 (61%)
PHOBIUS	96/121 (79%)	75/121 (62%)
<i>Based on multiple sequence alignments</i>		
PSI-Kyte-Doolittle	97/121 (80%)	–
ENSEMBLE 1.0	105/121 (87%)	92/121 (76%)
ENSEMBLE 2.0	105/121 (87%)	95/121 (79%)
MEMSAT	93/121 (77%)	90/121 (74%)
PRODIV	99/121 (82%)	93/121 (77%)

$Q_{\text{topography}}$: Number of proteins predicted with the correct number of transmembrane helices and correct position of transmembrane helices along the sequence (allowing that predicted and expected position of the helix are at least 50% overlapping). Q_{topology} : Number of proteins predicted with correct Topography and correct orientation with respect to the membrane plane

procedure allows the user to make also comparison among different predictors at the same time and at the same web site, and to assess whether the expected results are or are not in agreement with its own experimental finding. Alternatively different predictions, especially when in agreement, may enforce the expectation that a given chain is a membrane protein and in this case the putative topology may help in designing experiments in order to validate (or not) the number of transmembrane helices and the location of the N and C termini of the protein with respect to the plane of the membrane. This may be particularly useful when the chain has no homologous counterpart in the data base of sequences and may help in highlighting also its function. We then use the six state of the art predictors in order to identify the most probable integral transmembrane proteins. In this way the intersection of the prediction would be the most reliable set, while the union furnishes a very high level of integral membrane protein coverage when proteomes are filtered for detecting integral membrane proteins.

The list of the programs that have been wrapped and run locally for the transmembrane annotations comprises:

1. TMHMM2.0 which is the predictor of transmembrane helices in proteins based on hidden Markov models developed by Krogh et al. (2001). TMHMM2.0 has the great advantage of being very fast, being based on only single sequence information.
2. MEMSAT is the new version of the predictor of transmembrane helices in proteins developed by Jones et al. (1994). This version takes advantage of the evolutionary information derived by multiple sequence alignment.

3. ENSEMBLE is the predictor of transmembrane helices in proteins developed by Martelli et al. (2003). It is an ensemble of two hidden Markov models and one neural network. ENSEMBLE takes advantage of the evolutionary information derived by multiple sequence alignment.
4. PRODIV_TMhMM_0.91 is a predictor of transmembrane helices in proteins developed by Viklund and Elofsson (2004) which uses a hidden Markov model similar to TMhMM, but exploits the evolutionary information derived by multiple sequence alignment.
5. TMhMM DOMFIX is a predictor of transmembrane helices where according to the authors, topology predictions was substantially improved by constraining part of the protein to a given in/out location relative to the membrane using experimental data or other information (Bernsel and von Heijne, 2005). The predictor was trained with 367 SMART domains and it is represented by a profile HMM.
6. ENSEMBLE2.0, which is similar to ENSEMBLE, but it is endowed with a new direct method to assign topology (unpublished) that overcomes the prediction of topography by our consensus method, previously described (Martelli et al. 2003). The algorithm used to assign the topology is the Posterior-Viterbi (Fariselli et al. 2005a). The models are the amphipatic and hydrophilic HMMs described in Martelli et al. (2003).

Since it is quite common that signal peptides are mispredicted as transmembrane helices due to their hydrophobicity, before running any of the predictors described above, we test the presence of the signal peptide using SPEP, a signal peptide predictor (Fariselli et al. 2003b). In case of a positive answer, we cut the corresponding predicted segment and we process the remainder of the sequence using the six transmembrane predictors (Fig. 3). Furthermore in Bologna we embedded, in a TRAMPLE environment, a version of the Kyte-Doolittle predictor supplemented with a sequence profile input and both a neural network, and a HMM based predictor of all-beta membrane proteins (Fariselli et al. 2005b; www.biocomp.unibo.it).

6 Benchmarking the predictors of transmembrane topology

6.1 Testing on membrane proteins of known structure and topology

The predictors are tested by predicting the topography and topology of membrane proteins known with atomic resolution. As a matter of fact while performing the test, only the in-house implemented/trained predictors are predicting never-seen before proteins. This is obtained by adopting a leave-one-out procedure, in which the training test does not contain the sequence that is predicted. We also compared the overall

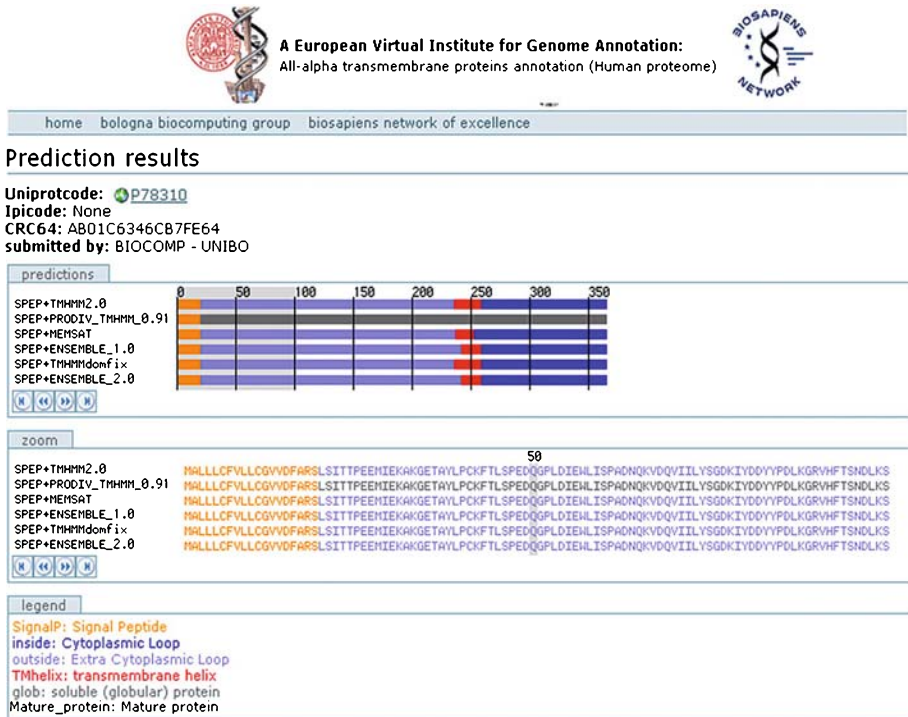


Fig. 3 Prediction of Human adenovirus receptor precursor with PONGO (Amico et al. 2006)

predictions to those obtained on the same test set with Phobius (Käll et al. 2007), a well performing method specifically suited to discriminate signal peptides from N-terminal located membrane spanning helices. The results are shown in Table 1, and listed depending on the input (single sequence versus sequence profile). We may conclude that inputs based on sequence profile are scoring higher than methods based on single sequence.

6.2 Topological experimental data

Even if the three dimensional structure of membrane proteins is very difficult to be experimentally resolved as the paucity of structural data testifies, techniques based on fusion with a reporter protein allow to determine the location of the C-terminus with respect to the membrane plane. Two model organisms were taken into consideration for large scale topology determination in the von Heijne's lab: *E. coli* and *S. cerevisiae*. In the case of the first organism, alkaline phosphatase (PhoA) and green fluorescent protein (GFP) were fused in parallel to the C-terminus of candidate inner membrane proteins.

Since GFP is fluorescent only in the cytoplasm and PhoA is active only in the periplasm, measurements on the activity of the two proteins lead to the determination of the location of the C-terminal region (Daley et al. 2005). In the case of Yeast the reporter genes are the histidinol dehydrogenase (His4C), which is active only in the cytoplasm, and the invertase 2 (Suc2), which contains eight N-glycosylation sites that are glycosylated only if the reporter is translocated to the lumen of the endoplasmic reticulum and then to the extracytoplasmic space. Activity and glycosylation assays on the fused membrane proteins are then able to determine the location of the C-terminus with respect to the membrane plane (Kim et al. 2006).

6.3 Validation towards experimental data

In the following test procedure predictions are compared to experimental data, concerning the location of the C terminus in the whole proteome of *E. coli* and *S. cerevisiae*.

As a general comment, one has to keep in mind that this test is not focused on correctly predicting the topological model of the protein sequence, but only the location of the C-terminus with respect to the membrane plane, and that the expected results have been produced experimentally as described in the previous section. We have therefore 613 chains from *E. coli*, and 505 chains from *S. cerevisiae*. For each organism, chains are divided in relation to their C-terminus being IN (inner) or OUT (outer) with respect to the plane of the membrane. For each predictor, identified from its official name, as explained above, all the correct and wrong predictions are listed, so the results for the different tables are to be considered as confusion matrices. The results of this experiment indicate that the performances of the predictors implemented in the Bologna DAS server for the annotation of membrane proteins can provide reliable predictions, when scored in a blind way, against sets of membrane proteins whose C-terminus position was experimentally detected.

From this set of data it appears that for the specific task at hand (the prediction of the location of C-terminus with respect to the membrane plane) the best performing predictors are ENSEMBLE2.0 (HMMM2 and HMM1) and PRODIV, with scores as high as 86.9%, 85.5%, 84.5% for the *E. coli* protein membrane set, and 83.4%, 82.4% and 83% for the one from *S. cerevisiae*, respectively.

Furthermore for each predictor the overall accuracy (Q), accuracy for the class (Q(IN); Q(OUT)), probability of correct predictions for the class (P(IN); P(OUT)) and Matthew's correlation coefficient are listed. The first set of data indicates predictions for 613 membrane proteins from *E. coli* with 480 sequences experimentally detected with C-terminus in, and 133 experimentally detected with C-terminus out. The second set of data indicates predictions for 505 membrane proteins from *S. cerevisiae* with 419 sequences experimentally detected with C-terminus in, and 86 experimentally detected with C-terminus out (Tables 2 and 3).

Table 2 Performance in predicting the C-terminal location in the *E. coli* membrane protein set

Method	C-ter IN			C-ter OUT			Global indexes	
	Pred IN	Pred OUT	Pred GLOB	Pred IN	Pred OUT	Pred GLOB	Overall accuracy	Correlation
TMHMM 2.0	360	119	1	36	97	0	74.6%	0.41
TMHMMdomfix	375	100	5	34	99	0	77.3%	0.46
MEMSAT	411	64	5	38	94	1	82.4%	0.53
PRODIV	418	52	10	33	100	0	84.5%	0.58
ENSEMBLE 1.0	391	88	1	40	93	0	78.8%	0.46
ENSEMBLE 2.0	434	45	1	34	99	0	86.9%	0.63

Performances are computed on 613 chains from *E. coli* whose topology has been experimentally determined by means of reporter fusions (Daley et al. 2005). The set contains 480 sequences with cytoplasmic C-terminal (IN) and 133 with extracytoplasmic C-terminal (OUT)

Table 3 Performance in predicting the C-terminal location in the *Saccharomyces cerevisiae* membrane protein set

Method	C-ter IN			C-ter OUT			Global indexes	
	Pred IN	Pred OUT	Pred GLOB	Pred IN	Pred OUT	Pred GLOB	Overall accuracy	Correlation
TMHMM 2.0	286	125	8	33	51	2	66.7%	0.22
TMHMMdomfix	275	136	8	34	50	2	64.4%	0.18
MEMSAT	328	84	7	30	52	4	75.2%	0.32
PRODIV	363	36	20	26	56	4	83.0%	0.47
ENSEMBLE 1.0	343	73	3	37	48	1	77.4%	0.33
ENSEMBLE 2.0	374	42	3	33	48	5	83.4%	0.43

Performances are computed on 505 chains whose topology has been experimentally determined by means of reporter fusions (Kim et al. 2006). The set contains 419 sequences with cytoplasmic C-terminal (IN) and 86 with extracytoplasmic C-terminal (OUT)

7 How many membrane proteins in the Human genome?

Methods for the prediction of membrane proteins can be used to find membrane proteins encoded by a genome. To this aim, the evaluation of the false positive and false negative rates is crucial. Due to their parametrization, different methods estimate different contents of membrane proteins.

We annotated with three methods the 33,860 proteins of the human genome, as reported in version 35 of Ensembl. Signal peptides were predicted with SPEP and cleaved before the prediction of transmembrane topology. The estimates range from 19.6% of TMHMM, corresponding to 6639 proteins, to 32.3% of MEM-SAT, corresponding to 10,945 proteins. ENSEMBLE predicts 7044 membrane proteins (20.8%). Considering the false positive and the false negative rates that for ENSEMBLE are both

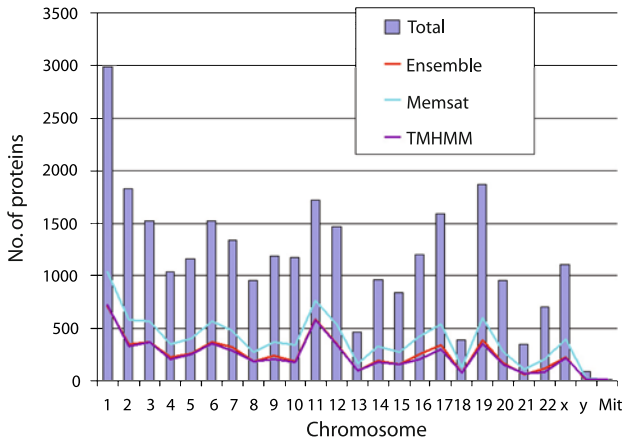


Fig. 4 Distribution of predicted TM proteins among the chromosomes. Human protein sequences are from release 35 of Ensembl

lower than to 4%, a reliable estimate of the content of membrane proteins in the Human genome ranges from 20 to 24%.

The comparison of the predictions shows an agreement of the three methods for 5752 proteins (17%). These figures highlight a set of chains predicted as membrane proteins by all the methods and a smaller portion of proteins with more blurred predictions (Fig. 4).

8 Membrane proteins and genetic diseases: PhD-SNP at work

Single Nucleotide Polymorphisms (SNPs) are the most frequent type of genetic variation in human population (Collins et al. 1998). Great interest is focused on non-synonymous coding SNPs (nsSNPs) that are responsible of protein single point mutations; these mutations occurring in coding regions may have a large effect on gene functionality. nsSNPs can be neutral or disease associated (Ng and Henikoff 2002; Capriotti et al. 2006). The question is therefore whether we know enough SNPs in membrane proteins to make it interesting to model topology in relation to the position of the mutation along the sequence.

A recent statistics of ours indicates that out of 1308 well annotated human proteins listed in Swiss Prot with disease-related variants, 393 are membrane proteins with some 5358 documented mutations. Noteworthy the most frequent disease-related mutations in membrane proteins are the G/R, L/P, R/W, Q, C, and P/L substitutions.

As an example of a relevant outcome of our analysis, we show the 3D/topological model of the Erythrocyte Band-3 anion transport protein (Fig. 5). This protein has a

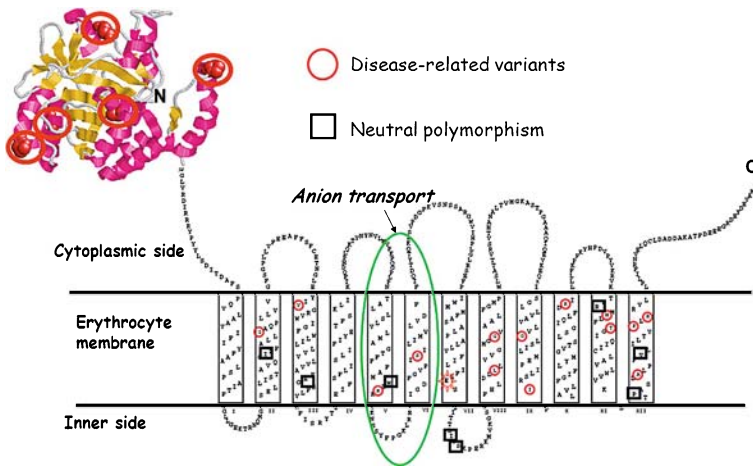


Fig. 5 Localization of SNPs in the Erythrocyte Band-3 anion transport protein (Gene map locus 17q21-q22). The atomic structure of the protein N-terminal domain is available (PDB code 1HYN). The transmembrane portion is predicted with ENSEMBLE (Martelli et al. 2003). SNPs are listed in the SwissProt entry (B3AT_HUMAN) and their correlation to diseases derives from OMIM (www.ncbi.nlm.nih.gov/omim/). 19 SNPs are related to diseases (familial distal renal tubular acidosis and hereditary spherocytosis), while the number of neutral polymorphisms is equal to 10. The green ellipse highlights putative helices involved in the anion transport, as experimentally derived (www.ncbi.nlm.nih.gov/omim/; references in the protein file)

known 3D structure of the soluble portion and a large transmembrane portion that was modeled with PONGO. The snake-viewer of the membrane embedded portion of the protein shows 12 transmembrane alpha helices where all the disease-related and neutral mutations are also highlighted. The model is consistent with the location of the anion transport channel as deduced from experiments, to which both helix V, VI and VII are contributing.

The present possibility of retrieving a large dataset of annotated SNPs from the Swiss-Prot Database prompted the application of machine learning techniques to predict the insurgence of human diseases due to single point protein mutation starting from the protein sequence (Ramensky et al. 2002). We developed a method based on support vector machines (SVMs) that starting from the protein sequence information and evolutionary information, when available, can predict whether a new phenotype derived from a nsSNP can be related to a genetic disease in humans. The system is based on two different SVMs: one is a SVM-sequence that performs predictions relying on sequence information alone; the other is a SVM-profile performing predictions on profile features when evolutionary information is available. Merging in a unique framework the two SVMs we got a hybrid predictive method.

On a recent dataset (June 2006) of 23,597 single point mutations, 58% of which are disease related, out of 4275 proteins, we show that our hybrid predictor can reach more

Table 4 Performances of PhD-SNP2.0 on neutral and disease-related SNPs in human proteins

	Accuracy	Correlation	Spec (D)	Sens (D)	Spec (N)	Sens (N)
All proteins	78%	0.58	80%	80%	78%	78%
Membrane proteins	80%	0.57	84%	83%	73%	73%

The “All proteins” set consists of 15,266 disease related SNPs (D) and 14,226 neutral polymorphisms (N) derived from 5976 sequences annotated in Swiss Prot (release 54.0). The “Membrane proteins” subset consists of 9094 disease-related SNPs (D) and 5556 neutral polymorphisms (N) out of 1992 membrane proteins annotated in the same release of Swiss Prot. Spec: Specificity is computed as the number of correct predictions over the total number of predictions in the class. Sens: Sensitivity is computed as the number of correct predictions over the total number of proteins belonging to the class

than 74% accuracy (with a correlation coefficient of 48%) in the specific task of predicting whether a single point mutation can be disease related or not. Our method, although based on less information, reaches the same accuracy, with a higher correlation coefficient, of the other web-available predictors implementing different approaches (Ng and Henikoff 2002; Ramensky et al. 2002). Moreover, differently from other methods, ours always gives a prediction (Capriotti et al. 2006).

We design a web server integrating our SVM predicting methods, called Predictor of human Deleterious Single Nucleotide Polymorphisms (PhD-SNP). The server is a user friendly resource that gives the possibility of retrieving predictions *via* e-mail. The submission form is very simple and the user has to paste the query sequence, to select the mutation position and the mutated residue in relative input boxes; furthermore he can choose the predictive method. Best results are obtained when evolutionary information is available and when it is possible to perform predictions using the hybrid predictive method (Capriotti et al. 2006).

The results on how well the system is performing when predicting disease-related SNPs on a set of membrane proteins are shown in Table 4, and compared to the same results on a set of globular proteins.

9 Last but not least: 3D MODELLING of membrane proteins

In principle all the strategies used for predicting the 3D conformation of globular proteins can be adopted to predict the structure of membrane ones. Given a target sequence, when a protein with similar sequence and known structure is available in the PDB, comparative modeling procedures can be successfully used. However, due to the scarcity of membrane proteins solved at atomic resolution, the search for homologous templates can be very unfortunate. For this reason the quest for a suitable template needs to include more information, such as functional features and strategies for remote homology search.

Tools for the prediction of membrane topology can support this task, by constraining the number of transmembrane segments of the suitable templates. This is particularly true in the case of beta-barrel fold, whose architecture is quite well defined when the number of transmembrane beta-strands is known. On the contrary, the conformations of different all-alpha membrane proteins having the same number of transmembrane segments can be largely different and functional features have to be carefully taken into consideration in choosing the templates (Casadio et al. 2006).

When suitable templates have been found, the overlap between the predicted topology of the target and the real transmembrane segments of the template guides the pairwise alignment procedure. A model is then built using standard comparative techniques and on the basis of the structural quality check the alignment can be iteratively refined, taking into consideration more features, such as the hydrophobicity of residues, or the location of functional sites.

Our group has successfully addressed the modeling of 3D structure of some membrane proteins starting from the sequence and validated them with ad hoc wet experiments, including site directed mutagenesis, fluorescence spectroscopy or gene expression. In particular the conformation of mitochondrial Voltage Dependent Anion Channels from different eukaryotic organisms was modeled on the basis of a 16-stranded bacterial porin (Casadio et al. 2002; Aiello et al. 2004). The topography of the target sequence was predicted and then a template with the same number of strands in the barrel was selected from the database of prokaryotic porins known at atomic resolution (Fig. 6). Alignment was done overlapping the predicted topography and a comparative modelling strategy was adopted. The resulting model was able to cast all the known functional features available in the literature. A similar strategy was also adopted for the all-alpha mitochondrial oxoglutarate carrier (Morozzo della Rocca et al. 2005).

For small all-alpha membrane proteins, when no suitable template can be found, a knowledge-based method can be applied to the structure prediction. This method searches for supersecondary structural fragments in a library of atomic-solved transmembrane alpha helices. In a second step the fragments are assembled with a simulated annealing algorithm taking into consideration classical energy terms and a statistical potential derived from an analysis of membrane proteins known at atomic resolution. The membrane potential is added to classical energy terms, and derived from a statistical analysis of a selected database of membrane proteins (Hurwitz et al. 2006).

10 What can currently be done in practice?

Presently we can take advantage of the available information in the data bases and in literature for annotating membrane proteins. We may recognize with a certain likelihood of success if a chain is or is not a membrane protein of either type. We can predict

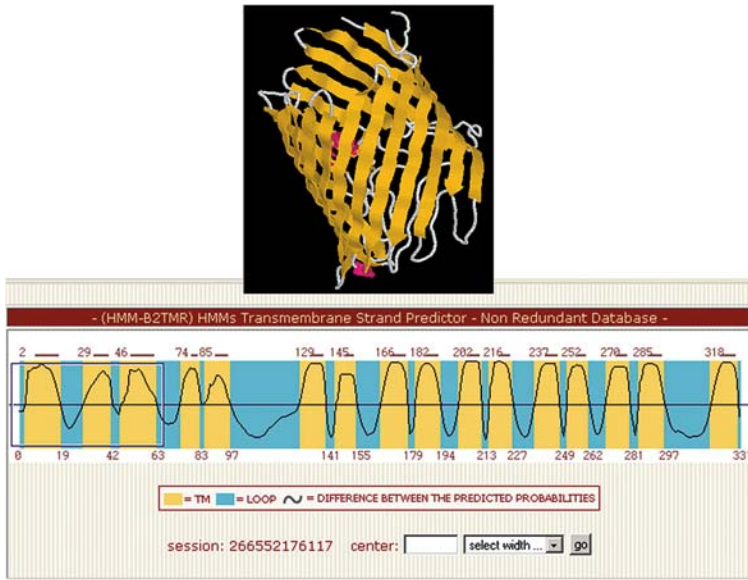


Fig. 6 Prediction of the topology of the OMP32 anion-selective porin from *Delftia acidovorans* (PDB code: 2FGQ). The topology was predicted with HMM-B2TMR (Martelli et al. 2002), using the TRAMPLE environment (<http://gpcr.biocomp.unibo.it/biodec/>) (Fariselli et al. 2005b). The protein was not contained in the training data set. The accuracy of the prediction, after the comparison with the known structure (shown), scores as high as 80%

and discover (after experimental validation) new membrane proteins (Marani et al. 2006). We can model the membrane protein topology, taking advantage of the constraints imposed by the membrane bilayer: we can predict the transmembrane regions of either type of structural architecture and then organize the transmembrane regions with respect to the plane of the membrane. Eventually we can also model them up to their 3D structure, depending on their structural type and we can also investigate in humans how their folding/misfolding is related to diseases.

11 Can we improve?

Always, provided that we are very careful in selecting our training/testing set, in avoiding redundancy and by keeping in mind that most of the problems that we discussed throughout the chapter have not yet been resolved. It is a matter of fact that experiments can always tell us how to improve our tools, and in turn, computations can tell us whether we do the most appropriate experiment.

References

- Aiello R, Messina A, Schiffler B, Benz R, Tasco G, Casadio R, De Pinto V (2004) Functional characterization of a second porin isoform in *Drosophila melanogaster*: DmPorin2 forms voltage-independent cation selective pores. *J Biol Chem* 279: 25364–25373
- Amico M, Finelli M, Rossi I, Zauli A, Elofsson A, Viklund H, von Heijne G, Jones D, Krogh A, Fariselli P, Luigi Martelli P, Casadio R (2006) PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res* 34: W169–W172
- Bagos PG, Liakopoulos TD, Hamodrakas SJ (2005) Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics* 6: 7
- Bell J (2004) Predicting disease using genomics. *Nature* 429: 453–456
- Bernsel A, Von Heijne G (2005) Improved membrane protein topology prediction by domain assignments. *Protein Sci* 14: 1723–1728
- Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22: 2729–2734
- Casadio R, Jacoboni I, Messina A, De Pinto V (2002) A 3D model of the voltage-dependent anion channel (VDAC). *FEBS Lett* 520: 1–7
- Casadio R, Fariselli P, Finocchiaro G, Martelli PL (2003a) Fishing new proteins in the twilight zone of genomes: The test case of outer membrane proteins in *Escherichia coli* K12, *Escherichia coli* O157: H7, and other Gram-negative bacteria. *Protein Sci* 11: 1158–1168
- Casadio R, Fariselli P, Martelli PL (2003b) In silico prediction of the structure of membrane proteins: is it feasible? *Brief Bioinf* 4: 341–348
- Casadio R, Fariselli P, Martelli PL, Tasco G (2006) Thinking the impossible: how to solve the protein folding problem with and without homologous structures and more. *Methods Mol Biol* 350: 305–320
- Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* 8: 1229–1231
- Daley DO, Rapp M, Granseth E, Melén K, Drew D, von Heijne G (2005) Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science* 27: 1321–1323
- Deisenhofer J, Epp O, Miki K, Huber R, Michel H (1984) X-ray structure analysis of a membrane protein complex. Electron density map at 3 Å resolution and a model of the chromophores of the photosynthetic reaction center from *Rhodospseudomonas viridis*. *J Mol Biol* 180: 385–398
- Elofsson A, von Heijne G (2007) Membrane protein structure: prediction versus reality. *Annu Rev Biochem* 76: 125–140
- Fariselli P, Finelli M, Marchignoli D, Martelli PL, Rossi I, Casadio R (2003a) MaxSubSeq: an algorithm for segment-length optimization. The case study of the transmembrane spanning segments. *Bioinformatics* 19: 500–505
- Fariselli P, Finocchiaro G, Casadio R (2003b) SPEFlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics* 19: 2498–2499
- Fariselli P, Martelli PL, Casadio R (2005a) A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics* 6 (Suppl 4): S12
- Fariselli P, Finelli M, Rossi I, Amico M, Zauli A, Martelli PL, Casadio R (2005b) TRAMPLE: the transmembrane protein labelling environment. *Nucl Acids Res* 33: W198–W201
- Hurwitz N, Pellegrini-Calace M, Jones DT (2006) Towards genome-scale structure prediction for transmembrane proteins. *Philos Trans R Soc Lond B Biol Sci* 361: 465–475
- Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 23: 538–544

- Jones DT, Taylor WR, Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 33: 3038–3049
- Käll L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction – the Phobius web server. *Nucleic Acids Res* 35: W429–W432
- Kim H, Melén K, Osterberg M, von Heijne G (2006) A global topology map of the *Saccharomyces cerevisiae* membrane proteome. *Proc Natl Acad Sci USA* 103: 11142–11147
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305: 567–580
- Marani P, Wagner S, Baars L, Genevaux P, de Gier JW, Nilsson I, Casadio R, von Heijne G (2006) New *Escherichia coli* outer membrane proteins identified through prediction and experimental verification. *Protein Sci* 15: 884–889
- Martelli PL, Fariselli P, Krogh A, Casadio R (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics* 18: S46–S53
- Martelli PL, Fariselli P, Casadio R (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics* 19: I205–I211
- Morozzo della Rocca B, Miniero DV, Tasco G, Dolce V, Falconi M, Ludovico A, Cappello AR, Sanchez P, Stipani I, Casadio R, Desideri A, Palmieri F (2005) Substrate-induced conformational changes of the mitochondrial oxoglutarate carrier: a spectroscopic and molecular modelling study. *Mol Membr Biol* 22: 443–452
- Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 12: 436–434
- Nilsson J, Persson B, von Heijne G (2005) Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins* 60: 606–616
- Punta M, Forrest LR, Bigelow H, Kernysky A, Liu J, Rost B (2007) Membrane protein prediction methods. *Methods* 41: 460–474
- Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 30: 3894–8900
- Viklund H, Elofsson A (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci* 13: 1908–1917

SECTION 6

Protein–protein complexes, pathways and networks

CHAPTER 6.1

Computational analysis of metabolic networks

P.-Y. Bourguignon¹, J. van Helden², C. Ouzounis³ and V. Schächter¹

¹CEA – Institut de Genomique – Genoscope, Evry, France

²Laboratory of Genome and Network Bioinformatics, Université Libre de Bruxelles, Bruxelles, Belgique

³Centre for Bioinformatics, King's College, London, UK

1 Introduction

The metabolism of a species results from the joint operation of a large network of biochemical reactions, almost all of which are catalyzed by enzymes encoded in the genome of that species. Metabolic databases such as KEGG (Ogata et al. 1998; Kanehisa et al. 2006) or MetaCyc (Karp et al. 1996; Caspi et al. 2007) contain information about thousands of such reactions together with the compounds they involve. For instance, the KEGG database (as of January 2007) contains 6,580 reactions, and 5355 compounds, linked together by 13,490 substrate-to-reaction and 13,956 reaction-to-product relationships. In total, the KEGG database thus contains 18,515 entities (the metabolites and reactions), and 27,446 links (the substrate-to-reaction and reaction-to-product relationships). Furthermore, genes whose products are known to encode enzymes are linked to the corresponding reactions.

These thousands of reactions are not independent from each other, but they rather form routes between metabolites called metabolic pathways. Along a pathway, some substrates are converted by a first reaction into a set of product metabolites, which are in turn converted into other compounds by the next reactions in the pathway. An example of a metabolic pathway (composed of two sub-pathways, methionine biosynthesis and sulfur reduction) is depicted on Fig. 1, and the aforementioned metabolic databases contain hundreds of such pathways.

Successive reactions along a pathway are adjacent in the sense that one product of the first reaction is a substrate of the second, just as successive compounds are adjacent when some reaction uses the first as a substrate to produce the second. Metabolic pathways like the one depicted on Fig. 1, however, are partial representations of the metabolism focused on a specific biochemical process, from which

Corresponding author: Vincent Schächter, CEA – Institut de Genomique – Genoscope, Evry, France (e-mail: vs@genoscope.cns.fr)

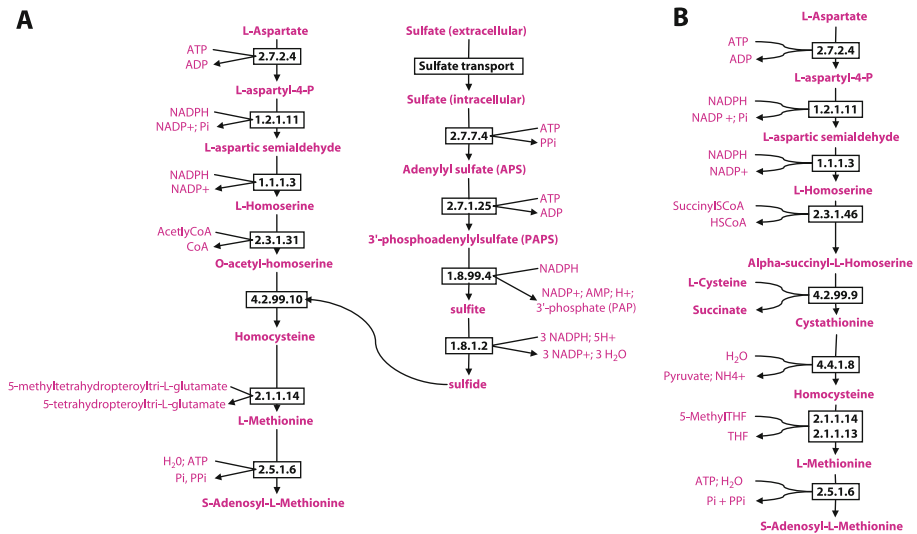


Fig. 1 Example of a metabolic pathway: methionine biosynthesis in the yeast *Saccharomyces cerevisiae* (A) and in the bacterium *Escherichia coli* (B)

links have been deleted for the sake of clarity. As some compounds are involved as substrate or product in as many as hundreds of reactions, it is indeed impractical to represent every possible production or consumption link in a diagram focused on a given process. A far more comprehensive view of the interplay between metabolites and reactions is provided by metabolic networks, in which most (but not necessary all) adjacency relationships are kept. In these comprehensive networks, compounds and reactions are linked by a very large number of possible paths. Computational approaches are thus required to address the complexity of such networks, and understand the relationship between their organization and the biological properties it supports.

Thanks to the availability of comprehensive databases on metabolic pathways, metabolic networks are now amenable to computer analyses. After an introduction to the most popular resources on metabolism, the first section of this chapter describes the typical procedure used to reconstruct the metabolic network of a given organism from the annotation of its genome. The second section introduces classical analyses of the topological properties of metabolic networks, and discusses the biological relevance of the corresponding results. Finally, the last section introduces genome-scale metabolic models as a means of assessing the completeness and quality of a reconstructed metabolic network against functional information, including the organism's known physiological properties.

2 Computational resources on metabolism

A prerequisite for computational analyses of large datasets relies on their availability as well-structured and easily accessible databases. In the case of metabolic networks, a key requirement is for those databases to describe reactions by their chemical equations using a non-redundant set of metabolites, in order to avoid node duplications with its corresponding loss of adjacency relationships. Two databases have acquired reference status within the scientific community: Kyoto Encyclopedia of Genes and Genomes (KEGG, <http://www.genome.jp/kegg>) (Kanehisa 1996), and BioCyc (SRI, <http://biocyc.org>) (Karp et al. 2005). In addition to the web interfaces browsing and querying the data offered by these systems, a variety of software tools dedicated to specific biological analyses have been developed. We provide below an overview of these resources.

2.1 Databases

2.1.1 KEGG

The KEGG database was first released in 1996. Its data model is quite complex, as KEGG is composed of several interconnected databases, each of them being dedicated to a specific realm of molecular biology.

The KEGG Genes database collects gene catalogs of all publicly available complete (and also some partial) genomes. Beyond the storage of raw sequence data, orthology relationships between genes are also computed (and stored in a dedicated database, KEGG SSDB) and made available to the scientific community. The resulting sets of orthologous genes (named KO) are named after the biological functions implemented by the corresponding genes, using a controlled vocabulary.

The controlled vocabulary for describing biological functions is structured in a hierarchical manner, and stored in a dedicated database called KEGG Brite. Brite is not limited to metabolism, but rather aims at providing a comprehensive classification system for biological functions.

Chemical details are provided by the Ligand database, which is actually a set of related databases, each of them describing a biochemical entity (Compounds, Drug, Glycan, Reaction, Rpair, and Enzyme). Reactions are linked to their substrates and products from the Compounds database, and enzymes are identified according to their EC numbers (a standard defined by the IUBMB/IUPAC nomenclature committee), but also linked to the reaction they catalyse. Sets of orthologous genes (KO) encoding an enzyme are referenced by the corresponding Enzyme entry. Considering these databases together allows the computation of direct associations between metabolic genes and the reaction catalysed by their products.

KEGG includes a Pathway database, which gathers manually drawn pathway maps of metabolism, but also of genetic information processing, environmental information

processing, cellular processes and human diseases. Metabolic pathway maps are focused on specific regions of the metabolism (for instance TCA cycle, carbon fixation, . . .) and hierarchically organized according to the corresponding BRITE classification.

In addition to its web interface, KEGG provides a two-level entry-point on its data: an advanced programmer interface (API), the KEGG API, which permits programmatic access to the data using the now standard SOAP technology; and desktop applications that provide advanced graphical representations of various analysis results using the KEGG maps. This suite comprises three software tools, dedicated to browsing and searching hierarchies, analyzing transcriptome and metabolome data in conjunction with KEGG pathways, and drawing of chemical compounds structures.

2.1.2 BioCyc

The BioCyc database collection originated from a database named EcoCyc (Karp et al. 2002), aimed at providing a comprehensive, highly curated resource on the metabolism of *Escherichia coli*. EcoCyc is structured as a Pathway-Genome Database (PGDB), relating information on the *E. coli* genome (chromosome, genes, and gene sequences) to its known metabolic network and set of transporters. This relationship is structured, so that complexes of gene products catalyzing reactions can be represented. The EcoCyc PGDB also contains information about the genetic network (operons, transcription factors and their interactions with DNA binding sites), imported from the RegulonDB database (Salgado et al. 2006). One of its key features is the high level of manual curation. For instance, links to the articles in which the experimental evidence used as a primary source for annotation was published are systematically provided.

PGDB for other species have been built since the release of EcoCyc, forming the BioCyc resource. The PGDBs of 20 species (so-called tier 2 PGDBs) have been manually curated, albeit less systematically than EcoCyc. For 349 other organisms, computationally-derived PGDB without any manual curation have been generated, and are known as tier 3.

An interesting feature of BioCyc is a cross-species pathway database, MetaCyc, which gathers pathways from around 900 organisms. MetaCyc thus includes pathways found in organisms for which no organism-specific PGDB exist. All pathways in MetaCyc have been manually curated, so that MetaCyc and EcoCyc together form the tier 1 of BioCyc, meaning that they reached the highest curation level. MetaCyc can thus be used reliably as a reference pathways database for the reconstruction of the metabolic networks of newly sequenced organisms (see below).

2.1.3 Reactome

Reactome is a distributed infrastructure for curating biological pathways, built upon a knowledge base and rich clients dedicated to pathway authoring and curation. Its focus

is not limited to metabolism, but includes signaling and regulation. For the time being, Reactome is mainly focusing on the description of biological processes in humans (Joshi-Tope et al. 2005), even though other eucaryotic organisms are represented, as well as the bacterium *Escherichia coli*. The data model underlying Reactome integrates several reference databases as primary sources (like Gene Ontology) or cross-references. The Reactome curation model has been quite successful in eliciting collaboration from community experts.

An appealing feature of Reactome resides in its cross-species approach, which is critical since most biological processes happening in humans are actually studied in model organisms, rather than in human cells directly. Pathways are curated by experts in the context of the model organism, and then projected on the human biology using a notion of inferred orthologous events. This mechanism prevents erroneous transfers of knowledge by keeping inferential links in the database. It is noteworthy that even though such transfers of knowledge are often performed across procaryotes for annotation and metabolic networks reconstruction purposes, this feature is absent from all resources on procaryotic metabolism to date.

2.1.4 Querying and exporting data

In order to apply the steadily growing family of software tools dedicated to metabolic network analysis, pathways data must be made accessible in computer-readable format. A first solution is to use database connectors such as those provided for BioCyc (Krummenacker et al. 2005), or software connectors providing the user with a higher-level querying solution on the databases (such as Cyclone for BioCyc (Le Fevré et al. 2007)).

A second solution is to export metabolic networks and pathway description information using a standard exchange format. Two such formats have gained significant recognition, together with import/export tools from and to the main pathways databases: the Systems Biology Markup Language (SBML) (Hucka et al. 2003), and the BioPAX format (Luciano and Stevens 2007). Both are based on XML technology, and result from open collaborative efforts. Even though both languages can encode biological networks, their foci differ. BioPAX aims at providing a robust format for describing biochemical compounds and processes. SBML is more oriented towards the exchange of models in systems biology, and is widely supported by modelling and simulation software tools.

2.2 Reconstruction of metabolic networks

A prerequisite for performing any computational study of the metabolic network of an organism is the reconstruction of that network, from the annotation of its genome together with additional information on its physiology and metabolism. This task may

seem trivial, as it consists in retrieving the set of reactions associated to annotated enzymatic genes. It is unfortunately not, since the initial annotation effort fails at assigning precise functions to around 50% of the enzymes in a typical genome. Reconstructed networks often involve reactions, the catalysers of which seem to be missing from the genome. As a consequence, metabolic networks often exhibit gaps, i.e. some reactions are totally disconnected, or paths that are known or thought to exist between metabolites are interrupted.

2.2.1 From annotated genomes to metabolic networks

One possible approach to enhancing the completeness of networks reconstructed from functional annotations is to retrieve entire pathways from metabolic databases instead of single reactions. But as soon as missing annotations are present, some criteria are necessary to decide which specific form of a pathway should be inferred among those harboring the annotated reactions. More generally, any well-defined notion of metabolic modules may facilitate the reconstruction of coherent networks. Modules should be small enough to occur in several species (so that their identification in one species benefits to newly annotated genomes), but large enough to help retrieve reactions that cannot be directly identified from the annotation.

The best known implementation of this strategy is the Pathologic program, available as part of the Pathway Tools software suite dedicated to querying and computing with BioCyc data (Paley and Karp 2002). Pathologic identifies pathways that are most likely to occur in a given species using the MetaCyc reference pathways database together with the annotation of the species. Networks reconstructed using Pathologic are thus assemblies of metabolic pathways identified and curated in different organisms.

2.2.2 Filling gaps

A corollary of the incorporation of reactions that were not annotated in the genome is the presence of gene gaps in pathways. Searching for these missing genes is one way of completing the annotation. A variety of approaches have been proposed, based on sequence analysis, but also on gene expression or mutant phenotype data analysis. Most methods rely on a combination of different types of evidence supporting the presence of reactions in the metabolic network, together with genome sequence or expression profile analysis. For instance, the Pathway Tools software suite includes Pathway Hole Filler (Green and Karp 2004), which generate gene candidates for filling holes in pathways. One can also exploit the fact that genes involved in a linear pathway are usually overexpressed in similar proportions when the pathway is activated, so that searching for genes, whose expression profiles are similar expression to those of the genes known to be involved in the remainder of the pathway, can shed light on candidate genes to fill the holes (Kharchenko et al. 2004).

Applying this procedure to a newly sequenced organism unavoidably results in an incomplete metabolic network. In the case of reconstructions performed using Pathologic, novel pathways that are absent from the reference pathways database used would obviously be missing from the reconstructed network. In other words, procedures such as the one performed by Pathologic are limited by the current knowledge on metabolism, as well as by our ability to predict genes enzymatic functions based on their sequence.

3 Basic notions of graph theory

In mathematical terms, a metabolic network is a graph: the entities among which adjacency is defined (metabolites and reactions) are represented by nodes, and edges connect adjacent nodes¹. Graph theory has developed a wealth of concepts and algorithms for studying graph topology: degrees of the nodes, shortest-path lengths distance between nodes, or graph diameters are only a sample of the most classical topological properties that can be computed in a graph, and distributions thereof over specific ensembles of graphs have been investigated. In 2000, in an influential paper, Jeong et al. (2000) applied this type of analysis to a metabolic network representing all the reactions catalyzed in 40 organisms, as well as their substrates and products. This article provoked a strong enthusiasm in the bioinformatics community, and gave rise to a fertile field of publications, where the same concepts were transposed to different types of biological networks: metabolism (Fell and Wagner 2000; Jeong et al. 2000; Ravasz et al. 2002; Ravasz and Barabasi 2003), protein interactions (Jeong et al. 2001), transcriptional regulation (Potapov et al. 2005). Some interesting topological properties seemed recurrent in biological networks, among which the three most popular are probably the power-law distribution of degree, the small world property, and the scale-freeness. Sometimes perceived as intrinsic properties of biological networks, their interpretations can be questioned. This motivates the following discussion of both pitfalls and strengths of some widespread ideas about network topology.

3.1 Metabolic networks as bipartite graphs

When metabolic networks represent both metabolites and reactions, only entities of different types (metabolites and reactions) may be connected. This feature defines a class of graphs known as bipartite graphs; furthermore, a metabolite m_i may be related to a reaction r_j as a substrate or as a product, so that metabolic networks are directed bipartite graphs. Reversible reactions are formally encoded as two irreversible reactions in the graph, corresponding to both possible directions.

¹ The graph structure of metabolic networks is actually slightly richer than that, as we shall see later.

A bipartite graph can be projected into component subgraphs, where only one type of entities is represented. In the projected subgraphs, two nodes (for instance two reactions) are connected by a directed edge if they can be connected by two consecutive edges in the bipartite subgraph (for instance if the product of a reaction is the substrate of the other). Metabolic networks can thus be projected into metabolite–metabolite or reaction–reaction graphs.

In either of these graphical representations of metabolic networks, some basic questions may be asked: how many neighbours has a node? How far apart are two metabolites? Or two reactions? In the rest of this section, the graph theoretical notions underlying these questions will be introduced.

3.2 Node degree

In a graph, the number of edges having a given node n as their target (respectively source) is called the incoming degree (respectively outgoing degree) of node n , and written $d_i(n)$ (respectively $d_o(n)$). The sum of the incoming and outgoing degrees of a node is called its total degree, and written $d(n) = d_i(n) + d_o(n)$.

3.3 Paths and distances

A path in a graph is a sequence of edges e_1, \dots, e_n such that the target of any edge is the source of the following one. The source $s(e_1)$ of the first edge is called the origin of the path, while the target of the last one is called the destination of the path. If some path has a node m as origin and a node n as destination, then m is connected to n .

The length of a path is defined as its number of edges. In general, a node m is connected to another node n by several paths of different lengths, but there is always at least one shortest path among them. Its length defines the shortest-path length distance $D(m, n)$ from m to n , which can be extended to the case where m is not connected to n with the convention that $D(m, n) = \infty$ in this case.

Note that in a directed graph, the shortest-path length distance is not symmetric, since a path connecting m to n may not connect n to m .

A notion of size of a graph can be derived from this distance, as the largest distance separating two nodes. This quantity is called the diameter of the graph, and holds mainly for graphs having all nodes connected to each other. Note that some authors use a different definition of diameter, as the average distance separating any two nodes in the graph.

4 Topological analysis of metabolic networks

We would like now to challenge some of the properties that were attributed to metabolic networks, or, at least to question some biological interpretations about these mathe-

matical properties. Our purpose is to stimulate a constructive debate, even though some of the statements below might seem provocative at first sight. In each case, we will discuss both pitfalls and strengths of some widespread ideas about network topology. Before entering in the debate, we briefly summarize hereafter the main concepts that will be treated. We will then treat them with more details in the forthcoming sections.

Power-law distribution of the degree. In a random network generated under the Erdős-Renyí model, the number of connections per node is expected to follow a Poisson distribution, whose right tail shows a very rapid decrease of the probability for increasing number of connections. However, in many biological networks, the degree distribution seems to follow a power law, whose right tail decreases much slower than for the Poisson distributions. Power-law networks contain many nodes with a very small number of connections, and a few very highly connected nodes, called hubs (Jeong et al. 2000).

Small world property. Due to the presence of these hubs, it is generally possible to connect any pair of node through relatively short paths. This small-world property can be characterized by computing the network diameter, here defined as the average length of the shortest path between any two nodes of the network. The diameter of the metabolic network has been estimated to correspond to 3 reactions (Fell and Wagner 2000; Jeong et al. 2000), suggesting that molecules can be inter-converted into other ones in a very few metabolic steps.

Scale-free property of the network diameter. Jeong and co-workers analyzed separately the metabolic networks from 40 bacteria, whose genome contained variable numbers of enzymes. Surprisingly, they observed that the diameter of the metabolic network does not vary with the number of enzymes. The network diameter appeared thus as a scale-free property of metabolic networks. The fact for a metabolic network to have a small diameter irrespective of its number of enzymes has been interpreted as an evolutive advantage, since all organisms would be able to efficiently respond to some change in their environment by metabolizing compounds in a few steps.

4.1 Node degree distribution

A global view of the topology of the graph is reflected by the distribution of node degrees across the network, which can be compared to the one observed in randomly generated graphs. A classical model of random graph is the so-called Erdős-Renyí (ER), where one starts from a set of q nodes, and progressively adds n edges that link two randomly selected nodes. All nodes have the same probability to be selected as source or target. In such networks, each node has a given probability to “catch” a given number of edges by chance: some nodes will end up with no edge, some others with 1, 2, 3, . . . edges. It can be shown that the probability for a node to obtain a given number of edges follows a Poisson law.

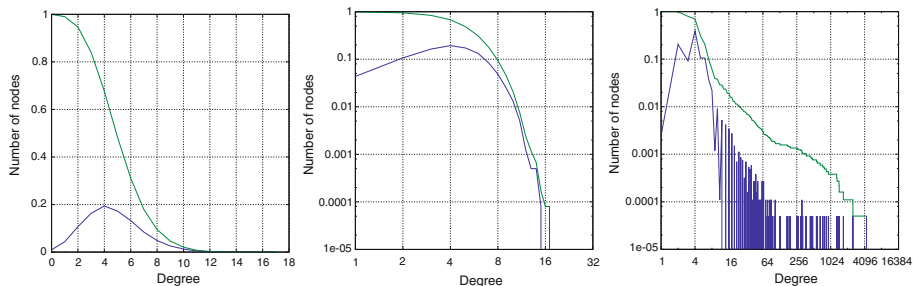


Fig. 2 Degree distributions. Left: Random graph generated according to an ER model, with mean degree $m = 4.6$. center: degree distribution of the same graph, with logarithmic scales. Right: Total degrees distribution of the compounds in the metabolic network derived from KEGG ($p = 11;935$; $n = 27;4446$; $m = 4.6$), in log-log scale. The blue curves represent the probability functions $P(k)$, i.e. the probabilities for a node to be connected to exactly k edges. The green curve shows the inverse cumulative distributions $P(K \geq k)$, i.e. the probabilities for a node to be connected by at least k edges

Figure 2 shows some typical degree distributions exhibited by an ER network and a metabolic network, respectively. The rightmost plot depicts the degree distribution (in log-log scale) observed for all the compounds of a metabolic network derived from the KEGG/LIGAND database. The degree distribution of the metabolic compounds is clearly distinct from the Poisson distribution followed by the degree in an ER network (central plot). In particular, some nodes have an extremely high degree (more than 2000 for the most connected), which would be extremely unlikely in a Poisson distribution. When smoothed, this distribution broadly follows a straight line, which led Jeong and co-workers (2000) to the conclusion that the degree distribution of metabolic compounds follows a power-law distribution, for which the probability of degree k is $P(k) \propto k^{-\alpha}$. Such degree distributions are typical of randomly generated networks where nodes and edges are progressively added, so that the more ancient nodes of the graph are likely to be connected to a larger number of nodes than the most recent nodes.

The power law distribution is characterized by the fact that most nodes have a very few connections, but a few nodes have a very high degree. Such a power-law property had been previously observed for the Internet network, and was subsequently found in basically any other type of network that could be extracted from biological database (protein interactions, transcriptional regulation). Paradoxically, all these reports were based on a visual inspection of the broad shape of the distribution. Recently, Khanin and Wit (2006) tested the goodness-of-fit of a power-law distribution onto 10 biological networks, and showed that not a single one of these networks passed the test! Would the power-law distribution of biological network be a myth, rather than a scientific theory? Actually, the power law property is not so obvious, even by simple visual inspection (Fig. 2C): the left tail of the distribution shows a plateau, whereas a linear decrease would be expected for a power law. This is significant, because each point of this left

Rank	ID	Name	In-degree	Out-degree	Total degree	Rank	ID	Name	In-degree	Out-degree	Total degree
1	C00001	H ₂ O	769	1444	2213	16	C00021	S-Adenosyl-L-homocysteine	227	9	236
2	C00080	H ⁺	809	460	1269	17	C00015	UDP	216	6	222
3	C00007	O ₂	43	817	860	18	C00028	Acceptor	55	134	189
4	C00006	NADP ⁺	318	406	724	19	C00030	Reduced acceptor	132	55	187
5	C00005	NADPH	405	316	721	20	C00027	H ₂ O ₂	142	21	163
6	C00003	NAD ⁺	160	503	663	21	C00026	2-Oxoglutarate	33	125	158
7	C00004	NADH	497	158	655	22	C00020	AMP	144	14	158
8	C00002	ATP	17	449	466	23	C00022	Pyruvate	101	50	151
9	C00011	CO ₂	378	49	427	24	C00024	Acetyl-CoA	35	101	136
10	C00009	Orthophosphate (Pi)	315	78	393	25	C00025	L-Glutamate	83	46	129
11	C00010	CoA	242	127	369	26	C00029	UDP-glucose	2	97	99
12	C00008	ADP	313	20	333	27	C00033	Acetate	80	17	97
13	C00014	NH ₃	253	43	296	28	C00042	Succinate	76	13	89
14	C00013	Pyrophosphate (PPi)	256	30	286	29	C00031	D-Glucose	65	21	86
15	C00019	S-Adenosyl-L-methionine	6	239	245	30	C00051	Glutathione	16	49	65

Fig. 3 Compounds with the highest degree in the KEGG/LIGAND metabolic network. Most are cofactors (NADP + = NADPH, NAD + = NADH, ATP), but the highest degrees are attained by universal metabolites (water, CO₂)

represents several thousands of compounds (its relative importance is somewhat masked by the logarithmic *Y* scale).

Interestingly, such deviations have sometimes been attributed to the incompleteness of databases rather than to the relevance of the statistical model. This is hopefully not the most frequent attitude, but the reader's attention is drawn on the dangers of turning some exciting hypothesis into a dogma, that can even not be questioned in the light of data!

Beyond this argument, it is important to wonder what these properties reveal us about the underlying biochemistry. If molecules are named instead of being qualified according to their topological properties (Fig. 3), one immediately notices that the most connected compounds are actually those ensuring some basic chemical operations such as redox (NAD⁺, NADH, NADP⁺, NADPH, H₂O, H⁺, O₂) or energy transfer (ATP, ADP, Pi, PPi). The presence of "hubs" in the metabolic network simply reflects the well-known fact that these basic operations are applied on several hundreds of different molecules.

4.1.1 Robustness to random deletions and targeted attacks

By computer-based simulations, it has been shown that power-law networks are robust to random deletions, but sensitive to targeted attacks (i.e. deleting hubs). Indeed, deletion of a randomly selected subset of nodes barely affects their smallworld property as far as the hubs are there to ensure short distances between any pair of nodes. On the contrary, if these hubs are dismantled by targeted attack, the network diameter rapidly increases. Such computer simulations offer a practical tool to estimate the impact of progressive hub removal on the behaviour of a network, and have some important consequences on the design of human-built networks such as Internet.

Some authors transposed these properties to metabolic networks, and inferred that hubs would play an important role in ensuring robustness to random deletions. However, the concept of hub removal simply does not apply to metabolic networks, for the simple reason that the "hubs" of these networks represent molecules, that can

simply not be removed from the network. For example, even a bacteria with a very small genome already contains several hundreds of enzymes that catalyze H_2O -producing reactions. Thus, removing a single hub (the H_2O molecule) from the metabolic network of such an organism would require the deletion of several hundreds of enzymes. For a geneticist, it is clear that the deletion of a few tens of enzyme-coding genes would already be lethal, and this would happen much before the last H_2O -producing enzyme has been deleted. Thus, targeted removal of “hub” compounds can thus neither be achieved naturally, nor by human intervention.

What about random deletions then? The “deletion” of poorly connected compounds seems practically feasible: if a specific compound is produced by one or a few enzymes only, these enzymes could be inactivated by spontaneous mutation or directed mutagenesis. Actually, this is exactly the procedure that has been followed by biochemists during the last 50 years, in order to isolate enzymes and decipher pathways: their preferred model organism (e.g. the bacteria *Escherichia coli* K12, or the yeast *Saccharomyces cerevisiae*) were submitted to a mutagenesis, followed by a screening to select colonies which had lost the capacity to grow in the absence of a given metabolite (e.g. methionine, lysine). Those mutants were called auxotrophic, to denote their dependency towards this metabolite. The simple fact that most of the enzymes that we currently know were isolated by their auxotrophy phenotype shows that the mutation of a single enzyme often suffices to block a whole metabolic pathway, which cannot be compensated by the other enzymes of the genome. It seems thus obvious that the high degree of connectivity of “hubs” such as H_2O , NADP, O_2 can by no means compensate for the absence of an enzyme like aspartate kinase, which catalyses a single reaction, essential to the biosynthesis of three amino acids (methionine, lysine and threonine) in bacteria. In summary, in the particular case of metabolic networks, the power-law property neither confers resistance to random deletions, nor sensitivity to targeted attacks. This illustrates the importance to be careful when transposing concepts from mathematical models and biological system.

4.1.2 Generative models for power-law networks

Another danger comes from the transposition of generative models to metabolic networks. As said before, it is possible to generate a random graph with power-law distribution of degree, by progressively adding nodes and edges to the existing graph, but with an increased probability for the ancient node to catch a new edge. By extension, some bioinformaticians proposed that the “hubs” of the metabolic network correspond to compounds that appeared earlier during the evolution of metabolism. This attractive hypothesis unfortunately does not hold when analyzing the structure of the most connected molecules of the metabolic network (Table 3). For example, ATP, which is involved in 466 reactions, is a quite complex molecule, comprising a sugar group + a heterocyclic base + 3 phosphate groups. It is obvious that thus molecule may not have

been present before the smaller molecules that are part of it, irrespective of the fact that these sub-components are involved in much less reactions than the “hubs”. Another illustrative example is S-Adenosyl-L-methionine, which is made from methionine and ATP. The same observation holds for many other “metabolic hubs”. As discussed above, that there is a perfectly understandable reason for the high connectivity of these molecules, and which has nothing to do with their supposed chronology of appearance: ATP is involved in energy transfer, and S-Adenosyl-L-methionine in methyl transfer. This example again emphasizes the importance of including as much biological knowledge for the interpretation of system properties.

4.2 Paths and distances in metabolic networks

The small-world and scale-freeness properties rely on a computation of the shortest path between all pairs of nodes. It would be tempting to interpret a shortest path as a metabolic pathway, in its common biochemical acceptance, i.e. a set of connected reactions that transform a set of input metabolites (substrates) into a set of output metabolites (products). This would however be erroneous, because the shortest paths in metabolic graph generally contain irrelevant inter-connections between reactions, as discussed in detail in previous publications (van Helden et al. 2002; Arita 2004, 2005; Croes et al. 2005, 2006). Metabolic hubs should generally not be used as intermediates between two reactions. As shown in Fig. 1, in the yeast *Saccharomyces cerevisiae* methionine is synthesized from L-aspartate in 6 reactions. The bacteria *Escherichia coli* also synthesizes methionine from L-aspartate, but the two pathways differ by a few intermediate steps. As a matter of test, we applied a k -shortest path finding algorithm to obtain the 5 shortest paths from L-aspartate to L-methionine in 3 metabolic networks.

1. A raw graph containing all compounds and reactions, including the “hubs” (Fig. 4A). In each one of the 5 shortest paths, L-aspartate is converted into L-methionine in no more than 2 steps. However, the returned paths are completely invalid on the biochemical point of view, because “hubs” are used as intermediate metabolites between two compounds, without sharing anything in common with their structure.
2. The filtered graph (Fig. 4B) contains the same molecules and reactions, except for 36 highly connected molecules that were discarded from the network (van Helden et al. 2002). The most trivial connections via H_2O or AMP are avoided, but the paths remain very short, and have nothing in common with the known methionine biosynthetic pathways (Fig. 1).
3. The weighted graph (Fig. 4C) contains all the molecules (including the hubs), but a weight is assigned to each compound, proportional to its degree (Croes et al. 2006). The program searches the k lightest paths, i.e. those having the lowest weight. The 4 top-ranking paths partly correspond to the bacterial pathway (Fig. 1B), whereas the 5th path corresponds to the yeast pathway (Fig. 1A).

A	1.	L-aspartic acid → <u>6.3.5.4</u> → <u>AMP</u> → <u>6.1.1.10</u> → L-methionine
	2.	L-aspartic acid → <u>3.5.1.15</u> → H ₂ O → <u>3.4.13.12</u> → L-methionine
	3.	L-aspartic acid → <u>3.5.1.15</u> → H ₂ O → <u>3.4.13.12</u> → L-methionine
	4.	L-aspartic acid → <u>4.3.1.1</u> → NH ₃ → <u>4.4.1.11</u> → L-methionine
	5.	L-aspartic acid → <u>3.5.1.15</u> → H ₂ O → <u>3.5.1.31</u> → L-methionine
B	1.	L-aspartic acid → <u>2.6.1.35</u> → glycine → <u>2.6.1.73</u> → L-methionine
	2.	L-aspartic acid → <u>2.6.1.12</u> → L- α -alanine → <u>2.6.1.44</u> → glycine → <u>2.6.1.73</u> → L-methionine
	3.	L-aspartic acid → <u>2.6.1.12</u> → L- α -alanine → <u>2.6.1.41</u> → d-methionine → <u>5.1.1.2</u> → L-methionine
	4.	L-aspartic acid → <u>2.6.1.12</u> → L- α -alanine → <u>2.6.1.2</u> → o-acetyl-L-homoserine → <u>2.5.1.49</u> → L-methionine
	5.	L-aspartic acid → <u>4.1.1.12</u> → L- α -alanine → <u>2.6.1.44</u> → glycine → <u>2.6.1.73</u> → L-methionine
C	1.	L-aspartic acid → 2.7.2.4 → L-4-aspartyl phosphate → 1.2.1.11 → L-aspartic 4-semialdehyde → 1.1.1.3 → L-homoserine → 2.3.1.31 → o-acetyl-L-homoserine → <u>2.5.1.49</u> → L-methionine
	2.	L-aspartic acid → 2.7.2.4 → L-4-aspartyl phosphate → 1.2.1.11 → L-aspartic 4-semialdehyde → 1.1.1.3 → L-homoserine → 2.3.1.31 → o-acetyl-L-homoserine → <u>2.5.1.49</u> → L-methionine
	3.	L-aspartic acid → <u>3.5.5.4</u> → L-beta-cyanoalanine → R03972 → L-2,4-diaminobutyrate → <u>2.6.1.46</u> → L-aspartic 4-semialdehyde → 1.1.1.3 → L-homoserine → 2.3.1.31 → o-acetyl-L-homoserine → <u>2.5.1.49</u> → L-methionine
	4.	L-aspartic acid → <u>3.5.5.4</u> → L-beta-cyanoalanine → R03972 → L-2,4-diaminobutyrate → <u>2.6.1.46</u> → L-aspartic 4-semialdehyde → 1.1.1.3 → L-homoserine → 2.3.1.31 → o-acetyl-L-homoserine → <u>2.5.1.49</u> → L-methionine
	5.	L-aspartic acid → 2.7.2.4 → L-4-aspartyl phosphate → 1.2.1.11 → L-aspartic 4-semialdehyde → 1.1.1.3 → L-homoserine → 2.3.1.46 → o-succinyl-L-homoserine → <u>2.5.1.48</u> → L-cystathionine → <u>2.5.1.49</u> → o-acetyl-L-homoserine → <u>2.5.1.49</u> → L-methionine

Fig. 4 The 5 shortest paths found between L-aspartate and L-methionine in the metabolic network derived from KEGG/LIGAND, in the raw graph containing all compounds (A), in a “filtered” graph from which 36 compounds had been excluded (B), and in a weighted graph containing all compounds (C). Inferred reactions and compounds that correspond to the annotated pathways are in bold. Incorrect reactions and compounds are underlined

A systematic evaluation based on 148 annotated pathways revealed that the correspondence between these pathways and those inferred by shortest path finding in the raw graph is only 30% and, in addition, these correspond to very short pathways (3–4 reactions including the seed nodes), where one or two reactions had to be inferred, no more. The removal of a selected subset of compounds increases the accuracy to 65%. The most convincing results were obtained with the weighted graph, where the average accuracy for the shortest reaches 85% (Croes et al. 2006; Brohée et al. 2008; Brohée submitted).

5 Assessing reconstructed metabolic networks against physiological data

Given its enzymatic arsenal, an organism may be able to grow on various media, but unable to grow on some others, depending on its capacity to catalyze a set of reactions that together can produce all biomass precursors from the available nutrients. Since at least some of the media on which an organism of interest is able to grow are usually known, checking whether the reconstructed metabolic network of an organism indeed permits the conversion of the nutrients into biomass can be used as a mean of assessing its completeness and correctness. Since a reaction can occur only when all of its substrates are present, topological analyses are not able to answer such questions. The

constraints based modelling framework introduced below is precisely aimed at tackling these questions.

The Flux balance analysis approach considers steady-state distributions of the reaction fluxes across the metabolic network for which the net production rate of all the intermediate metabolites is zero: fluxes must be balanced around each metabolite (Schilling and Palsson 2000). Instantaneous reaction rates should be interpreted as averages over some long time period of growth, and the flux balance assumption as their material possibility without lacking (or, conversely, accumulating continuously) any intermediate metabolite.

Constraints-based models of metabolism do not involve further knowledge of the metabolism of an organism than the stoichiometric coefficients of all the reactions. Thank to this simplicity, the flux balance analysis can perform growth phenotype predictions by integrating the biochemical knowledge on a species at the genome scale, which makes it an efficient procedure for assessing a reconstructed metabolic network against growth phenotypes on various media, and/or gene essentiality data. Furthermore, any inconsistency between data and predictions can be solved by only one type of modifications of the network (addition or deletion of a reaction, or, in the case of essentiality data, increase or decrease of the impact of the deletion of one gene on a reaction).

5.1 Constraints-based models of metabolism

5.1.1 The flux balance hypothesis

We consider the metabolic network of an organism with p metabolites m_1, \dots, m_p and q reactions r_1, \dots, r_q . The stoichiometric matrix S has entry $S(i, j)$ equal to the stoichiometric coefficient of the metabolite m_i in the reaction r_j . $S(i, j)$ is set to zero if m_i is neither a substrate or a product of reaction r_j , and is by convention negative (resp. positive) if m_i is a substrate (resp. a product) of r_j .

Denote by n the flux distribution, which is a vector formed by all reaction rates ν_j , $j = 1, \dots, q$. The net production rate of any metabolite m_i is then given by the i^{th} entry of the vector $S \cdot n$. The flux-balance hypothesis imposes the nullity of all entries of $S \cdot n$ associated to internal metabolites. Writing S_{int} for the stoichiometric matrix from which all rows associated to non-internal metabolites were removed, this property can be written:

$$S_{\text{int}} \cdot n = 0 \tag{1}$$

In mathematical terms, the set C formed by all vectors n such that (1) holds is a linear subspace of R^q , called the kernel of the matrix S_{int} .

At this point, we did not account for irreversible reactions. But setting the additional constraint that $\nu_j = 0$ whenever r_j is irreversible discards from C all flux distributions

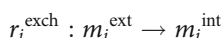
having at least one reaction rate incompatible with the known direction of an irreversible reaction, which solves this issue. After this operation, C ceases being invariant under any linear combination of its elements, but only under linear combination with positive coefficients: it is a cone. The magnitude of a flux distribution thus has no intrinsic meaning, and when flux optimizations are performed, its explosion is prevented by imposing upper bounds on the uptake rates. The set C has then the structure of a polyhedral cone.

An important point to be made here is that turning a metabolic network into a stoichiometric model is not straightforward at the time of writing. Major hurdles in this process include the presence of generic metabolites and reactions in the databases, e.g. “an alcohol and an acetyl-CoA are converted into an acetic ester, releasing a coenzyme A”. Since the metabolism of a species usually involves several different alcohols, including such reactions would prevent from balancing each of them. A similar issue results from the polymerisation processes, which are also frequently represented in databases by single reactions corresponding to one step of extension of the polymer. Another shortcoming of metabolic models reconstruction is related to the inference of the transporters, which are usually poorly predicted compared to enzymes. This is crucial for metabolic models since many transport systems are coupled to a reverse transport of other molecules. The TransportDB database gathers descriptions of several transport systems (Ren et al. 2004), and is a key resource for reconstructing metabolic models. Last but not least, even bacterial metabolism is best described if the different compartments of the cell are modelled. In the case of constraints-based models, compartments are represented by the addition of transport reactions across membranes, and by duplicating metabolites appearing in several compartments so that they can be balanced in each of them.

5.1.2 Modelling the growth medium

The non-internal metabolites evoked earlier are metabolites that the organism can take from the environment, or that it can secrete to. These exchanges are represented in the metabolic model using transport reactions, that merely convert external metabolites into their internal equivalent. Even though these processes are most often omitted in metabolic networks, they may involve cofactors and thus interfere with the balance of internal metabolites.

Formally, it requires to duplicate each transportable metabolite m_i into its external m_i^{ext} and internal forms m_i^{int} , and to augment the reaction set with the corresponding transport reactions:



Their reaction rates may be subjected to irreversibility constraints, but other constraints on them are induced by the composition of the medium. Indeed, metabolites that are not

provided in the growth medium may not be used by the cell to feed its metabolism. Writing M for the set of metabolites available from the medium, the following constraints are added:

$$\forall i \in M, \quad v_i^{\text{exch}} \leq 0$$

The constraints defining an attainable metabolic flux distribution can thus be encoded in the model using only linear equality and inequality constraints.

5.1.3 Biomass function

In addition to the exchangeable metabolites provided in the medium, the components of the cell itself (nucleotides, amino-acids, and lipids), which are known to accumulate when the cell grows and divides, should not be subjected to the flux balance constraint.

Based on biochemical analyses of dried cells, the proportion of those components in the biomass can be determined. This composition can be reproduced in the model by adding a virtual irreversible reaction, the biomass assembly reaction, that assembles all precursors into a virtual metabolite, the biomass. The stoichiometry of this reaction is chosen to reflect the known composition of the biomass, and includes an energetic cost through the degradation of a certain amount of energy carriers (like ATP).

When a biomass reaction is defined, biomass precursors are considered as normal internal metabolites, of which the net production rate should be zero. But since the biomass reaction collects the end products of all biosynthesis pathways, it is responsible for strongly constraining the flux distribution.

5.2 Predicting metabolic capabilities

5.2.1 Predicting growth on a defined medium

Returning to the original question that motivated this section, we can now ask whether an organism is able to produce biomass when it is fed with a given medium. The composition of the medium being translated into additional constraints, the question to ask is now: which biomass production rate is attainable by the organism? Of course, the absolute value of this rate has to be normalized against the uptake rate of one of the nutrients (usually the carbon source) in order to have any meaning.

The simplest way to answer this question is to perform a flux balance analysis (Schilling et al. 2001), i.e. to search for the maximal achievable growth rate, the normalizing uptake rate being forced to 1. If this maximal rate is too close from zero, then the organism is predicted not to be able to grow on the medium. Otherwise, its growth is considered possible, although other phenomena (for instance transcriptional or metabolic regulation) may prevent it.

An alternative approach, known as the metabolite producibility check, does not take into account the stoichiometry of the biomass reaction, but rather checks whether the maximal production rate attainable is far from zero for all of its substrates. Since the removal of the biomass reaction from the model releases constraints, an organism is more easily predicted to be able to grow by this method than by FBA.

All of these methods rely on an optimisation of some flux in the flux cone, and can be easily implemented using linear programming algorithms. They have been performed on several organisms, and yielded satisfactory results (Edwards et al. 2001; Schilling and Palsson 2000; Jamshidi et al. 2001). Actually, this method is not only able to predict whether growth is achievable on some medium. It has also been shown that the predicted flux distribution (i.e. the one achieving the optimal growth rate) across the network is coherent with experimental measurements (such as the ratio of oxygen to carbon consumptions) (Edwards et al. 2001).

5.2.2 Predicting gene essentiality

When constrained to a null flux, some reactions completely prevent the optimal biomass production rate from being significantly different from zero. These reactions are called essential, since the organism cannot grow if they cannot proceed. Thus, gene deletion mutants lacking a gene (or several genes) encoding the enzyme (or enzymes) required for an essential reaction to take place are not expected to be viable.

Precisely defining the sets of genes whose deletion would inhibit a reaction is thus a key element for predicting the growth capacities of gene-deletion mutants, which is formalized using gene-product-reaction (GPR) relationships (Joyce et al. 2006). A GPR is a boolean predicate involving the genes presence or absence defined for each catalyzed reaction in the model, which evaluates to true if the gene presence profile indeed allows the reaction to be catalyzed. It encodes the complexation of proteins using and operations, while isozymes are encoded using the or operator.

For instance, the GPR of a reaction r_i catalyzed by the complex formed by the products of gene A and gene B, and a third protein among the products of the genes C and D would be : A and B and (C or D). Predicting all the enzymatic complexes in an organism remains a difficult task, and constitutes an additional obstacle on the path from a metabolic network to a metabolic model.

When the GPR relations are known, the impact of gene deletions on the capacity to grow on some medium can be assessed as follows. For each reaction whose GPR evaluates to false, an additional constraint imposes the nullity of its flux. Then the FBA analysis is performed, and the theoretically maximal attainable biomass yield on a given growth medium can be derived in exactly the same manner as previously.

It is noteworthy that as the deletion of a gene coding an enzyme (or the blocking of a reaction) is translated into an additional constraint on the flux distribution, it may only result in a decreased biomass yield. This is a consequence of the approach, that considers

only the metabolic capacities conferred by the enzymatic arsenal an organism disposes of, without regards to regulatory effects (even though extensions of this modelling framework to include regulatory constraints have been proposed (Covert and Palsson 2002)).

This approach has been applied to several organisms, such as *Escherichia coli* using gene essentiality data from the Keio collection (Bara et al. 2006; Joyce et al. 2006), or *Acinetobacter baylyi* (Durot et al. 2007; de Berardinis et al. 2008).

5.3 Assessing and correcting models using experimental data

When experimental growth phenotypes are available, their comparison to the corresponding predictions of the models may reveal two types of discrepancies: a false viability prediction (the organism or mutant is predicted to grow on the medium, but does not experimentally), or a false non-viability prediction. In the first case, the absence of growth may be due to two types of causes: a regulatory limitation, or the absence of a reaction that is in the model. In the second case, an essential reaction for the organism to grow on the medium is necessarily missing from the model.

When compared to gene essentiality data, mispredictions may also be due to erroneous GPR relationships. A false viability prediction of a mutant may indeed also reveal an under-estimated impact of the gene deletion on some reaction, while the opposite occurs when a gene-deletion is erroneously predicted to inhibit a reaction.

Since the predicted phenotypes are changed in such a monotonic manner when reactions are added or removed (and when the impact of a gene deletion is increased or decreased), discrepancies between predictions and growth phenotypes or gene essentiality data indicate directions for correcting the model. This fact has been leveraged in the course of the work on *Acinetobacter baylyi* mentioned above, for which a novel method for enumerating GPR relationships compatible with both the metabolic network and the essentiality data has been developed, yielding candidate GPR and/or network corrections (Schächter and Durot, in preparation).

5.4 Structural properties of the flux cone

The set of attainable flux distributions features much richer properties than just an optimal growth rate. Several approaches have indeed been proposed to investigate the structure of the flux cone, typically by characterizing correlated behaviors among sets of reactions.

An example of such a structural property is provided by the flux coupling analysis (Burgard et al. 2004), which searches for couple of reactions whose fluxes are linked all across the flux cone. Fully coupled reactions pairs have proportional fluxes, while directionally coupled reaction pairs are linked by inequality relationships. Since the full coupling relationship is symmetric and transitive, it defines fully coupled sets of reactions, whose fluxes can be described using a single number without loss of generality.

The same idea applied to the characterization of larger sets of reactions yielded the notion of elementary modes (Pfeiffer et al. 1999; Schuster et al. 1999), which are minimal sets of reactions through which a non-null flux is attainable, the remainder of the metabolic network being inactive.

Such properties are said structural because they hold for any attainable flux distribution, and thus reveal sets of reactions that coherently participate in the metabolic processes. The rather subjective definition of pathway mentioned in the introduction is indeed absent from the constraints-based formalism, but elementary modes as well as fully coupled sets provide model-derived alternatives. Furthermore, these structures not only simplify the biological interpretation of flux distributions, but also reveal strong constraints the organism has to cope with.

5.5 Working with constraints-based models

Computing with constraints-based metabolic models at genome-scale requires adequate software tools. The first such tool to be widely adopted is actually a set of scripts for the Matlab environment called FluxAnalyzer (Klamt et al. 2003), which recently evolved into CellNetAnalyzer (Klamt et al. 2007). One appealing feature of this tool is the possibility it offers of displaying flux distributions on metabolic maps.

Several other environments dedicated to stoichiometric modeling have been developed, most of them consisting in sets of scripts performing a variety of constraints-based related analyses (such as ScrumPy (Poolman 2006), or the cobra toolbox (Becker et al. 2007)). The Sympheny software platform, developed by the Genomatica company, advertises a large-set of reconstruction and analyses functionalities, but is not freely available.

A software platform dedicated to constraints-based modelling of metabolism based on a robust data model, called NemoStudio (Combe et al. in preparation), aims at providing the user with a complete *in silico* workbench supporting extensions and scripting languages. It also features a web interface allowing to perform online simulations of the models.

Conclusion

With the dramatic increase of the number of microbial genomes sequenced each year, it is now critical to speed-up the reconstruction process of a metabolic network from a genome annotation. It is expected that adequate definitions of metabolic modules, subnetworks defined as performing a given metabolic function with possible species- or taxon-specific variations in the precise set of reactions implementing that function, would (a) facilitate automated network reconstruction by reducing the search space and (b) permit “transversal” curation of partial metabolic networks across species

(Overbeek et al. 2005). The SEED database pioneered this vision by introducing the notions of subsystems and variants, with the latter goal of transversal curation as its main focus. Following the same path, the KEGG database recently released the KEGG modules (Kanehisa et al. 2008), while SwissProt is progressively adopting the data model of its UniPathway database as its ontology for metabolic gene annotations.

Downstream analyses of metabolic networks using constraints-based models are also expected to be facilitated by these evolutions of pathways databases. Conversely, metabolic models are also expected to be increasingly used as framework for assessing reconstructed metabolic networks against experimental data. Especially promising in that respect are gene essentiality datasets, growth phenotypes acquired on several media, and measurements of intra-cellular concentrations of metabolites (metabolomics). Deducing additional constraints on fluxes, and in some cases on metabolite concentration ranges via the use of additional thermodynamics constraints, is a very active research topic (Yang et al. 2005; Kummel et al. 2006).

References

- Arita M (2004) The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA* 101(6): 1543–1547
- Arita M (2005) Scale-freeness and biological networks. *J Biochem* 138(1): 1–4
- Bara T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko K, Tomita M, Wanner B, Mori H (2006) Construction of *Escherichia coli* k12 in-frame, singlegene knockout mutants: the keio collection. *Mol Syst Biol* 1: 2:2006.0008
- Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox. *Nat Protocols* 2: 727–738
- de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, Cruaud C, Samair S, Lechaplais C, Gyapay G, Richez C, Durot M, Kreimeyer A, Le Fevré F, Schächter V, Pezo V, Doring V, Scarpelli C, Medigué C, Cohen GN, Marlieré P, Salanoubat M, Weissenbach J (2008) A complete collection of single-gene deletion mutants of acinetobacter baylyi adp1. *Mol Syst Biol* (in press)
- Brohée S, Faust K, Vanderstocken G, van Helden J (2008) Network analysis tools: from biological networks to clusters and pathways (submitted)
- Brohée S, Faust K, Lima-Mendez G, Sand O, Janky R, Vanderstocken G, Deville Y, van Helden J (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res*
- Burgard AP, Nikolaev EV, Schilling CH, Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* 14: 301–312
- Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD (2007) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res* 36 (Database issue): D623–D631
- Combe C, Le Fevré F, Smidtas S, Schächter V (in preparation) Nemostudio: a software platform for constraints-based modelling of metabolism
- Covert MW, Palsson BO (2002) Transcriptional regulation in constraints-based models of *Escherichia coli*. *J Bio Chem* 277(31): 28,058–28,064

- Croes D, Couche F, Wodak SJ, van Helden J (2005) Metabolic pathfinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res* 33(Web Server issue): W326–W330
- Croes D, Couche F, Wodak SJ, van Helden J (2006) Inferring meaningful pathways in weighted metabolic networks. *J Mol Biol* 356(1): 222–236
- Durot M, Le Fevré F, De Berardinis V, Kreimeyer A, Weissenbach J, Schächter V (2007) Reconstruction of a global model of acinetobacter baylyi metabolism using genome-wide conditional essentiality data on several media. In: 2nd ASM Conference on Integrating Metabolism and Genomics, Am Soc Microbiol
- Edwards JS, Ibarra RU, Palsson BO (2001) In silico predictions of *escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 19: 125–130
- Fell DA, Wagner A (2000) The small world of metabolism. *Nat Biotechnol* 18(11): 1121–1122
- Green ML, Karp PD (2004) A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5: 76
- van Helden J, Wernisch L, Gilbert D, Wodak SJ (2002) Graph-based analysis of metabolic networks. In: al MHWe (ed) Ernst Schering Res Found Workshop, Springer-Verlag, pp 245–274
- 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, Novere NL, Loew LM, Lucio D, Mendes P, Minch E, Mjølness ED, Nakayama Y, adn P F Nielse MRN, Sakurada T, Schaff JC, Shapiro BE, Shimizu T, 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(4): 524–531
- Jamshidi N, Edwards JS, Fahland T, Church GM, Palsson BO (2001) Dynamic simulation of the human red blood cell metabolic network. *Bioinformatics* 213(1): 286–287
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. *Nature* 407(6804): 651–654
- Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411(6833): 41–42
- Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 33(Database Issue): D428–D432
- Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BO, Agarwalla S (2006) Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *J Bacteriol* 188(23): 8259–8271
- Kanehisa M (1996) Toward pathway engineering: a new database of genetic and molecular pathways. *Sci Technol Japan* 59: 34–38
- Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res* 34(Database issue): D354–D357
- Kanehisa M, Araki M, Goto S, Hattori M, Hirawaka M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) Kegg for linking genomes to life and environment. *Nucleic Acids Res* 36(Database issue): D480–D484
- Karp P, Riley M, Saier M, Paulsen I, Paley S, Pellegrini-Toole A (2002) The ecocyc database. *Nucleic Acids Res* 30(1): 56–58
- Karp P, Ouzounis C, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N (2005) Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 33(19): 6083–6089
- Karp PD, Riley M, Paley SM, Pelligrini-Toole A (1996) Ecocyc: an encyclopedia of *escherichia coli* genes and metabolism. *Nucleic Acids Res* 24(1): 32–39
- Khanin R, Wit E (2006) How scale-free are biological networks. *J Comput Biol* 13(3): 810–818

- Kharchenko P, Vitkup D, Church GM (2004) Filling gaps in a metabolic network using expression information. *Bioinformatics* 20(Suppl 1): i178–i185
- Klamt S, Saez-Rodriguez J, Ginkel M, Gilles E (2003) Fluxanalyzer: exploring structure, pathways and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics* 19(2): 261–269
- Klamt S, Saez-Rodriguez J, Gilles E (2007) Structural and functional analysis of cellular networks with cellnetanalyzer. *BMC Syst Biol* 1: 2
- Krummenacker M, Paley S, Yan T, Karp PD (2005) Querying and computing with biocyc databases. *Bioinformatics* 21(16): 3454–3455
- Kummel A, Panke S, Heinemann M (2006) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome. *Mol Syst Biol* 2:2006.0034
- Le Fevré F, Smidtas S, Schächter V (2007) Cyclone: Java-based querying and computing with pathway genome databases. *Bioinformatics* 23(10): 1299–1300
- Luciano JS, Stevens RD (2007) e-Science and biological pathway semantics. *BMC Bioinformatics* 8(S3)
- Ogata H, Goto S, Fujibuchi W, Kanehisa M (1998) Computation with the kegg pathway database. *Biosystems* 47(1–2): 119–128
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33(17): 5691–5702
- Paley S, Karp PD (2002) Evaluation of computational metabolic-pathway predictions for *h. pylori*. *Bioinformatics* 18: 715–724
- Pfeiffer T, Sanchez-Valdenebro I, Nuno JC, Montera F, Schuster S (1999) Metatool: for studying metabolic networks. *Bioinformatics* 15(3): 251–257
- Poolman MG (2006) Metabolic modelling with python. *IEEE Proc Syst Biol* 153: 375–378
- Potapov AP, Voss N, Sasse N, Wingender E (2005) Topology of mammalian transcription networks. *Genome Inform* 16(2): 270–278
- Ravasz E, Barabasi AL (2003) Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 67(2 Pt 2): 026,112
- Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297(5586):1551–1555
- Ren Q, Kang KH, Paulsen IT (2004) Transportdb: a relational database of cellular membrane transport systems. *Nucleic Acids Res* 1(32(Database issue)): D274–D279
- Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J (2006) Regulondb (version 5.0): *Escherichia coli* k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 34(Database issue): D394–D397
- Schächter V, Durot M (in preparation) Systematic refinement of genome-scale metabolic models using gene essentiality data
- Schilling CH, Palsson BO (2000) Assessment of the metabolic capabilities of haemophilus influenza rd through a genome-scale pathway analysis. *J Theor Biol* 203(3): 249–283
- Schilling CH, Edwards JS, Letscher D, Palsson B (2001) Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol Bioeng* 71(4): 286–306
- Schuster S, Dandekar T, Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol* 17(2): 53–60
- Yang F, Qian H, Beard DA (2005) Ab initio prediction of thermodynamically feasible reaction directions from biochemical network stoichiometry. *Metab Eng* 7: 251–259

CHAPTER 6.2

Protein–protein interactions: analysis and prediction

D. Frishman^{1,2}, M. Albrecht³, H. Blankenburg³, P. Bork^{4,5},
E. D. Harrington⁴, H. Hermjakob⁶, L. Juhl Jensen^{4,7}, D. A. Juan⁸,
T. Lengauer³, P. Pagel¹, V. Schachter⁹ and A. Valencia^{8f}

¹Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising, Germany

²Institute for Bioinformatics and Systems Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany

³Department of Computational Biology and Applied Algorithmics, Max-Planck-Institute for Informatics, Saarbrücken, Germany

⁴Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

⁵Max-Delbrück-Centre for Molecular Medicine, Berlin-Buch, Berlin, Germany

⁶European Molecular Biology Laboratory Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

⁷Novo Nordisk Foundation Center for Protein Research, Panum Institute, Copenhagen, Denmark

⁸Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernandez Almagro, Madrid

⁹Computational Systems Biology Group – Genoscope – CEA, Evry, France

1 Introduction

Proteins represent the tools and appliances of the cell – they assemble into larger structural elements, catalyze the biochemical reactions of metabolism, transmit signals, move cargo across membrane boundaries and carry out many other tasks. For most of these functions proteins cannot act in isolation but require close cooperation with other proteins to accomplish their task. Often, this collaborative action implies physical interaction of the proteins involved. Accordingly, experimental detection, *in silico* prediction and computational analysis of protein–protein interactions (PPI) have attracted great attention in the quest for discovering functional links among proteins and deciphering the complex networks of the cell.

Proteins do not simply clump together – binding between proteins is a highly specific event involving well defined binding sites. Several criteria can be used to further classify interactions (Nooren and Thornton 2003). Protein interactions are not mediated by covalent bonds and, from a chemical perspective, they are always

Corresponding author: Dmitrij Frishman, Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany (e-mail: d.frishman@wzw.tum.de)

reversible. Nevertheless, some PPI are so persistent to be considered irreversible (obligatory) for all practical purposes. Other interactions are subject to tight regulation and only occur under characteristic conditions. Depending on their functional role, some protein interactions remain stable for a long time (e.g. between proteins of the cytoskeleton) while others last only fractions of a second (e.g. binding of kinases to their targets). Protein complexes formed by physical binding are not restricted to so called binary interactions which involve exactly two proteins (dimer) but are often found to contain three (trimer), four (tetramer), or more peptide chains. Another distinction can be made based on the number of distinct proteins in a complex: homo-oligomers contain multiple copies of the same protein while hetero-oligomers consist of different protein species. Sophisticated “molecular machines” like the bacterial flagellum consist of a large number of different proteins linked by protein interactions.

2 Experimental methods

The focus of this chapter is on the computational methods for analyzing and predicting protein–protein interactions. Nevertheless, some basic knowledge about experimental techniques for detecting these interactions is highly useful for interpreting results, estimating potential biases, and judging the quality of the data we use in our work.

Many different types of methods have been developed but the vast majority of interactions in the literature and public databases come from only two classes of approaches: co-purification and two-hybrid methods. Co-purification methods (Rigaut et al. 1999) are carried out *in vitro* and involve three basic steps. First, the protein of interest is “captured” from a cell lysate – e.g. by attaching it to an immobile matrix. This may be done with specific antibodies, affinity tags, epitope tags along with a matching antibody, or by other means. Second, all other proteins in the solution are removed in a washing step in order to purify the captured protein. Under suitable conditions, protein–protein interactions are preserved. In the third step, any proteins still attached to the purified protein are detected by suitable methods (e.g. Western-blot or mass spectrometry). Hence, the interaction partners are co-purified, as the name of the method implies.

The two-hybrid technique (Fields and Song 1989) uses a very different approach – it exploits the fact that transcription factors such as Gal4 consist of two distinct functional domains. The DNA-binding domain (BD) recognizes the transcription factor (TF) binding site in the DNA and attaches the protein to it while the activation domain (AD) triggers transcription of the gene under the control of the factor. When expressed as separate protein chains, both domains remain fully functional: the BD still binds the DNA but lacks a way of triggering transcription. The AD could trigger transcription but has no means of binding to the DNA. For a two-hybrid test, two proteins X and Y are *fused* to these domains resulting in two hybrids: X-BD and Y-AD. If X binds to Y, the

resulting protein complex turns out to be a fully functional transcription factor. Accordingly, an interaction is revealed by detecting transcription of the *reporter gene* under the control of the TF. In contrast to co-purifications, the interaction is tested *in vivo* in the two-hybrid system (usually in yeast, but other systems exist).

The above description refers to small-scale experiments testing one pair of proteins at a time, but both approaches have successfully been extended to large-scale experiments testing thousands of pairs in very short time. While such high-throughput data is very valuable, especially for computational biology which often requires comprehensive input data, a word of caution is necessary. Even with the greatest care and a maximum of thoughtful controls, high-throughput data usually suffer from a certain degree of false-positive results as well as false-negatives compared to carefully performed and highly optimized individual experiments.

The ultimate source of information about protein interactions is provided by high-resolution three-dimensional structures of interaction complexes, such as the one shown in Fig. 1. Spatial architectures obtained by X-ray crystallography or NMR spectroscopy provide atomic-level detail of interaction interfaces and allow for mechanistic understanding of interaction processes and their functional implications. Additional kinetic, dynamic and structural aspects of protein interactions can be elucidated by electron and atomic force microscopy as well as by fluorescence resonance energy transfer.

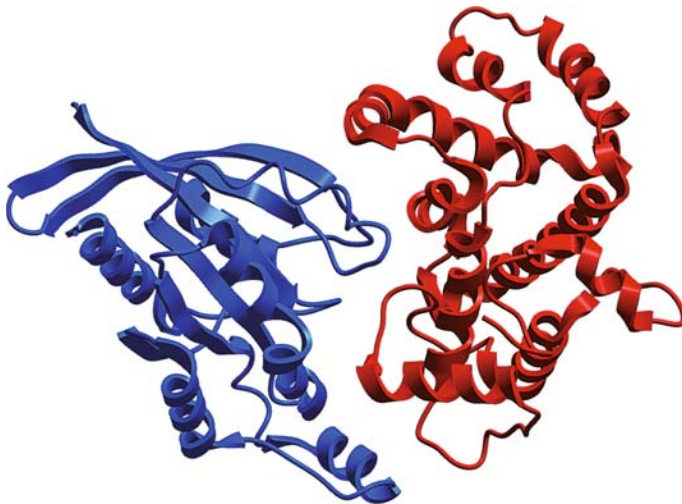


Fig. 1 Structural complex between RhoA, a small GTP protein belonging to the Ras superfamily, and the catalytic GTPase activating domain of RhoGAP (Graham et al. 2002)

3 Protein interaction databases

A huge number of protein–protein interactions has been experimentally determined and described in numerous scientific publications. Public protein interaction databases that provide interaction data in form of structured, machine-readable datasets organized according to well documented standards have become invaluable resources for bioinformatics, systems biology and researchers in experimental laboratories. The data in these databases generally originate from two major sources: large-scale datasets and manually curated information extracted from the scientific literature. As pointed out above, the latter is considered substantially more reliable and large bodies of manually curated PPI data are often used as the gold standard against which predictions and large-scale experiments are benchmarked. Of course, these reference data are far from complete and strongly biased. Many factors, including experimental bias, preferences of the scientific community, and perceived biomedical relevance influence the chance of an interaction to be studied, discovered and published. In the manual annotation process it is not enough to simply record the interaction as such. Additional information such as the type of experimental evidence, citations of the source, experimental conditions, and more need to be stored in order to convey a faithful picture of the data. Annotation is a highly labor intensive task carried out by specially trained database curators.

PPI databases can be roughly divided in two classes: specialized databases focusing on a single organism or a small set of species and general repositories which aim for a comprehensive representation of current knowledge. While the former are often well integrated with other information resources for the same organism, the latter strive for collecting all available interaction data including datasets from specialized resources. The size of these databases is growing constantly as more and more protein interactions are identified. As of writing (November 2007), global repositories are approaching 200,000 pieces of evidence for protein interactions in various species.

All of these databases offer convenient web interfaces that allow for interactively searching the database. In addition, the full datasets are usually provided for download in order to enable researchers to use the data in their own computational analyses. Table 1 gives an overview of some important PPI databases.

4 Data standards for molecular interactions

Until relatively recently, molecular interaction databases like the ones listed in Table 1 acted largely independently from each other. While they provided an extremely valuable service to the community in collecting and curating available molecular interaction data from the literature, they did so largely in an uncoordinated manner. Each database had its own curation policy, feature set, and data formats. In 2002, the Proteomics Standards Initiative (PSI), a work group of the Human Proteome Organization (HUPO), set out to

Table 1 A selection of protein–protein interaction databases

Name	Focus	URL	Reference
BioGrid	global	www.thebiogrid.org	(Stark et al. 2006)
BIND/BOND	global	bond.unleashedinformatics.com	(Bader et al. 2003)
DIP	global	dip.doe-mbi.ucla.edu	(Salwinski et al. 2004)
IntAct	global	www.ebi.ac.uk/intact/	(Kerrien et al. 2007a)
MINT	global	mint.bio.uniroma2.it	(Chatr-aryamontri et al. 2007)
HPRD	Human	www.hprd.org	(Mishra et al. 2006)
IM	<i>D. melanogaster</i> , <i>C. jejunii</i>	proteome.wayne.edu/PIMdb.html	(Pacifico et al. 2006)
MPact/MIPS	<i>S. cerevisiae</i>	mips.gsf.de/genre/proj/mpact/	(Guldener et al. 2006)
MPPI	Mammals	mips.gsf.de/proj/ppi/	(Pagel et al. 2005)

improve this situation, with contributions from a broad range of academic and commercial organizations, among them BIND, Cellzome, DIP, GlaxoSmithKline, Hybrigenics SA, IntAct, MINT, MIPS, Serono, and the Universities of Bielefeld, Bordeaux, and Cambridge. In a first step, a community standard for the representation of protein–protein interactions was developed, the PSI MI format 1.0 (Hermjakob et al. 2004). Recently, version 2.5 of the PSI MI format has been published (Kerrien et al. 2007b), extending the scope of the format from protein–protein interactions to molecular interactions in general, allowing to model for example protein–RNA complexes.

The PSI MI format is a flexible XML format representing the interaction data to a high level of detail. N-ary interactions (complexes) can be represented as well as experimental conditions and technologies, quantitative parameters and interacting domains. The XML format is accompanied by detailed controlled vocabularies in OBO format (Harris et al. 2004). These vocabularies are essential for standardizing not only the syntax, but also the semantics of the molecular interaction representation. As an example, the “yeast two-hybrid technology” described above is referred to in the literature using many different synonyms, for example Y2H, 2H, “yeast-two-hybrid”, etc. While all of these terms refer to the same technology, filtering interaction data from multiple different databases based on this set of terms is not trivial. Thus, the PSI MI standard provides a set of now more than 1000 well-defined terms relevant to molecular interactions. Figure 2 shows the IntAct advanced search tool with a branch of the hierarchical PSI MI controlled vocabulary. Figure 3 provides a partial graphical representation of the annotated XML schema, combined with an example dataset in PSI MI XML format, reprinted from Kerrien et al. (2007b).

For user-friendly distribution of simplified PSI data to end users, the PSI MI 2.5 standard also defines a simple tabular representation (MITAB), derived from the BioGrid format (Breitkreutz et al. 2003). While this format necessarily excludes details

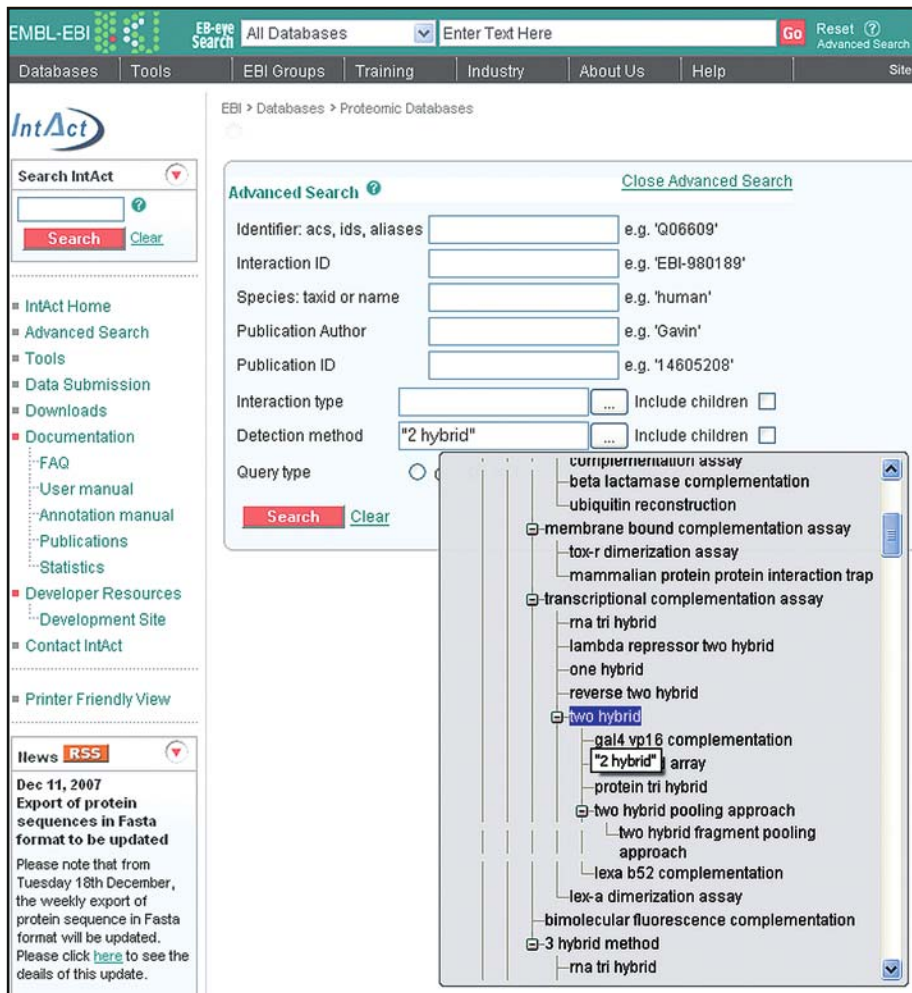


Fig. 2 IntAct advanced search

of interaction data like interacting domains, it provides a means to efficiently access large numbers of basic binary interaction records.

The PSI MI format is now widely implemented, with data available from BioGrid, DIP, HPRD, IntAct, MINT, and MIPS, among others. Visualization tools like Cytoscape (Shannon et al. 2003) can directly read and visualize PSI MI formatted data. Comparative and integrative analysis of interaction data from multiple sources has become easier, as has the development of analysis tools which do not need to provide a plethora of input parsers any more. The annotated PSI MI XML schema, a list of tools and

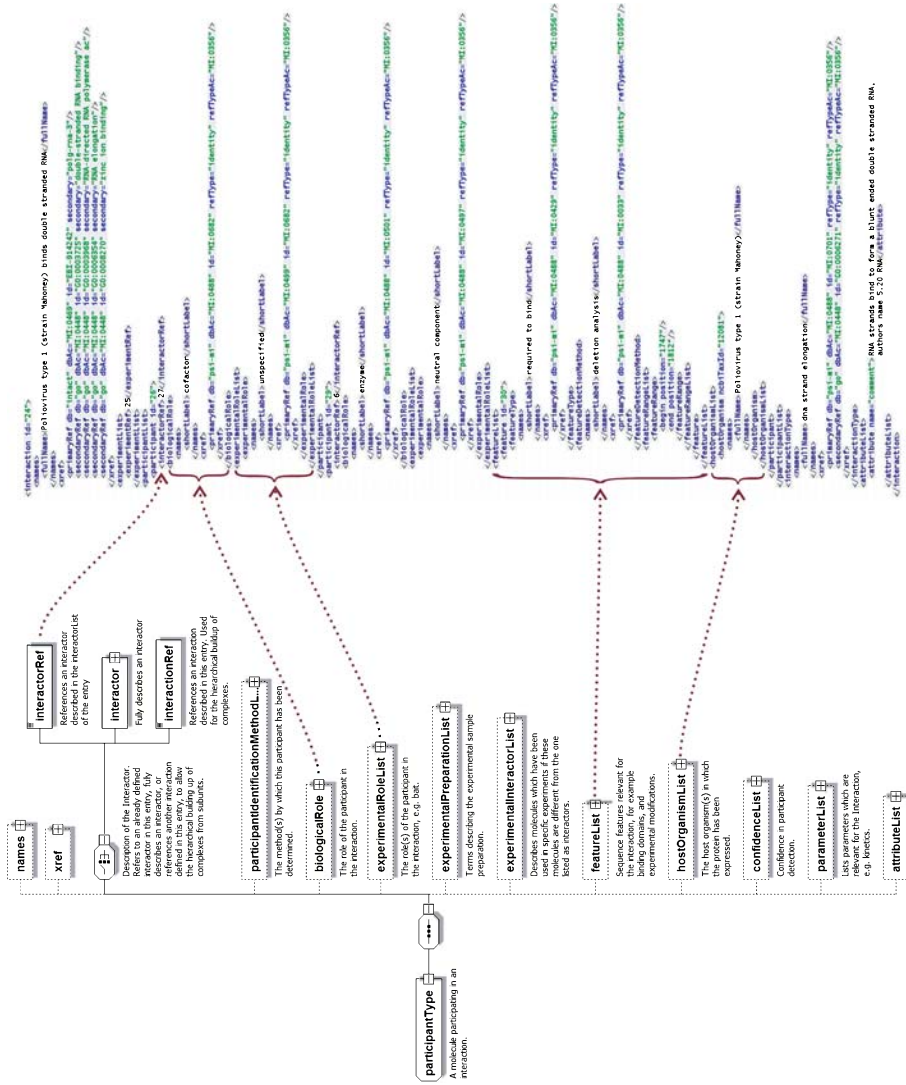


Fig. 3 Partial graphical representation of the annotated PSI MI XML schema, combined with an example dataset in PSI MI XML format (reprinted from Kerrien et al. (2007b))

databases implementing it, as well as further information, are available from <http://www.psidev.info/>.

However, the development and implementation of a common data format is only one step towards the provision of consistent molecular interaction data to the scientific community. Another key step is the coordination of the data curation process itself between different molecular interaction databases. Without such synchronization, independent databases will often work on the same publications and insert the data into their systems, according to different curation rules, thus doing redundant work on some publications, while neglecting others. Recognizing this issue, the DIP, IntAct, and MINT molecular interaction databases are currently synchronizing their curation efforts in the context of the IMEx consortium (<http://imex.sf.net>). These databases are now applying the same curation rules to provide a consistent high level of curation quality, and are synchronizing their fields of activity, each focusing on literature curation from a non-overlapping set of scientific journals. For these journals, the databases aim to insert all published interactions into the database shortly after publication. Regular data exchange of all newly curated data between IMEs databases is currently in the implementation phase.

To support the systematic representation and capture of relevant molecular interaction data supporting scientific publications, the HUPO Proteomics Standards Initiative has recently published “The minimum information required for reporting a molecular interaction experiment (MIMIx)” (Orchard et al. 2007b), detailing data items considered essential for the authors to provide, as well as a practical guide to efficient deposition of molecular interaction data in IMEx databases (Orchard et al. 2007a).

The IMEx databases are also collaborating with scientific journals and funding agencies, to increasingly recommend data producers to deposit their data in an IMEx partner database prior to publication. Database deposition prior to publication not only ensures public availability of the data at the time of publication, but also provides important quality control, as database curators often assess the data in much more detail than reviewers. The PSI journal collaboration efforts are starting to show first results. Nature Biotechnology, Nature Genetics, and Proteomics are now recommending that authors deposit molecular interaction data in a relevant public domain database prior to publication, a key step to a better capture of published molecular interaction data in public databases, and to overcome the current fragmentation of molecular interaction data.

5 The IntAct molecular interaction database

As an example of a molecular interaction database implementing the PSI MI 2.5 standard, we will provide a more detailed description of the IntAct molecular interaction database (Kerrien et al. 2007a), accessible at <http://www.ebi.ac.uk/intact>. IntAct

is a curated molecular interaction database active since 2002. IntAct follows a full text curation policy, publications are read in full by the curation team, and all molecular interactions contained in the publication are inserted into the database, containing basic facts like the database accession numbers of the proteins participating in an interaction, but also details like experimental protein modifications, which can have an impact on assessments of confidence in the presence or absence of interactions. Each database record is cross-checked by a senior curator for quality control. On release of the record, the corresponding author of the publication is automatically notified (where an email address is available), and requested to check the data provided. Any corrections are usually inserted into the next weekly release. While such a detailed, high quality approach is slow and limits coverage, the provision of high quality reference datasets is an essential service both for biological analysis, and for the training and validation of automatic methods for computational prediction of molecular interactions.

As it is impossible for any single database, or even the collaborating IMEx databases, to fully cover all published interactions, curation priorities have to be set. Any direct data depositions supporting manuscripts approaching peer review have highest priority. Next, for some journals (currently Cell, Cancer Cell, and Proteomics) IntAct curates all molecular interactions published in the journal. Finally, several special curation topics are determined in collaboration with external communities or collaborators, where IntAct provides specialized literature curation and collaborates in the analysis of experimental datasets, for example around a specific protein of interest (Camargo et al. 2006).

As of November 2007, IntAct contains 158,000 binary interactions supported by ca. 3,000 publications. The IntAct interface implements a standard “simple search” box, ideal for search by UniProt protein accession numbers, gene names, species, or PubMed identifiers. The advanced search tool (Fig. 2) provides field-specific searches as well as a specialized search taking into account the hierarchical structure of controlled vocabularies. A default search for the interaction detection method “2 hybrid” returns 30,251 interactions, while a search for “2 hybrid” with the tickbox “include children” activated returns more than twice that number, 64,589 interactions. The hierarchical search automatically includes similarly named methods like “two hybrid pooling approach”, but also “gal4 vp16 complement”. Search results are initially shown in a tabular form based on the MITAB format, which can also be directly downloaded. Each pairwise interaction is only listed once, with all experimental evidence listed in the appropriate columns. The final column provides access to a detailed description of each interaction as well as a graphical representation of the interaction in its interaction neighborhood graph. For interactive, detailed analysis, interaction data can be loaded into tools like Cytoscape (see below) via the PSI 2.5 XML format.

All IntAct data is freely available via the web interface, for download in PSI MI tabular or XML format, and computationally accessible via web services. IntAct software is open source, implemented in Java, with Hibernate (www.hibernate.org/)

for the object-relational mapping to OracleTM or Postgres, and freely available under the Apache License, version 2 from <http://www.ebi.ac.uk/intact>.

6 Interaction networks

On a global scale, protein–protein interactions participate in the formation of complex biological networks which, to a large extent, represent the paths of communication and metabolism of an organism. These networks can be modeled as graphs making them amenable to a large number of well established techniques of graph theory and social network analysis. Even though interaction networks do not directly encode cellular processes nor provide information on dynamics, they do represent a first step towards a description of cellular processes, which is ultimately dynamic in nature. For instance, protein–interaction networks may provide useful information on the dynamics of complex assembly or signaling. In general, investigating the topology of protein interaction, metabolic, signaling, and transcriptional networks allows researchers to reveal the fundamental principles of molecular organization of the cell and to interpret genome data in the context of large-scale experiments. Such analyses have become an integral part of the genome annotation process: annotating genomes today increasingly means annotating networks.

A *protein–protein interaction network* summarizes the existence of both stable and transient associations between proteins as an (undirected) graph: each protein is represented as a node (or vertex), an edge between two proteins denotes the existence of an interaction. Interactions known to occur in the actual cell (Fig. 4a) can thus be represented as an abstract graph of interaction capabilities (Fig. 4b). As such a graph is limited by definition to binary interactions, its construction from a database of molecular interactions may involve arbitrary choices. For instance, an n-ary interaction measured by co-purification can be represented using either the clique (all binary interactions between the n proteins are retained) or the spoke model (only edges connecting the “captured” protein to co-purified proteins are retained).

Once a network has been reconstructed from protein interaction data, a variety of statistics on network topology can be computed, such as the distribution of vertex degrees, the distribution of the clustering coefficient and other notions of density, the distribution of shortest path length between vertex pairs, or the distribution of network motifs occurrences (see (Barabasi and Oltvai 2004) for a review). These measures can be used to describe networks in a concise manner, to compare, group or contrast different networks, and to identify properties characteristic of a network or a class of network under study. Some topological properties may be interpreted as ‘traces’ of underlying biological mechanisms, shedding light on their dynamics, their evolution, or both and helping connect structure to function (see the “Network Modules” section below). For instance, most interaction networks seem to exhibit scale-free topology (Jeong et al.

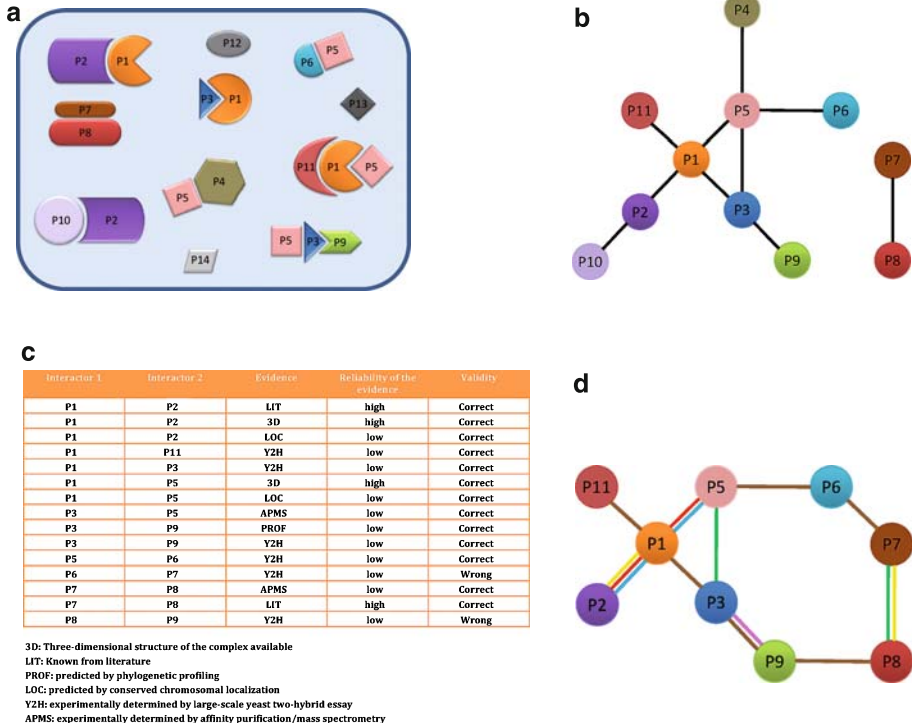


Fig. 4 Graph representation of interaction networks. (a) Hypothetical protein interactions in the living cell. Interacting proteins are denoted as P1, P2, etc. (b) A graph representation of the protein interactions shown in a. Each node represents a protein, and each edge connects proteins that interact. (c) Information on protein interactions obtained by different methods. (d) Protein interaction network derived from experimental evidence shown in c. As in a, each node is a protein, and edges connect interactors. Edges are colored according to the source of evidence: red – 3D, green – APMS, brown – Y2H, magenta – PROF, yellow – LIT, blue – LOC

2001; Yook et al. 2004), i.e. their degree distribution (the probability that a node has exactly k links) approximates a power law $P(k) \sim k^{-\gamma}$, meaning that most proteins have few interaction partners but some, the so-called “hubs”, have many.

As an example of derived evolutionary insight, it is easy to show that networks evolving by growth (addition of new nodes) and preferential attachment (new nodes are more likely to be connected to nodes with more connections) will exhibit scale-free topology (degree distribution approximates a power-law) and hubs (highly connected nodes). A simple model of interaction network evolution by gene duplication, where a duplicate initially keeps the same interaction partners as the original, generates preferential attachment, thus providing a candidate explanation for the scale-free nature and the existence of hubs in these networks (Barabasi and Oltvai 2004).

A corresponding functional interpretation of hubs and scale-free topology has been proposed in terms of robustness. Scale-free networks are robust to component failure, as random failures are likely to affect low degree nodes and only failures affecting hub nodes will significantly change the number of connected components and the length of shortest paths between node pairs. Deletion analyses have, perhaps unsurprisingly, confirmed that highly connected proteins are more likely to be essential (Winzeler et al. 1999; Giaever et al. 2002; Gerdes et al. 2003).

Most biological interpretations that have been proposed for purely topological properties of interaction networks have been the subject of heated controversies, some of which remain unsolved to this day (e.g. (He and Zhang 2006; Yu et al. 2007) on hubs). One often cited objection to any strong interpretation is the fact that networks reconstructed from high-throughput interaction data constitute very rough approximations of the “real” network of interactions taking place within the cell. As illustrated in Fig. 4c, interaction data used in a reconstruction typically result from several experimental methods, often complemented with prediction schemes. Each specific method can miss real interactions (false negatives) and incorrectly identify other interactions (false positives), resulting in biases that are clearly technology-dependent (Gavin et al. 2006; Legrain and Selig 2000). Assessing false-negative and false-positive rates is difficult since there is no ‘gold standard’ for positive interactions (protein pairs that are known to interact) or, more importantly, for negative interactions (protein pairs that are known not to interact). Using less-than-ideal benchmark interaction sets, estimates of 30-60% false positives and 40-80% false negatives have been proposed for yeast-two-hybrid and co-purification based techniques (Aloy and Russell 2004). In particular, a comparison of several high-throughput interaction datasets on yeast, showing low overlap, has confirmed that each study covers only a small percentage of the underlying interaction network (von Mering et al. 2002) (see also “Estimates of the number of protein interactions” below).

Integration of interaction data from heterogeneous sources towards interaction network reconstruction can help compensate for these limitations. The basic principle is fairly simple and rests implicitly on a multigraph representation: several interaction networks to be integrated, each resulting from a specific experimental or predictive method, are defined over the same set of proteins. Integration is achieved by merging them into a single network with several types of links – or edge colors – each drawn from one of the component networks. Some edges in the multigraph may be incorrect, while some existing interactions may be missing from the multigraph, but interactions confirmed independently by several methods can be considered reliable. Figure 4d shows the multigraph that corresponds to the evidence from Fig. 4c and can be used to reconstruct the actual graph in Fig. 4b.

In practice, integration is not always straightforward: networks are usually defined over subsets of the entire gene or protein complement of a species, and meaningful integration requires that the overlap of these subsets be sufficiently large.

In addition, if differences of reliability between network types are to be taken into account, an integrated reliability scoring scheme needs to be designed (Jansen et al. 2003; von Mering et al. 2007) with the corresponding pitfalls and level of arbitrariness involved in comparing apples and oranges. Existing methods can significantly reduce false positive rates on a subset of the network, yielding a subnetwork of high-reliability interactions.

7 Visualization software for molecular networks

The tremendous amounts of available molecular interaction data raise the important issue of how to visualize them in a biologically meaningful way. A variety of tools have been developed to address this problem; two prominent examples are VisANT (Hu et al. 2005) and Cytoscape (Shannon et al. 2003). A recent review of further network visualization tools is provided by Suderman and Hallett (2007). In this section, we focus on Cytoscape (<http://www.cytoscape.org>) and demonstrate its use for the investigation of protein–protein interaction networks. For a more extensive protocol on the usage of Cytoscape, see (Cline et al. 2007).

Cytoscape is a stand-alone Java application that is available for all major computer platforms. This software provides functionalities for (i) generating biological networks, either manually or by importing interaction data from various sources, (ii) filtering interactions, (iii) displaying networks using graph layout algorithms, (iv) integrating and displaying additional information like gene expression data, and (v) performing analyses on networks, for instance, by calculating topological network properties or by identifying functional modules.

One advantage of Cytoscape over alternative visualization software applications is that Cytoscape is released under the open-source Lesser General Public License (LGPL). This license basically permits all forms of software usage and thus helps to build a large user and developer community. Third-party Java developers can easily enhance the functionality of Cytoscape by implementing own plug-ins, which are additional software modules that can be readily integrated into the Cytoscape platform. Currently, there are more than forty plug-ins publicly available, with functionalities ranging from interaction retrieval and integration across topological network analysis, detection of network motifs, protein complexes, and domain interactions, to visualization of subcellular protein localization and bipartite networks. A selection of popular Cytoscape plug-ins is listed in Table 2. In the following, we will describe the functionalities of Cytoscape in greater detail.

The initial step of generating a network can be accomplished in different ways. First, the user can import interaction data that are stored in various flat file or XML formats such as BioPax, SBML, or PSI-MI, as described above. Second, the user can directly retrieve interactions from several public repositories from within Cytoscape. A number

Table 2 Brief descriptions of popular Cytoscape plug-ins with web links to their project sites

Plug-in	Description	Project web site
Agilent Literature Search	Network generation based on text-mining of scientific publications	http://cytoscape.org/plugins/
APID2NET	Network generation and analysis based on the Agile Protein Interaction DataAnalyzer (APID)	http://bioinfow.dep.usal.es/apid/apid2net.html
BiLayout	Generation of bipartite network layouts	http://bilayout.bioinf.mpi-inf.mpg.de/
BiNGO	Determination of overrepresented Gene Ontology (GO) terms	http://www.psb.ugent.be/cbd/papers/bingo/
BiNoM	Manipulation of networks represented in standardized formats like SBML and BioPAX	http://bioinfo-out.curie.fr/projects/binom/
BubbleRouter	Incremental layout generation based on various attributes	http://www.genmapp.org/BubbleRouter/manual.htm
CABIN	Exploratory analysis and integration of multiple interaction networks	http://www.sysbio.org/capabilities/compbio/cabin.stm
Cerebral	Layout generation based on subcellular protein localizations	http://www.pathogenomics.ca/cerebral/
DomainGraph	Decomposition of protein networks into domain-domain interaction networks	http://domaingraph.bioinf.mpi-inf.mpg.de
Enhanced Search	Sophisticated search functionality within a network	http://conklinwolf.ucsf.edu/genmappwiki/Google_Summer_of_Code_2007/Maital
GenePro	Analysis of functional modules and clusters	http://genepro.ccb.sickkids.ca/
GOlorize	Network visualization based on Gene Ontology (GO) categories (only in combination with BiNGO plug-in)	http://www.pasteur.fr/recherche/unites/Biolsys/GOlorize/
GroupTool	Combination of nodes and edges into groups	http://www.rbvi.ucsf.edu/Research/cytoscape/
jActiveModules	Determination of expression activated subnetworks and modules	http://cytoscape.org/plugins/
MCODE	Determination of highly connected clusters and putative complexes	http://baderlab.org/Software/mcode
MetaNode-Plugin2	Abstraction of nodes into meta nodes that can be expanded or collapsed	http://www.rbvi.ucsf.edu/Research/cytoscape/
MiMIplugin	Network generation based on the Michigan Molecular Interaction Database (MiMI)	http://mimi.ncibi.org/cytoscape/
MiSink	Network generation based on the Database of Interacting Proteins (DIP)	http://dip.doe-mpi.ucla.edu/dip/Software.cgi
NamedSelection	Temporary storage of node and edge selections	http://www.rbvi.ucsf.edu/Research/cytoscape/
NetworkAnalyzer	Computation of topological network parameters	http://med.bioinf.mpi-inf.mpg.de/networkanalyzer/
StructureViz	Linkage to macromolecular structures and sequences provided by UCSF Chimera	http://www.cgi.ucsf.edu/Research/cytoscape/structureViz/

of plug-ins exists that facilitate querying certain databases for interactions related to specific genes/proteins or species (APID2NET, MiMIplugin, MiSink; Table 2). Third, the user can utilize a text-mining plug-in that builds networks based on associations found in publication abstracts (Agilent Literature Search; Table 2). While these associations are not as reliable as experimentally derived interactions, they can be helpful when the user is investigating species that are not well covered yet in the current data repositories. Fourth, the user can directly create or manipulate a network by manually adding or removing nodes (genes, proteins, domains, etc.) and edges (interactions or relationships). In this way, expert knowledge that is not captured in the available data sets can be incorporated into the loaded network.

Generated networks can be further refined by applying selections and filters in Cytoscape. The user can select nodes or edges by simply clicking on them or framing a selection area. In addition, starting with at least one selected node, the user can incrementally enlarge the selection to include all direct neighbor nodes. Cytoscape also provides even sophisticated search and filter functionality for selecting particular nodes and edges in a network based on different properties; in particular, the Enhanced Search plug-in (Table 2) improves the built-in search functionality of Cytoscape. Filters select all network parts that match certain criteria, for instance, all human proteins or all interactions that have been detected using the yeast two-hybrid system. Once a selection has been made, all selected parts can be removed from the network or added to another network.

The main purpose of visualization tools like Cytoscape is the presentation of biological networks in an appropriate manner. This can usually be accomplished by applying graph layout algorithms. Sophisticated layouts can assist the user in revealing specific network characteristics such as hub proteins or functionally related protein clusters. Cytoscape offers various layout algorithms, which can be categorized as circular, hierarchical, spring-embedded (or force-directed), and attribute-based layouts (Fig. 5). Further layouts can be included using the Cytoscape plug-in architecture, for example, to arrange protein nodes according to their subcellular localization or to their pathways assignments (BubbleRouter, Cerebral; Table 2).

Some layouts may be more effective than others for representing molecular networks of a certain type. The spring-embedded layout, for instance, has the effect of exposing the inherent network structure, thus identifying hub proteins and clusters of tightly connected nodes. It is noteworthy that current network visualization techniques have limitations, for example, when displaying extremely large or dense networks. In such cases, a simple graphical network representation with one node for each interaction partner, as it is initially created by Cytoscape, can obfuscate the actual network organization due to the sheer number of nodes and edges. One potential solution to this problem is the introduction of meta-nodes (MetaNode plug-in; Table 2). A meta-node combines and replaces a group of other nodes. Meta-nodes can be collapsed to increase clarity of the visualization and expanded to increase the level of detail (Fig. 6).

Cytoscape Desktop (Sessions/layouts.cysp)

File Edit View Select Layout Plugins Help

Search: []

Control Panel

Network | Workspace | Editor | Filters

Current Visual Style

Style []

Defaults

Source — Interaction — Target

Visual Mapping Browser

- Edge Visual Mapping
- Edge Color
- Edge Line Style
- Unsaved Properties
- Edge Font Face
- Edge Font Size
- Edge Label
- Edge Label Color
- Edge Label Opacity
- Edge Line Width
- Edge Opacity
- Edge Source Arrow Color
- Edge Source Arrow Opacity
- Edge Source Arrow Shape
- Edge Target Arrow Color
- Edge Target Arrow Opacity
- Edge Target Arrow Shape
- Edge Tooltip
- Node Border Color
- Node Border Opacity
- Node Color
- Node Font Face
- Node Font Size
- Node Height
- Node Label
- Node Label Color
- Node Label Opacity

Control Panel

Visual Mapping Browser

- Edge Visual Mapping
- Edge Color
- Edge Line Style
- Unsaved Properties
- Edge Font Face
- Edge Font Size
- Edge Label
- Edge Label Color
- Edge Label Opacity
- Edge Line Width
- Edge Opacity
- Edge Source Arrow Color
- Edge Source Arrow Opacity
- Edge Source Arrow Shape
- Edge Target Arrow Color
- Edge Target Arrow Opacity
- Edge Target Arrow Shape
- Edge Tooltip
- Node Border Color
- Node Border Opacity
- Node Color
- Node Font Face
- Node Font Size
- Node Height
- Node Label
- Node Label Color
- Node Label Opacity

Workspace

(a) (b) (c) (d) (e) (f)

Toolbar

Visual Mapping Browser

- Edge Visual Mapping
- Edge Color
- Edge Line Style
- Unsaved Properties
- Edge Font Face
- Edge Font Size
- Edge Label
- Edge Label Color
- Edge Label Opacity
- Edge Line Width
- Edge Opacity
- Edge Source Arrow Color
- Edge Source Arrow Opacity
- Edge Source Arrow Shape
- Edge Target Arrow Color
- Edge Target Arrow Opacity
- Edge Target Arrow Shape
- Edge Tooltip
- Node Border Color
- Node Border Opacity
- Node Color
- Node Font Face
- Node Font Size
- Node Height
- Node Label
- Node Label Color
- Node Label Opacity

Data Panel

ID	Function	Localization	Canonical...
MYD88	Adaptor	cytoplasm	MYD88
BTK	Kinase	cytoplasm	BTK
IRAK1	Kinase	cytoplasm	IRAK1
TLR4	Receptor	plasma	TLR4

Node Attribute Browser | Edge Attribute Browser | Network Attribute Browser

Welcome to Cytoscape 3.5

Right-click to drag | Middle-click to drag to PAN

An overview of established and novel visualization techniques for biological networks on different scales is presented in (Hu et al. 2007).

All layouts generated by Cytoscape are zoomable, enabling the user to increase or decrease the magnification, and they can be further customized by aligning, scaling, or rotating selected network parts. Additionally, the user can define the graphical network representation through visual styles. These styles define the colors, sizes, and shapes of all network parts.

A powerful feature of Cytoscape is its ability of visually mapping additional attribute values onto network representations. Both nodes and edges can have arbitrary attributes, for example, protein function names, the number of interactions (node degree), expression values, the strength and type of an interaction, or confidence values for interaction reliability. These attributes can be used to adapt the network illustration by dynamically changing the visual styles of individual network parts (Fig. 7). For example, this feature enables highlighting trustworthy interactions by assigning

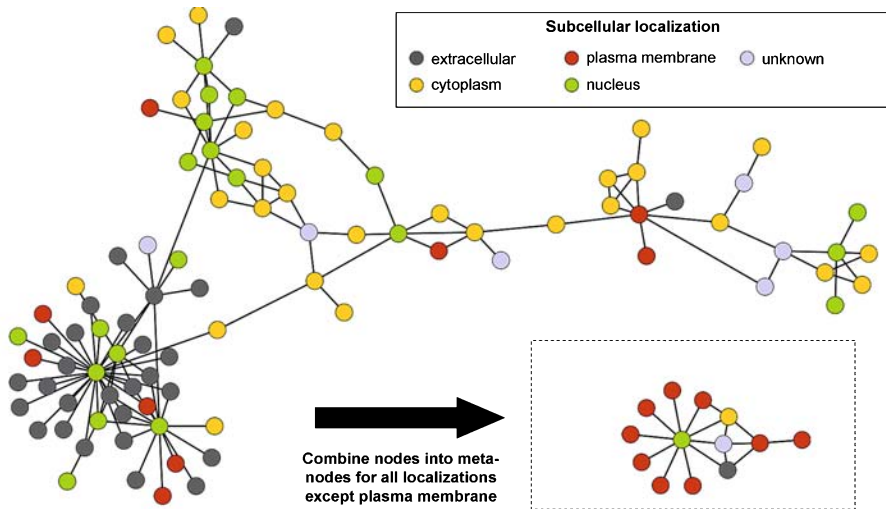


Fig. 6 Combination of nodes into meta-nodes using the Cytoscape plug-in *MetaNode* (Table 2). All protein nodes with subcellular localizations different from plasma membrane are combined into meta-nodes. These meta-nodes can be collapsed or expanded to increase clarity or detailedness, respectively

Fig. 5 The Cytoscape desktop. The workspace (middle) shows six identical networks with different layouts. The toolbar (top) contains basic control buttons for zooming and filtering/searching. The Control Panel (left) displays the VizMapper that defines the graphical network representation. The Data Panel (bottom) lists node attributes of the four selected nodes (yellow) in network (b). The different network layouts are: (a) grid, (b) circular with several circles, (c) spring-embedded or force-directed, (d) circular with one circle, (e) attribute-based, (f) hierarchical

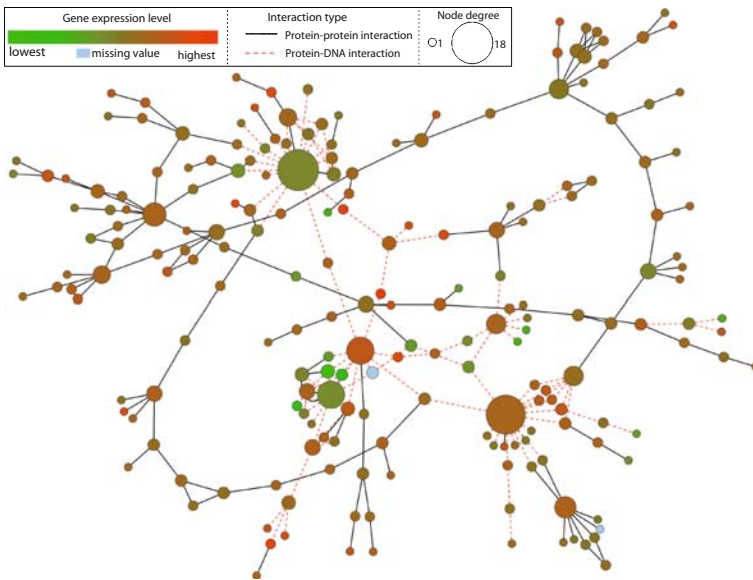


Fig. 7 Visual representation of a subset of the GAL4 network in yeast. The protein nodes are colored with a red-to-green gradient according to their expression value; green represents the lowest, red the highest value, and blue a missing value. The node size indicates the number of interactions (node degree); the larger a node, the higher is its degree. The colors and styles of the edges represent different interaction types; solid black lines represent protein-protein, dashed red lines protein-DNA interactions

different line styles or sizes to different experiment types (discrete mapping of an edge attribute), to spot network hubs by changing the size of a node according to its degree (discrete or continuous mapping of a node attribute), or to identify functional network patterns by coloring protein nodes with a color gradient according to their expression level (continuous mapping of a node attribute). Hence, it is possible to simultaneously visualize different data types by overlaying them with a network model.

In order to generate new biological hypotheses and to gain insights into molecular mechanisms, it is important to identify relevant network characteristics and patterns. For this purpose, the straightforward approach is the visual exploration of the network. Table 2 lists a selection of Cytoscape plug-ins that assist the user in this analysis task, for instance, by identifying putative complexes (MCODE), by grouping proteins that show a similar expression profile (jActiveModules), or by identifying overrepresented GO terms (BiNGO, Golorize). However, the inclusion of complex data such as time-series results or diverse Gene Ontology (GO) terms into the network visualization might not be feasible without further software support. Particularly in case of huge, highly connected, or dynamic networks, more advanced visualization techniques will be required in the future.

In addition to the visual presentation of interaction networks, Cytoscape can also be used to perform statistical analyses. For instance, the NetworkAnalyzer plug-in (Assenov et al. 2008) computes a large variety of topology parameters for all types of networks. The computed simple and complex topology parameters are represented as single values and distributions, respectively. Examples of simple parameters are the number of nodes and edges, the average number of neighbors, the network diameter and radius, the clustering coefficient, and the characteristic path length. Complex parameters are distributions of node degrees, neighborhood connectivities, average clustering coefficients, and shortest path lengths. These computed statistical results can be exported in textual or graphical form and are additionally stored as node attributes. The user can then apply the calculated attributes to select certain network parts or to map them onto the visual representation of the analyzed network as described above (Fig. 7). It is also possible to fit a power law to the node degree distribution, which can frequently indicate a so-called scale-free network with few highly connected nodes (hubs) and many other nodes with a small number of interactions. Scale-free networks are especially robust against failures of randomly selected nodes, but quite vulnerable to defects of hubs (Albert 2005).

8 Estimates of the number of protein interactions

How many PPIs exist in a living cell? The yeast genome encodes approximately 6300 gene products which means that the maximal possible number of interacting protein pairs in this organism is close to 40 million, but what part of these potential interactions are actually realized in nature? For a given experimental method, such as the two-hybrid essay, the estimate of the total number of interactions in the cell is given by

$$N_{\text{int}} = N_{\text{measured}} \times R_{\text{fp}} \times R_{\text{fn}}^{-1}$$

where N_{measured} is the number of interactions identified in the experiment, and R_{fp} and R_{fn} are false positive and false negative rates of the method. R_{fn} can be roughly estimated based on the number of interactions known with confidence (e.g., those confirmed by three-dimensional structures) that are being recovered by the method. Assessing R_{fp} is much more difficult because no experimental information on proteins that *do not* interact is currently available. Since it is known that proteins belonging to the same functional class often interact, one very indirect way of calculating R_{fn} is as the fraction of functionally related proteins not found to be interacting.

An even more monumental problem is the estimation of the total number of unique structurally equivalent interaction *types* existing in nature. An interaction type is defined as a particular mutual orientation of two specific interacting domains. In some cases homologous proteins interact in a significantly different fashion while in other cases proteins lacking sequence similarity engage in interactions of the same type.

In general, however, interacting protein pairs sharing a high degree of sequence similarity (30–40% or higher) between their respective components almost always form structurally similar complexes (Aloy et al. 2003). This observation allows utilization of available atomic resolution structures of complexes for building useful models of closely related binary complexes.

The total number of interaction *types* can then be estimated as follows:

$$N_{\text{types}} = N_{\text{measured}} \times R_{\text{fp}} \times R_{\text{fn}}^{-1} \times C \times E_{\text{All-species}}$$

where the interaction similarity multiplier C reflects the clustering of all interactions of the same type, and $E_{\text{All-species}}$ extrapolates from one biological species to all organisms. Aloy and Russel (2004) derived an estimate for C by grouping interactions between proteins that share high sequence similarity, as discussed above. C depends on the number of paralogous sequences encoded in a given genome. For small prokaryotic organisms it is close to 1 while for larger and more redundant genomes it adopts smaller values, typically in the range of 0.75–0.85. The multiplier for all species $E_{\text{All-species}}$ can be derived by assessing what fraction of known protein families is encoded in a given genome. Based on the currently available data this factor is close to 10 for bacteria, which means that a medium size prokaryotic organism contains around one tenth of all protein families. For eukaryotic organisms $E_{\text{All-species}}$ lies between 2 and 4. For the comprehensive two-hybrid screen of yeast by (Uetz 2000) in which 936 interactions between 987 proteins were identified, Aloy and Russell (2004) estimated C , R_{fp} , and R_{fn}^{-1} , and $E_{\text{All-species}}$ to be 0.85, 3.92, 0.55, and 3.35 respectively, leading to an estimated 1715 different interaction types in yeast alone, and 5741 over all species. Based on the two-hybrid interaction map of the fly (Giot 2003) the number of all interaction types in nature is estimated to be 9962. It is thus reasonable to expect the total number of interaction types to be around 10,000, and only 2000 are currently known.

9 Multi-protein complexes

Beyond binary interactions, proteins often form large molecular complexes involving multiple subunits (Fig. 8). These complexes are much more than a random snapshot of a group of interacting proteins – they represent large functional entities which remain stable for long periods of time. Many such protein complexes have been elucidated step by step over time and recent advances in high-throughput technology have led to large-scale studies revealing numerous new protein complexes. The preferred technology for this kind of experiment is initial co-purification of the complexes followed by the identification of the member proteins by mass spectrometry.

As the baker's yeast *S. cerevisiae* is one of the most versatile model organisms used in molecular biology, it is not surprising that the first large-scale complex datasets were obtained in this species (Gavin et al. 2002; Ho et al. 2002; Gavin et al. 2006; Krogan et al.

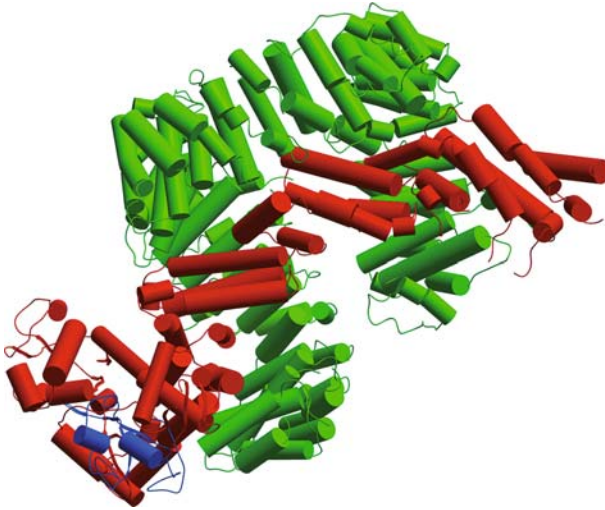


Fig. 8 Ternary complex between the Cand1 protein (green) and the catalytic core of the ubiquitin ligase consisting of cullin (red) and Roc1 (blue) (Goldenberg et al. 2004)

2006). The yeast protein interaction database MPact (Guldener et al. 2006) provides access to 268 protein complexes based on careful literature annotation composed of 1237 different proteins plus over 1000 complexes from large-scale experiments which contain more than 2000 distinct proteins. These numbers contain some redundancy with respect to complexes, due to slightly different complex composition found by different groups or experiments. Nevertheless, the dataset covers about 40% of the *S.cerevisiae* proteome. While many complexes comprise only a small number of different proteins, the largest of them features an impressive 88 different protein species.

A novel manually annotated database, CORUM (Ruepp et al. 2008) contains literature-derived information about 1750 mammalian multi-protein complexes. Over 75% of all complexes contain between three and six subunits, while the largest molecular structure, the spliceosome, consists of 145 components (Fig. 9).

10 Network modules

Modularity has emerged as one of the major organizational principles of cellular processes. Functional modules are defined as molecular ensembles with an autonomous function (Hartwell et al. 1999). Proteins or genes can be partitioned into modules based on shared patterns of regulation or expression, involvement in a common metabolic or regulatory pathway, or membership in the same protein complex or subcellular structure. Modular representation and analysis of cellular processes allows for inter-

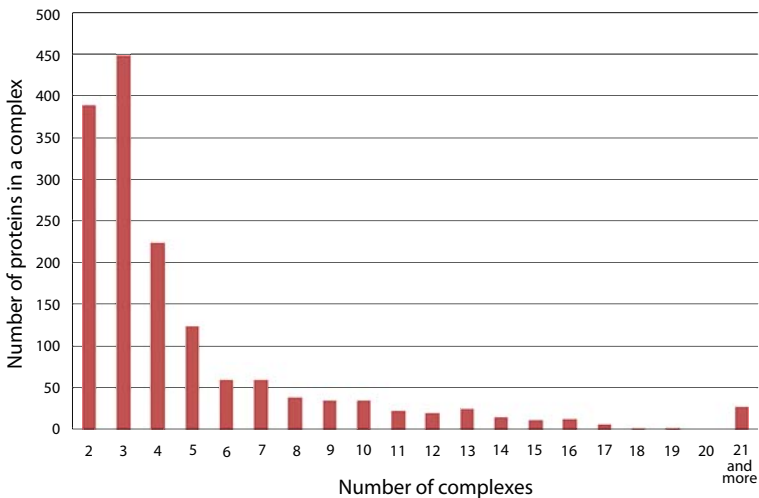


Fig. 9 Number of proteins in the CORUM complexes

pretation of genome data beyond single gene behavior. In particular, analysis of modules provides a convenient framework for studying the evolution of living systems (Snel and Huynen 2004). Multiprotein complexes represent one particular type of functional modules in which individual components engage in physical interactions to execute a specific cellular function.

Algorithmically, modular architectures can be defined as densely interconnected groups of nodes on biological networks (for an excellent review of available methods see (Sharan et al. 2007). Statistically significant functional subnetworks are characterized by a high degree of local clustering. The density of a cluster can be represented as a function $Q(m,n) = 2m/(n(n-1))$, where m is the number of interactions between the n nodes of the cluster (Spirin and Mirny 2003). Q thus takes values between 0 for a set of unconnected nodes and 1 for a fully connected cluster (clique). The statistical significance of Q strongly depends on the size of the graph. It is obvious that random clusters with $Q = 1$ involving just three proteins are very likely while large clusters with $Q = 1$ or even with values below 0.5 are extremely unlikely. In order to compute the statistical significance of a cluster with n nodes and m connections Spirin and Mirny calculate the expected number of such clusters in a comparable random graph and then estimate the likelihood of having m or more interactions within a given set of n proteins given the number of interactions that each of these proteins has. Significant dense clusters identified by this procedure on a graph of protein interactions were found to correspond to functional modules most of which are involved in transcription regulation, cell-cycle/cell-fate control, RNA processing, and protein transport. However, not all of them constitute physical protein complexes and, in general, it is not possible to predict

whether a given module corresponds to a multiprotein complex or just to a group of functionally coupled proteins involved in the same cellular process.

The search for significant subgraphs can be further enhanced by considering evolutionary conservation of protein interactions. With this approach protein complexes are predicted from binary interaction data by network alignment which involves comparing interaction graphs between several species (Sharan et al. 2005). First, proteins are grouped by sequence similarity such that each group contains one protein from each species, and each protein is similar to at least one other protein in the group. Then a composite interaction network is created by joining with edges those pairs of groups that are linked by at least one conserved interaction. Again, dense clusters on such network alignment graph are often indicative of multiprotein complexes.

An alternative computational method for deriving complexes from noisy large-scale interaction data relies on a “socio-affinity” index which essentially reflects the frequency with which proteins form partnerships detected by co-purification (Gavin et al. 2006). This index was shown to correlate well with available three-dimensional structure data, dissociation constants of protein–protein interactions, and binary interactions identified by the two-hybrid techniques. By applying a clustering procedure to a matrix containing the values of the socio-affinity index for all yeast protein pairs found to associate by affinity purification, 491 complexes were predicted, with over a half of them being novel and previously unknown. However, dependent on the analysis parameters distinct complex variants (isoforms) are found that differ from in terms of their subunit composition. Those proteins present in most of the isoforms of a given complex constitute its core while variable components present only in a small number of isoforms can be considered “attachments” (Fig. 10). Furthermore, some stable, typically smaller protein groups can be found in multiple attachments in which case they are

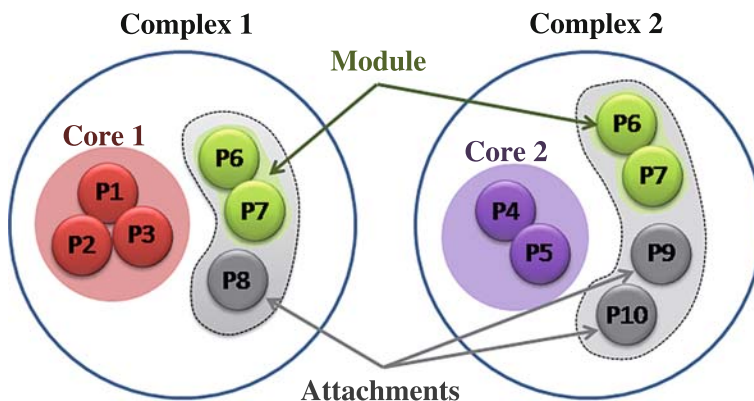


Fig. 10 Definitions of complex cores, attachments, and modules. Redrawn and modified with permission from (Gavin et al. 2006)

called “modules”. Stable functional modules can thus be flexibly used in the cell in a variety of functional contexts. Proteins frequently associated with each other in complex cores and modules are likely to be co-expressed and co-localized.

11 Diseases and protein interaction networks

In this section, we offer a computational perspective on utilizing protein network data for molecular medical research. The identification of novel therapeutic targets for diseases and the development of drugs has always been a difficult, time-consuming and expensive venture (Ruffner et al. 2007). Recent work has charted the current pharmacological space using different networks of drugs and their protein targets (Paolini et al. 2006; Keiser et al. 2007; Kuhn et al. 2008; Yildirim et al. 2007) based on biochemical relationships like ligand binding energy and molecular similarity or on shared disease association. Above all, since many diseases are due to the malfunctioning of proteins, the systematic determination and exploration of the human interactome and homologous protein networks of model organisms can provide considerable new insight into pathophysiological processes (Giallourakis et al. 2005).

Knowledge of protein interactions can frequently improve the understanding of relevant molecular pathways and the interplay of various proteins in complex diseases (Fishman and Porter 2005). This approach may result in the discovery of a considerable number of novel drug targets for the biopharmaceutical industry, possibly affording the development of multi-target combination therapeutics. Observed perturbations of protein networks may also offer a refined molecular description of the etiology and progression of disease in contrast to phenotypic categorization of patients (Loscalzo et al. 2007). Molecular network data may help to improve the ability of cataloging disease unequivocally and to further individualize diagnosis, prognosis, prevention, and therapy. This will require a network-based approach that does not only include protein interactions to differentiate pathophenotypes, but also other types of molecular interactions as found in signaling cascades and metabolic pathways. Furthermore, environmental factors like pathogens interacting with the human host or the effects of nutrition need to be taken into account.

After large-scale screens identified enormous amounts of protein interactions in organisms like yeast, fly, and worm (Goll and Uetz 2007), which also serve as model systems for studying many human disease mechanisms (Giallourakis et al. 2005), experimental techniques and computational prediction methods have recently been applied to generate sizable networks of human proteins (Cusick et al. 2005; Stelzl and Wanker 2006; Assenov et al. 2008; Ramírez et al. 2007). In addition, comprehensive maps of protein interactions inside pathogens and between pathogens and the human host have been compiled for bacteria like *E. coli*, *H. pylori*, *C. jejuni*, and other species (Noirot and Noirot-Gros 2004), for many viruses such as herpes viruses, the Epstein-

Table 3 Selection of pathogenic organisms for which comprehensive protein interaction maps are available

Organism	References
Bacteria	
<i>Escherichia coli</i>	(Butland et al. 2005)
<i>Helicobacter pylori</i>	(Colland et al. 2001)
<i>Campylobacter jejuni</i>	(Parrish et al. 2007)
Viruses	
Herpesvirus family	(Uetz et al. 2006)
Epstein-Barr virus	(Calderwood et al. 2007)
SARS coronavirus	(von Brunn et al. 2007)
HIV-1	(Wheeler et al. 2007)
Hepatitis C virus	(Flajolet et al. 2000)
Parasite	
<i>Plasmodium falciparum</i>	(LaCount et al. 2005)

Barr virus, the SARS coronavirus, HIV-1, the hepatitis C virus, and others (Uetz et al. 2004), and for the malaria parasite *P. falciparum* (Table 3). Those extensive network maps can now be explored to identify potential drug targets and to block or manipulate important protein–protein interactions.

Furthermore, different experimental methods are also used to expand the known interaction networks around pathway-centric proteins like epidermal growth factor receptors (EGFRs) (Tewari et al. 2004; Oda et al. 2005; Jones et al. 2006), Smad and transforming growth factor- β (TGF β) (Colland and Daviet 2004; Tewari et al. 2004; Barrios-Rodiles et al. 2005), and tumor necrosis factor- α (TNF α) and the transcription factor NF- κ B (Bouwmeester et al. 2004). All of these proteins are involved in sophisticated signal transduction cascades implicated in various important disease indications ranging from cancer to inflammation. The immune system and Toll-like receptor (TLR) pathways were the subject of other detailed studies (Oda and Kitano 2006). Apart from that, protein networks for longevity were assembled to research ageing-related effects (Xue et al. 2007).

High-throughput screens are also conducted for specific disease proteins causative of closely related clinical and pathological phenotypes to unveil molecular interconnections between the diseases. For example, similar neurodegenerative disease phenotypes are caused by polyglutamine proteins like huntingtin and over twenty ataxins. Although they that are not evolutionarily related and their expression is not restricted to the brain, they are responsible for inherited neurotoxicity and age-dependent dementia only in specific neuron populations (Ralser et al. 2005). Yeast two-hybrid screens revealed an unexpectedly dense interaction network of those disease proteins forming interconnected subnetworks (Fig. 11), which suggests common pathways affected in disease (Goehler et al. 2004; Lim et al. 2006). Some of the protein–protein interactions may be

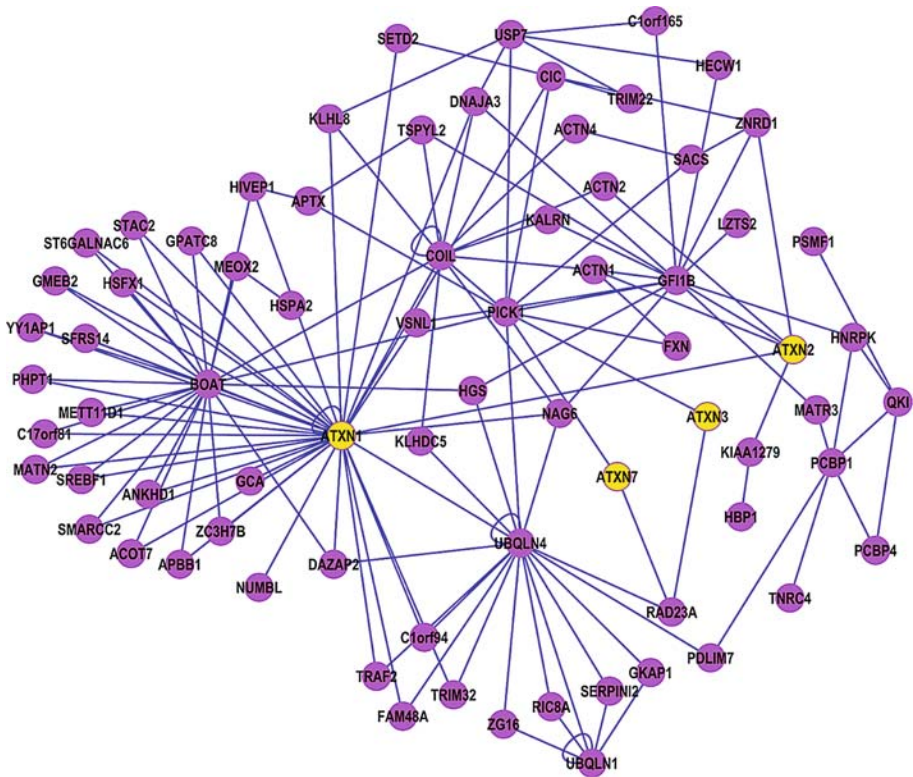


Fig. 11 Part of the protein interaction network around the four yellow-colored ataxins causative of neurodegenerative diseases

involved in mediating neurodegeneration and thus may be tractable for drug inhibition, and several interaction partners of ataxins could additionally be shown to be potential disease modifiers in a fly model (Kaltenbach et al. 2007).

A number of methodological approaches concentrate on deriving correlations between common topological properties and biological function from subnetworks around proteins that are associated with a particular disease phenotype like cancer. Recent studies report that human disease-associated proteins with similar clinical and pathological features tend to be more highly connected among each other than with other proteins and to have more similar transcription profiles (Gandhi et al. 2006; Xu and Li 2006; Goh et al. 2007). This observation points to the existence of disease-associated functional modules. Interestingly, in contrast to disease genes, essential genes whose defect may be lethal early on in life are frequently found to be hubs central to the network.

Further work focused on specific disease-relevant networks. For instance, to analyze experimental asthma, differentially expressed genes were mapped onto a protein

interaction network (Lu et al. 2007). Here, highly connected nodes tended to have smaller expression changes than peripheral nodes. This agrees with the general notion that disease-causing genes are typically not central in the network. Similarly, a comprehensive protein network analysis of systemic inflammation in human subjects investigated blood leukocyte gene expression patterns when receiving an inflammatory stimulus, a bacterial endotoxin, to identify functional modules perturbed in response to this stimulus (Calvano et al. 2005). Topological criteria and gene expression data were also used to search protein networks for functional modules that are relevant to type 2 diabetes mellitus (Liu et al. 2007) or to different types of cancer (Jonsson and Bates 2006; Cui et al. 2007; Lin et al. 2007; Pujana et al. 2007). Moreover, it was recently demonstrated that the integration of gene expression profiles with subnetworks of interacting proteins can lead to improved prognostic markers for breast cancer outcome that are more reproducible between patient cohorts than sets of individual genes selected without network information (Chuang et al. 2007).

In drug discovery, protein networks can help to design selective inhibitors of protein-protein interactions which target specific interactions of a protein, but do not affect others (Wells and McClendon 2007). For example, a highly connected protein (hub) may be a suitable target for an antibiotic whereas a more peripheral protein with few interaction partners may be more appropriate for a highly specific drug that needs to avoid side effects. Thus, topological network criteria are not only useful for characterizing disease proteins, but also for finding drug targets. The diversity of interactions of a targeted protein could also help in predicting potential side effects of a drug. Apart from that, it is remarkable that some potential drugs have been found to be less effective than expected due to the intrinsic robustness of living systems against perturbations of molecular interactions (Kitano 2007). Furthermore, mutations in proteins cause genetic diseases, but it is not always easy to distinguish protein interactions impaired by mutated binding sites from other disease causes like structural instability induced by amino acid mutations.

Nowadays many genome-wide association and linkage studies for human diseases suggest genomic loci and linkage intervals that contain candidate genes encoding SNPs and mutations of potential disease proteins (Kann 2007). Since the resultant list of candidates frequently contain dozens or even hundreds of genes, computational approaches have been developed to prioritize them for further analyses and experiments. In the following, we will demonstrate the variety of available prioritization approaches by explicating three recent methods that utilize protein interaction data in addition to the inclusion of other sequence and function information. All methods capitalize on the above described observation that closely interacting gene products often underlie polygenic diseases and similar pathophenotypes (Oti and Brunner 2007).

Using protein-protein interaction data annotated with reliability values, Lage et al. (2007) first predict human protein complexes for each candidate protein. They then score the pairwise phenotypic similarity of the candidate disease with all proteins within each complex that are associated with any disease. The scoring function basically

measures the overlap of the respective disease phenotypes as recorded in text entries of OMIM (Online Mendelian Inheritance in Man) (Hamosh et al. 2005) based on the vocabulary of UMLS (Unified Medical Language System) (Bodenreider 2004). Lastly, all candidates are prioritized by the probability returned by a Bayesian predictor trained on the interaction data and phenotypic similarity. Therefore, this method depends on the premise that the phenotypic effects caused by any disease-affected member in a predicted protein complex are very similar to each other.

Another prioritization approach by Franke et al. (2006) does not make use of overlapping disease phenotypes and primarily aims at connecting physically disjoint genomic loci associated with the same disease using molecular networks. At the beginning, their method Prioritizer performs a Bayesian integration of three different network types of gene/protein relationships. The latter are derived from functional similarity using Gene Ontology annotation, microarray coexpression, and protein–protein interaction. This results in a probabilistic human network of general functional links between genes. Prioritizer then assesses which candidate genes contained in different disease loci are closely connected in this gene-gene network. To this end, the score of each candidate is initially set to zero, but it is increased iteratively during network exploration by a scoring function that depends on the network distance of the respective candidate gene to candidates inside another genomic loci. This procedure finally yields separate prioritization lists of ranked candidate genes for each genomic loci.

In contrast to the integrated gene-gene network used by Prioritizer, the Endeavour system (Aerts et al. 2006) directly compares candidate genes with known disease genes and creates different ranking lists of all candidates using various sources of evidence for annotated relationships between genes or proteins. The evidence can be derived from literature mining, functional associations based on Gene Ontology annotations, co-occurrence of transcriptional motifs, correlation of expression data, sequence similarity, common protein domains, shared metabolic pathway membership, and protein–protein interactions. At the end, Endeavour merges the resultant ranking lists using order statistics and computes an overall prioritization list of all candidate genes.

Finally, it is important to keep in mind that current datasets of human protein interactions may still contain a significant number of false interactions and thus biological and medical conclusions derived from them should always be taken with a note of caution, in particular, if no good confidence measures are available.

12 Sequence-based prediction of protein interactions

A comprehensive atlas of protein interactions is fundamental for a better understanding of the overall dynamic functioning of the living organisms. These insights arise from the integration of functional information, dynamic data and protein interaction networks. In order to fulfill the goal of enlarging our view of the protein interaction network,

several approaches must be combined and a crosstalk must be established among experimental and computational methods. This has become clear from comparative evaluations which show similar performances for both types of methodologies. In fact, over the recent years this field has grown into one of the most appealing fields in bioinformatics. Evolutionary signals result from restrictions imposed by the need to optimize the features that affect a given interaction and the nature of these features can differ from interaction to interaction. Consequently, a number of different methods have been developed based a range of different evolutionary signals. This section is devoted to a brief review of some of these methods.

12.1 Phylogenetic profiling

These techniques are based on the similarity of absence/presence profiles of interacting proteins. In its original formulation (Gaasterland and Ragan 1998; Huynen and Bork 1998; Pellegrini et al. 1999; Marcotte et al. 1999a) the phylogenetic profiles were codified as 0/1 vectors for each reference protein according to the absence/presence of proteins of the studied family in a set of fully sequenced organisms (see Fig. 12a). The vectors for different reference sequences are compared by using the Hamming distance (Pellegrini et al. 1999) between vectors. This measure counts the number of differences between two binary vectors. The rationale for this method is that both interacting proteins must be present in an organism and that reductive evolution will remove unpaired proteins in the rest of the organisms. Proposed improvements include the inclusion of quantitative measures of sequence divergence (Marcotte et al. 1999b; Date and Marcotte 2003) and the ability to deal with biases in the taxonomic distribution of the organisms used (Date and Marcotte 2003; Barker and Pagel 2005). These biases are due to the intuitive fact that evolutionarily similar organisms will share a higher number of protein and genomic features (in this case presence/absence of an orthologue).

To reduce this problem, Date et al. used Mutual Information from sequence divergent profiles for measuring the amount of information shared by both vectors. Mutual Information is calculated as:

$$MI(P1, P2) = H(P1) + H(P2) - H(P1, P2),$$

where $H(P1) = -\sum p(P1) \ln p(P1)$ is the marginal entropy of the probability distribution of protein P1 sequence distances and $H(P1, P2) = -\sum \sum p(P1, P2) \ln p(P1, P2)$ is the joint entropy of the probability distributions of both protein P1 and P2 sequence distances. The corresponding probabilities are calculated from the whole distribution of orthologue distances for the organisms. In this way, the most likely evolutionary distances between orthologues from a pair of organisms will produce smaller entropies and consequently smaller values of Mutual Information. This formulation should implicitly reduce the effect of taxonomic biases. In an interesting work, published recently by Barker et al. (2007), the authors showed that detection of correlated gene-

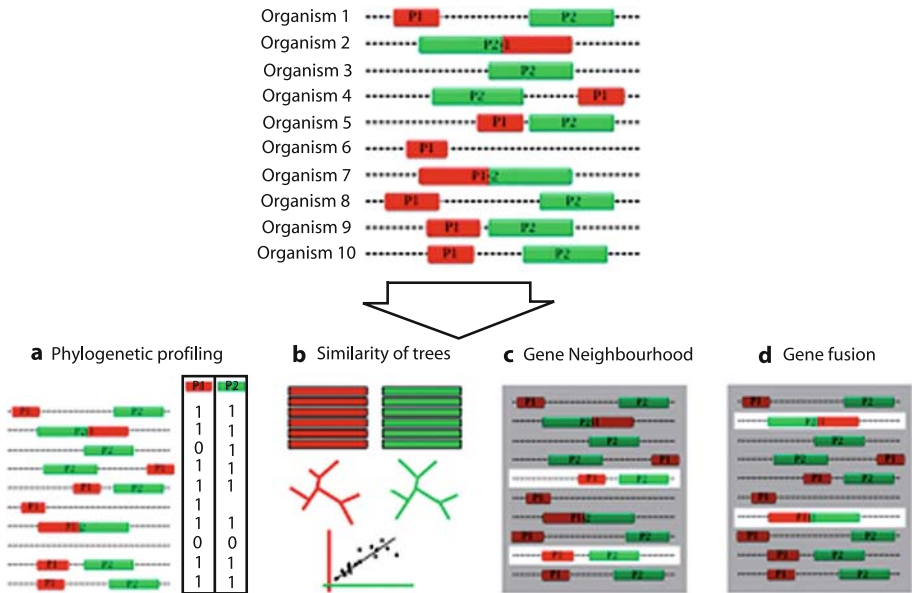


Fig. 12 Prediction of protein interactions based on genomic and sequence features. Information coming from the set of close homologs of the proteins P1 and P2 from the Organism 1 in other organisms can be used to predict an interaction between these proteins. (a) Phylogenetic profiling. Presence/absence of a homolog of both proteins in different organisms is coded as the corresponding two ‘1/0’ profiles (most simple approach) and an interaction is predicted for very similar profiles. (b) Similarity of phylogenetic trees. Multiple sequence alignments are built for both sets of proteins and phylogenetic trees are derived from the proteins with a possible partner present in its organism. Proteins with highly similar trees are predicted to interact. (c) Gene neighbourhood conservation. Genome closeness is checked for those genes coding for both sets of homologous proteins. Interaction is predicted if gene pairs are recurrently close to each other in a number of organisms. (d) Gene fusion. Finding the proteins containing different sequence regions homologous to each of the two proteins is used to predict an interaction between them

gain/gene-loss events improves the predictions by reducing the number of false positives due to taxonomic biases.

The phylogenetic profiling approach has been shown to be quite powerful, because its simple formulation has allowed the exploration of a number of alternative interdependencies between proteins. This is the case for enzyme “displacement” in metabolic pathways detected as anti-correlated profiles (Morett et al. 2003), and for complex dependence relations among triplets of proteins (Bowers et al. 2004). Phylogenetic profiles have also been correlated with bacterial traits to predict the genes related to particular phenotypes (Korbel et al. 2005). The main drawbacks of these methods are the difficulty of dealing with essential proteins (where there is no absence information) and the requirement for the genomes under study to be complete (to establish the absence of a family member).

12.2 Similarity of phylogenetic trees

Similarity in the topology of phylogenetic trees of interacting proteins has been qualitatively observed in a number of cases (Fryxell 1996; Pages et al. 1997; Goh et al. 2000). The extension of this observation to a quantitative method for the prediction of protein interactions requires measuring the correlation between the similarity matrices of the explored pairs of protein families (Goh et al. 2000). This formulation allows systematic evaluation of the validity of using the original observation as a signal of protein interaction (Pazos and Valencia 2001).

The general protocol for these methods is illustrated in Fig. 12b. It includes the building of the multiple sequence alignment for the set of orthologues (one per organism) related to every query sequence, the calculation of all protein pair evolutionary distances (derived from the corresponding phylogenetic trees) and finally the comparison of evolutionary distance matrices of pairs of query proteins using Pearson's correlation coefficient. Protein pairs with highly correlated distance matrices are predicted to be more likely to interact.

Although this signal has been shown to be significant, the underlying process responsible for this similarity is still controversial (Chen and Dokholyan 2006). There are two main hypotheses for explaining this phenomenon. The first hypothesis suggests that this evolutionary similarity comes from the mutual adaptation (co-evolution) of interacting proteins and the need to retain interaction features while sequences diverge. The second hypothesis implicates external factors. In this scenario, the restrictions imposed by evolution on the functional process implicating both proteins would be responsible for the parallelism of their phylogenetic trees.

Although the relative importance of both factors is still not clear, the predictive power of similarities in phylogenetic trees is not affected. Indeed, a number of developments have improved the original formulation (Pazos et al. 2005; Sato et al. 2005). The first advance involved managing the intrinsic similarity of the trees because of the common underlying taxonomic distribution (due to the speciation processes). This effect is analogous to the taxonomic biases discussed above. In these cases, the approach followed was to correct both trees by removing this common trend. For example, Pazos et al. subtracted the distances of the 16S rRNA phylogenetic tree to the corresponding distances for each protein tree. The correlations for the resulting distance matrices were used to predict protein interactions.

Additionally some analyses have focused on the selection of the sequence regions used for the tree building (Jothi et al. 2006; Kann et al. 2007). For example, it has been shown that interacting regions, both defined as interacting residues (using structural data) and as the sequence domain involved in the interaction, show more clear tree similarities than the whole proteins (Mintseris and Weng 2005; Jothi et al. 2006). Other interesting work showed that prediction performance can be improved by removing poorly conserved sequence regions (Kann et al. 2007).

Finally, in a very recent work (Juan et al. 2008) the authors have suggested a new method for removing noise in the detection of tree similarity signals and detecting different levels of evolutionary parallelism specificity. This method introduces the new strategy of using the global network of protein evolutionary similarity for a better calibration of the evolutionary parallelism between two proteins. For this purpose, they define a protein ‘co-evolutionary’ profile as the vector containing the evolutionary correlations between a given protein tree and all the rest of the protein trees derived from sequences in the same organism. This co-evolutionary profile is a more robust and comparable representation of the evolution of a given protein (it involves hundreds of distances) and can be used to deploy a new level of evolutionary comparison. The authors compare these co-evolutionary profiles by calculating Pearson’s correlation coefficient for each pair. In this way, the method detects pairs of proteins for which high evolutionary similarities are supported by their similarities with the rest of proteins of the organism. This approach significantly improves the predictive performance of the tree similarity-based methods so that different degrees of co-evolutionary specificity are obtained according to the number of proteins that might be influencing the co-evolution of the studied pair. This is done by extending the approach of Sato et al. (2006), that uses partial correlations and a reduced set of proteins for determining specific evolutionary similarities. Juan et al. calculated the partial correlation for each significant evolutionary similarity with respect to the remaining proteins in the organism and defined levels of co-evolutionary specificity according to the number of proteins that are considered to be co-evolving with each studied protein pair. With this strategy, it’s possible to detect a range of evolutionary parallelisms from the protein pairs (for very specific similarities) up to subsets of proteins (for more relaxed specificities) that are highly evolution dependent. Interestingly, if specificity requirements are relaxed, protein relationships among components of macro-molecular complexes and proteins involved in the same metabolic process can be recovered. This can be considered as a first step in the application of higher orders of evolutionary parallelisms to decode the evolutionary impositions over the protein interaction network.

12.3 Gene neighbourhood conservation

This method exploits the well-known tendency of bacterial organisms to organize proteins involved in the same biochemical process by clustering them in the genome. This observation is obviously related to the operon concept and the mechanisms for the coordination of transcription regulation of the genes present in these modules. These mechanisms are widespread among bacterial genomes. Therefore the significance of a given gene proximity can be established by its conservation in evolutionary distant species (Dandekar et al. 1998; Overbeek et al. 1999).

The availability of fully sequenced organisms makes computing the intergenic distances between each pair of genes easy. Genes with the same direction of transcrip-

tion and closer than 300 bases are typically considered to be in the same genomic context (see Fig. 12c). The conservation of this closeness must be found in more than two highly divergent organisms to be considered significant because of the taxonomic biases.

While this signal is strong in bacterial genomes, its relevance is unclear in eukaryotic genomes. This is the main drawback of these methodologies. In fact, this signal only can be exploited for eukaryotic organisms by extrapolating genomic closeness of bacterial genes to their homologues in eukaryotes. Obviously, this extrapolation leads to a considerable reduction in the confidence and number of obtained predictions for this evolutionary lineage. However, conserved gene pairs that are transcribed from a shared bidirectional promoter can be detected by similar methods and can be found in eukaryotes as well as prokaryotes (Korbel et al. 2004)

12.4 Gene fusion

A further use of evolutionary signals in protein function and physical interaction prediction has been the tendency of interacting proteins to be involved in gene fusion events. Sequences that appear as independently expressed ORFs in one organism become ‘fused’ as part of the same polypeptide sequence in another organism. These fusions are strong indicators of functional and structural interaction that have been suggested to increase the effective concentration of interacting functional domains (Enright et al. 1999; Marcotte et al. 1999b). This hypothesis proposes that gene fusion could remove the effect of diffusion and relative correct orientation of the proteins forming the original complex.

These fusion events are typically detected when sequence searches for two non-homologous proteins obtain a significant hit in the same sequence. Cases matching to the same region of the hit sequence are removed (these cases are schematically represented in Fig. 12d).

In spite of the strength of this signal, gene fusion seems to not be a habitual event in bacterial organisms. The difficulty of distinguishing protein interactions belonging to large evolutionary families is the main drawback of the automatic application of these methodologies.

13 Integration of experimentally determined and predicted interactions

As described above, there are many both experimental techniques and computational methods for determining and predicting interactions. To obtain the most comprehensive interaction networks possible, as many as possible of these sources of interactions should be integrated. The integration of these resources is complicated by the fact that

the different sources are not all equally reliable, and it is thus important to quantify the accuracy of the different evidence supporting an interaction.

In addition to the quality issues, comparison of different interaction sets is further complicated by the different nature of the datasets: yeast two-hybrid experiments are inherently binary, whereas pull-down experiments tend to report larger complexes. To allow for comparisons, complexes are typically represented by binary interaction networks; however, it is important to realize that there is not a single, clear definition of a “binary interaction”. For complex pull-down experiments, two different representations have been proposed: the matrix representation, in which each complex is represented by the set of binary interactions corresponding to all pairs of proteins from the complex, and the spoke representation, in which only bait-prey interactions are included (von Mering et al. 2002). The binary interactions obtained using either of these representations are somewhat artificial as some interacting proteins might in reality never touch each other and others might have too low an affinity to interact except in the context of the entire complex bringing them together. Even in the case of yeast two-hybrid assays, which inherently report binary interactions, not all interactions correspond to direct physical interactions.

The database STRING (“Search Tool for the Retrieval of Interacting Genes/Proteins”) (von Mering et al. 2007) represents an effort to provide many of the different types of evidence for functional interactions under one common framework with an integrated scoring scheme. Such an integrated approach offers several unique advantages: 1) various types of evidence are mapped onto a single, stable set of proteins, thereby facilitating comparative analysis; 2) known and predicted interactions often partially complement each other, leading to increased coverage; and 3) an integrated scoring scheme can provide higher confidence when independent evidence types agree.

In addition to the many associations imported from the protein interaction databases mentioned above (Bader et al. 2003; Salwinski et al. 2004; Guldener et al. 2006; Mishra et al. 2006; Stark et al. 2006; Chatr-aryamontri et al. 2007), STRING also includes interactions from curated pathway databases (Vastrik et al. 2007; Kanehisa et al. 2008) and a large body of predicted associations that are produced *de novo* using many of the methods described in this chapter (Dandekar et al. 1998; Gaasterland and Ragan 1998; Pellegrini et al. 1999; Marcotte et al. 1999c). These different types of evidence are obviously not directly comparable, and even for the individual types of evidence the reliability may vary. To address these two issues, STRING uses a two-stage approach. First, a separate scoring scheme is used for each evidence type to rank the interactions according to their reliability; these raw quality scores cannot be compared between different evidence types. Second, the ranked interaction lists are benchmarked against a common reference to obtain probabilistic scores, which can subsequently be combined across evidence types.

To exemplify how raw quality scores work, we will here explain the scoring scheme used for physical protein interactions from high-throughput screens. The two funda-

mentally different types of experimental interaction data sets, complex pull-downs and binary interactions are evaluated using separate scoring schemes. For the binary interaction experiments, e.g. yeast two-hybrid, the reliability of an interaction correlates well with the number of non-shared interaction partners for each interactor. STRING summarizes this in the following raw quality score:

$$S_1 = \log((N_1+1) \cdot (N_2+1)),$$

where N_1 and N_2 are the numbers of non-shared interaction partners. This score is similar to the IG1 measure suggested by Saito et al. (2002). In the case of complex pull-down experiments, the reliability of the inferred binary interactions correlates better with the number of times the interactors were co-purified compared to what would be expected at random:

$$S_2 = \log((N_{12} \cdot N)/((N_1+1) \cdot (N_2+1))),$$

where N_{12} is the number of purifications containing both proteins, N_1 and N_2 are the numbers of purifications containing either protein 1 or 2, and N is the total number of purifications. For this purpose, the bait protein was counted twice to account for bait-prey interactions being more reliable than prey-prey interactions. These raw quality scores are calculated for each individual high-throughput screen. Scores vary within one dataset, because they include additional, intrinsic information from the data itself, such as the frequency with which an interaction is detected. For medium sized data sets that are not large enough to apply the topology based scoring schemes, the same raw score is assigned to all interactions within a dataset. Finally, very small data sets are pooled and considered jointly as a single interaction set.

We similarly have different scoring schemes for predicted interactions based on co-expression in microarray expression studies, conserved gene neighborhood, gene fusion events and phylogenetic profiles. Based on these raw quality scores, a confidence score is assigned to each predicted association by benchmarking the performance of the predictions against a common reference set of trusted, true associations. STRING uses as reference the functional grouping of proteins maintained at KEGG (Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al. 2008)). Any predicted association for which both proteins are assigned to the same “KEGG pathway” is counted as a true positive. KEGG pathways are particularly suitable as a reference because they are based on manual curation, are available for a number of organisms, and cover several functional areas. Other benchmark sets could also be used, for example “Biological Process” terms from Gene Ontology (Ashburner et al. 2000) or Reactome pathways (Vastrik et al. 2007). The benchmarked confidence scores in STRING generally correspond to the probability of finding the linked proteins within the same pathway or biological process.

The assignment of probabilistic scores for all evidence types solves many of the issues of data integration. First, incomparable evidence types are made comparable by

assigning a score that represents how well the evidence type can predict a certain type of interactions (the type being specified by the reference set used). Second, the separate benchmarking of interactions from, for example, different high-throughput protein interaction screens accounts for any differences in reliability between different studies. Third, use of raw quality scores allows us to separate more reliable interactions from less reliable interactions even within a single dataset. The probabilistic nature of the scores also makes it easy to calculate the combined reliability of an interaction given multiple lines of evidence. It is computed under the assumption of independence for the various sources, in a naïve Bayesian fashion.

In addition to having a good scoring scheme, it is crucial to make the evidence for an interaction transparent to the end users. To achieve this, the STRING interaction network is made available via a user-friendly web interface (<http://string.embl.de>). When performing a query, the user will first be presented with a network view, which provides a first, simplified overview (Fig. 13). From here the user has full control over parameters such as the number of proteins shown in the network (nodes) and the minimal reliability required for an interaction (edge) to be displayed. From the network, the user also has the ability to drill down on the evidence that underlies any given interaction using the dedicated viewer for each evidence type. For example, it is possible to inspect the publications that support a given interaction, the set of protein that were

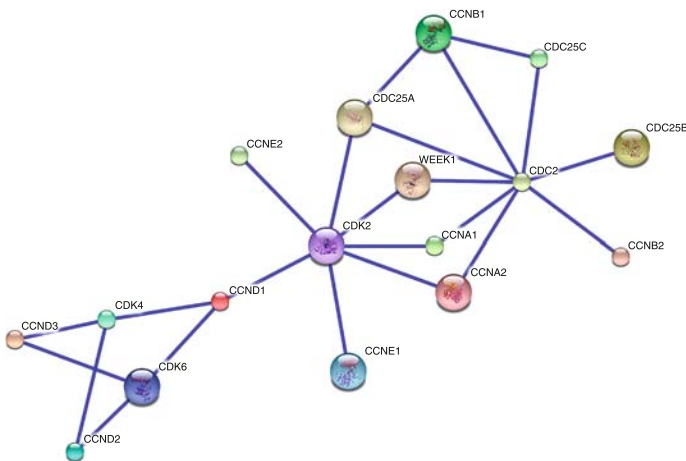


Fig. 13 Protein interaction network of the core cell-cycle regulation in human. The network was constructed by querying the STRING database (von Mering et al. 2007) for very high confidence interactions (conf. score > 0.99) between four cyclin-dependent kinases, their associated cyclins, the WEE1 kinase and the CDC25 phosphatases. The network correctly recapitulates CDC2 interacts with cyclin-A/B, CDK2 with cyclin-A/E, and CDK4/6 with cyclin-D. It also shows that the WEE1 and CDC25 phosphatases regulate CDC2 and CDK2 but not CDK4 and CDK6. Moreover, the network suggests that CDC25A phosphatase regulates CDC2 and CDK2, whereas CDC25B and CDC25C specifically regulate CDC2

co-purified in a particular experiment and the phylogenetic profiles or genomic context based on which an interaction was predicted.

14 Domain–domain interactions

Protein binding is commonly characterized by specific interactions of evolutionarily conserved domains (Pawson and Nash 2003). Domains are fundamental units of protein structure and function (Aloy and Russell 2006), which are incorporated into different proteins by genetic duplications and rearrangements (Vogel et al. 2004). Globular domains are defined as structural units of fifty and more amino acids that usually fold independently of the remaining polypeptide chain to form stable, compact structures (Orengo and Thornton 2005). They often carry important functional sites and determine the specificity of protein interactions (Fig. 14). Essential information on

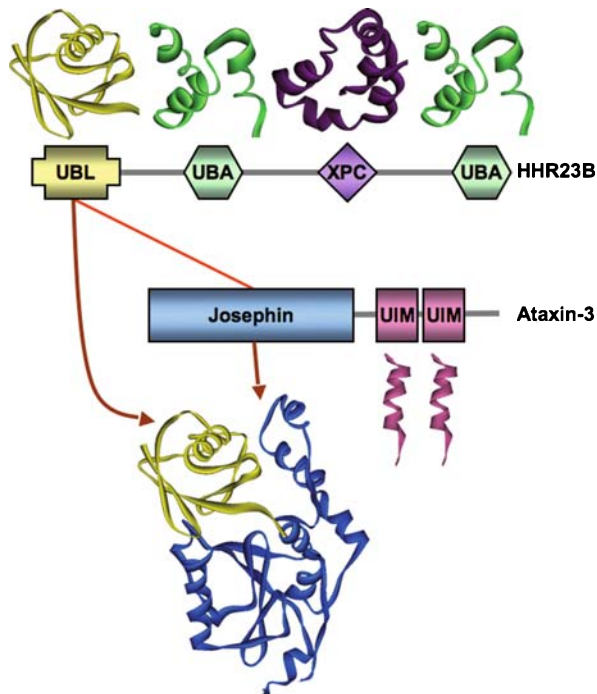


Fig. 14 Exemplary interaction between the two human proteins HHR23B and ataxin-3. Each protein domain commonly adopts a particular 3D structure and may fulfill a specific molecular function. Generally, the domains responsible for an observed protein-protein interaction need to be determined before further functional characterizations are possible. In the depicted protein-protein interaction, it is known from experiments that the ubiquitin-like domain UBL of HHR23B (yellow) forms a complex with de-ubiquitinating Josephin domain of ataxin-3 (blue) (Nicastro et al. 2005)

the cellular function of specific protein interactions and complexes can often be gained from the known functions of the interacting protein domains. Domains may contain binding sites for proteins and ligands such as metabolites, DNA/RNA, and drug-like molecules (Xia et al. 2004). Widely spread domains that mediate molecular interactions can be found alone or combined in conjunction with other domains and intrinsically disordered, mainly unstructured, protein regions connecting globular domains (Dunker et al. 2005). According to Apic et al. (2001) multi-domain proteins constitute two thirds of unicellular and 80% of metazoan proteomes. One and the same domain can occur in different proteins, and many domains of different types are frequently found in the same amino acid chain.

Much effort is being invested in discovering, annotating, and classifying protein domains both from the functional (Pfam (Finn et al. 2006), SMART (Letunic et al. 2006), CDD (Marchler-Bauer et al. 2007), InterPro (Mulder et al. 2007) and structural (SCOP (Andreeva et al. 2004), CATH (Greene et al. 2007)) perspective. Notably, it may be confusing that the term ‘domain’ is commonly used in two slightly different meanings. In the context of domain databases such as Pfam and SMART, a domain is basically defined by a set of homologous sequence regions, which constitute a domain family. In contrast, a specific protein may contain one or more domains, which are concrete sequence regions within its amino acid sequence corresponding to autonomously folding units. Domain families are commonly represented by Hidden Markov Models (HMMs), and highly sensitive search tools like HMMER (Eddy 1998) are used to identify domains in protein sequences.

Different sources of information about interacting domains with experimental evidence are available. Experimentally determined interactions of single-domain proteins indicate domain–domain interactions. Similarly, experiments using protein fragments help identifying interaction domains, but this knowledge is frequently hidden in the text of publications and not contained in any database. However, domain databases like Pfam, SMART, and InterPro may contain some annotation obtained by manual literature curation. In the near future, high-throughput screening techniques will result in even larger amounts of protein fragment interaction data to delineate domain borders and interacting protein regions (Colland and Daviet 2004).

Above all, three-dimensional structures of protein domain complexes are experimentally solved by X-ray crystallography or NMR and are deposited in the PDB database (Berman et al. 2007). Structural contacts between two interacting proteins can be derived by mapping sequence positions of domains onto PDB structures. Extensive investigations of domain combinations in proteins of known structures (Apic et al. 2001) as well as of structurally resolved homo- or heterotypic domain interactions (Park et al. 2001) revealed that the overlap between intra- and intermolecular domain interactions is rather limited. Two databases, iPfam (Finn et al. 2005) and 3did (Stein et al. 2005), provide pre-computed structural information about protein interactions at the level of Pfam domains.

Analysis of structural complexes suggests that interactions between a given pair of proteins may be mediated by different domain pairs in different situations and in different organisms. Nevertheless, many domain interactions, especially those involved in basic cellular processes such as DNA metabolism and nucleotide binding, tend to be evolutionarily conserved within a wide range of species from prokaryotes to eukaryotes (Itzhaki et al. 2006). In yeast, Pfam domain pairs are associated with over 60% of experimentally known protein interactions, but only 4.5% of them are covered by iPfam (Schuster-Bockler and Bateman 2007).

Domain interactions can be inferred from experimental data on protein interactions by identifying those domain pairs that are significantly overrepresented in interacting proteins compared to random protein pairs (Deng et al. 2002; Ng et al. 2003a; Riley et al. 2005; Sprinzak and Margalit 2001) (Fig. 15). However, the predictive power of such an approach is strongly dependent on the quality of the data used as the source of information for protein interactions, and the coverage of protein sequences in terms of domain assignments. Basically, the likelihood of two domains, D_i and D_j , to interact

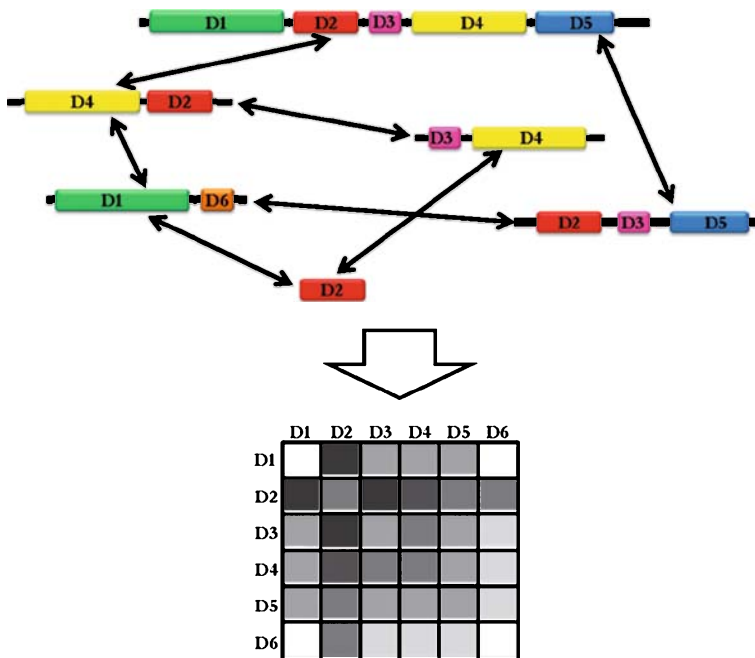


Fig. 15 Deriving the likelihood of domain interactions from experimental data of protein interactions. Six different proteins are shown containing domains D1., D2., . . . , D6 in different combinations. Known interactions between these proteins are shown as black arrows. The matrix in the bottom part of the figure shows the likelihood for each pair of domains to interact – from low (white) to high (dark)

can be estimated as the fraction of protein pairs known to interact among all proteins in the dataset containing this domain pair.

This basic idea has been improved upon by using a maximum-likelihood (ML) approach based on the expectation-maximization (EM) algorithm. This method finds the maximum likelihood estimator of the observed protein–protein interactions by an iterative cycle of computing the expected likelihood (E-step) and maximizing the unobserved parameters (domain interaction propensities) in the M-step. When the algorithm converges (i.e. the total likelihood cannot be further improved by the algorithm), the ML estimate for the likelihood of the unobserved domain interactions is found (Deng et al. 2002; Riley et al. 2005). Riley and colleagues further improved this method by excluding each potentially interacting domain pair from the dataset and re-computing the ML-estimate to obtain an additional confidence value for the respective domain–domain interaction. This domain pair exclusion (DPEA) method measures the contribution of each domain pair to the overall likelihood of the protein interaction network based on domain–domain interactions. In particular, this approach enables the prediction of specific domain–domain interactions between selected proteins which would have been missed by the basic ML method. Another ML-based algorithm is InSite which takes differences in the reliability of the protein–protein interaction data into account (Wang et al. 2007a). It also integrates external evidence such as functional annotation or domain fusion events.

An alternative method for deriving domain interactions is through co-evolutionary analysis that exploits the notion that mutations of residue pairs at the interaction interfaces are correlated to preserve favorable physico-chemical properties of the binding surface (Jothi et al. 2006). The pair of domains mediating interactions between two proteins P1 and P2 may therefore be expected to display a higher similarity of their phylogenetic trees than other, non-interacting domains (Fig. 16). The degree of agreement between the evolutionary history of two domains, D_i and D_j , can be computed by the Pearson's correlation coefficient r_{ij} between the similarity matrices of the domain sequences in different organisms:

$$r_{ij} = \frac{\sum_{p=1}^{n-1} \sum_{q=p+1}^n (M_{pq}^i - \bar{M}^i)(M_{pq}^j - \bar{M}^j)}{\sqrt{\sum_{q=p+1}^{n-1} \sum_{q=p+1}^n ((M_{pq}^i - \bar{M}^i))^2 \sum_{p=1}^{n-1} \sum_{q=p+1}^n ((M_{pq}^j - \bar{M}^j))^2}},$$

where n is the number of species, M_{pq}^i and M_{pq}^j are the evolutionary distances between species, and \bar{M}^i and \bar{M}^j are the mean values of the matrices, respectively. In Figure 16 the evolutionary tree of the domain D2 is most similar to those of D5 and D6, corroborating the actual binding region.

A well-known limitation of the correlated mutation analysis is that it is very difficult to decide whether residue co-variation happens as a result of functional co-evolution

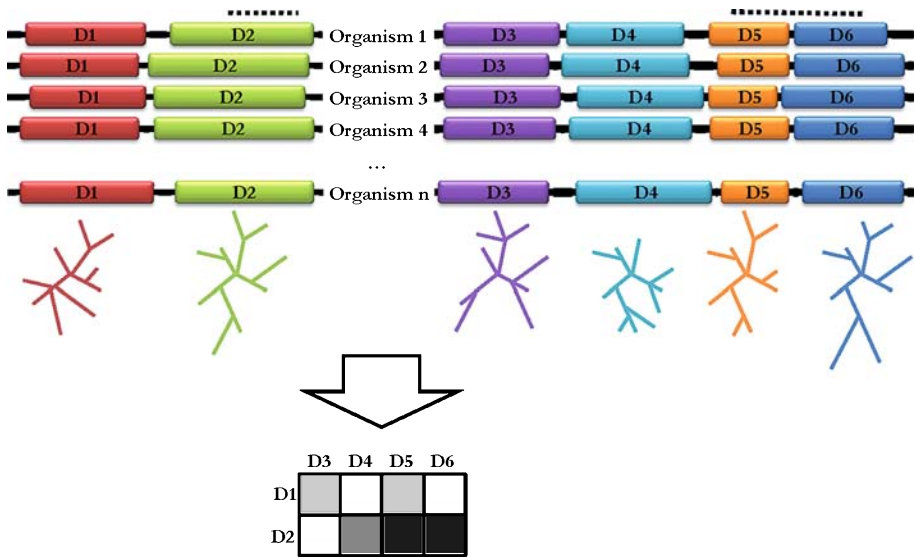


Fig. 16 Co-evolutionary analysis of domain interactions. Two orthologous proteins from different organisms known to interact with each other are shown. The first protein consists of two domains, D1 and D2, while the second protein includes the domains D3, D4, D5, and D6. Evolutionary trees for each domain are shown, their similarity serves as an indication of interaction likelihood that is encoded in the interaction matrix

directed at preserving interaction sites, or because of sequence divergence due to speciation. To address this problem, (Kann et al. 2007) suggested to distinguish the relative contribution of conserved and more variable regions in aligned sequences to the co-evolution signal based on the hypothesis that functional co-evolution is more prominent in conserved regions.

Finally, interacting domains can be identified by phylogenetic profiling, as described above for full-chain proteins. As in the case of complete protein chains, the similarity of evolutionary patterns shared by two domains may indicate that they interact with each other directly or at least share a common functional role (Pagel et al. 2004). As illustrated in Fig. 17, clustering protein domains with similar phylogenetic profiles allows researchers to build domain interaction networks which provide clues for describing molecular complexes. Similarly, the DomainTeam method (Pasek et al. 2005) considers chromosomal neighborhoods at the level of conserved domain groups.

A number of resources provide and combine experimentally derived and predicted domain interaction data. InterDom (<http://interdom.i2r.a-star.edu.sg/>) integrates domain-interaction predictions based on known protein interactions and complexes with domain fusion events (Ng et al. 2003b). DIMA (<http://mips.gsf.de/genre/proj/dima2>) is another database of domain interactions, which integrates experimentally demon-

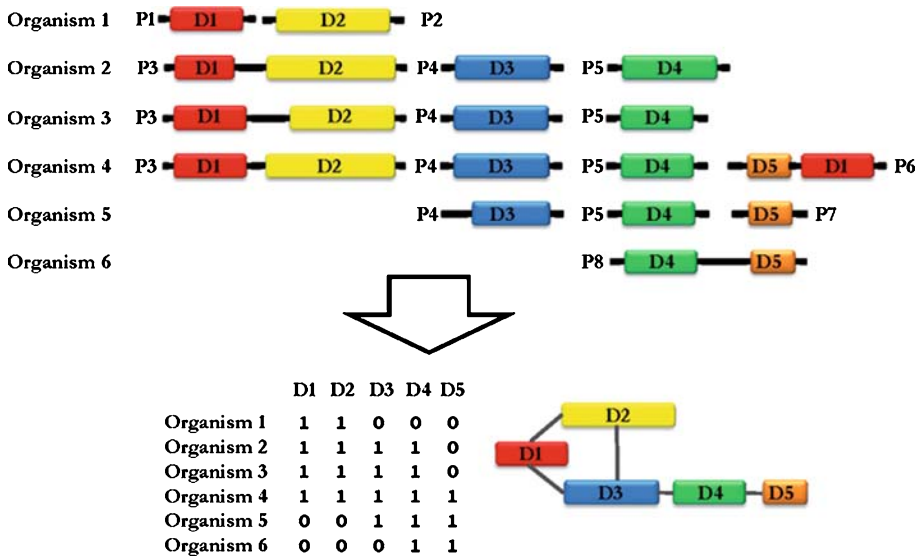


Fig. 17 Similarity of domain phylogenetic profiles can be used to build a domain interaction network

strated domain interactions from iPfam and 3did with predictions based on the DPEA algorithm and phylogenetic domain profiling (Pagel et al. 2007). Recently, two new comprehensive resources, DOMINE (<http://domine.utdallas.edu>) (Raghavachari et al. 2008) and DASMI (<http://www.dasmi.de>) (Blankenburg et al. 2008, submitted), were introduced and are available online. These resources contain iPfam and 3did data and predicted domain interactions taken from several other publications. Predictions are based on several methods for deriving domain interactions from protein interaction data, phylogenetic domain profiling data and domain coevolution. With the availability of an increasing number of predictions the task of method weighting and quality assessment becomes crucial. A thorough analysis of the quality of domain interaction data can be found in Schlicker et al. (2007).

Beyond domain–domain contacts, an alternative mechanism of mediating molecular recognition is through binding of protein domains to short sequence regions (Santonico et al. 2005), typically from three to eight residues in length (Zarrinpar et al. 2003; Neduva et al. 2005). Such linear recognition motifs can be discovered from protein interaction data by identifying amino acid sequence patterns overrepresented in proteins that do not possess significant sequence similarity, but share the same interacting partner (Yaffe 2006). Web services like EML (<http://elm.eu.org> (Puntervoll et al. 2003)), support the identification of linear motifs in protein sequences.

As described above, specific adapter domains can mediate protein–protein interactions. While some of these interaction domains recognize small target peptides, others

are involved in domain–domain interactions. As short binding motifs have a rather high probability of being found by chance and the exact mechanisms of binding specificity for this mode of interaction are not understood completely, predictions of protein–protein interactions based on binding domains is currently limited to domain–domain interactions for which reliable data is available.

Predicting PPIs from domain interactions may simply be achieved by reversing the ideas discussed above, that is, by using the domain composition of proteins to evaluate the interaction likelihood of proteins (Bock and Gough 2001; Sprinzak and Margalit 2001; Wojcik and Schachter 2001). In a naive approach, domain interactions are treated as independent, and all protein pairs with a matching pair of interacting domains are predicted to engage in an interaction. Given that protein interactions may also be mediated by several domain interactions simultaneously, more advanced statistical methods take into account dependencies between domains and exploit domain combinations (Han et al. 2004) and multiple interacting domain pairs (Chen and Liu 2005).

Exercising and validating these prediction approaches revealed that the most influential factor for PPI prediction is the quality of the underlying data. This suggests that, as for most biological predictions in other fields, the future of prediction methods for protein and domain interactions may lie in the integration of different sources of evidence and weighting the individual contributions based on calibration to gold-standard data. Further methodological improvements may include the explicit consideration of cooperative domains, that is, domain pairs that jointly interact with other domains (Wang et al. 2007b).

15 Biomolecular docking

Basic interactions between two or up to a few biomolecules are the basic elements of the complex molecular interaction networks that enable the processes of life and, when thrown out of their intended equilibrium, manifest the molecular basis of diseases. Such interactions are at the basis of the formation of metabolic, regulatory or signal transduction pathways. Furthermore the search for drugs boils down to analyzing the interactions between the drug molecule and the molecular target to which it binds, which is often a protein.

For the analysis of a single molecular interaction, we do not need complex biological screening data. Thus it is not surprising that the analysis of the interactions between two molecules, one of them being a protein, has the longest tradition in computational biology of all problems involving molecular interactions, dating back over three decades. The basis for such analysis is the knowledge of the three-dimensional structure of the involved molecules. To date, such knowledge is based almost exclusively on experimental measurements, such as X-ray diffraction data or NMR spectra. There are

also a few reported cases in which the analysis of molecular interactions based on structural models of protein has led to successes.

The analysis of the interaction of two molecules based on their three-dimensional structure is called molecular docking. The input is composed of the three-dimensional structures of the participating molecules. (If the involved molecule is very flexible one admissible structure is provided.) The output consists of the three-dimensional structure of the molecular complex formed by the two molecules binding to each other. Furthermore, usually an estimate of the differential free energy of binding is given, that is, the energy difference ΔG between the bound and the unbound conformation. For the binding event to be favorable that difference has to be negative.

15.1 Protein-ligand docking

This slight misnomer describes the binding between a protein molecule and a small molecule. The small molecule can be a natural substrate such as a metabolite or a molecule to be designed to bind tightly to the protein such as a drug molecule. Protein-ligand docking is the most relevant version of the docking problem because it is a useful help in searching for new drugs. Also, the problem lends itself especially well to computational analysis, because in pharmaceutical applications one is looking for small molecules that are binding very tightly to the target protein, and that do so in a conformation that is also a low-energy conformation in the unbound state. Thus, subtle energy differences between competing ligands or binding modes are not of prime interest. For these reasons there is a developed commercial market for protein-ligand docking software.

Usually the small molecule has a molecular weight of up to several hundred Daltons and can be quite flexible. Typically, the small molecule is given by its 2D structure formula, e.g., in the form of a SMILES string (Weininger 1988). If a starting 3D conformation is needed there is special software for generating such a conformation (see, e.g. (Pearlman 1987; Sadowski et al. 1994)).

Challenges of the protein ligand problem are (i) finding the correct conformation of the usually highly flexible ligand in the binding site of the protein, (ii) determining the subtle conformational changes in the binding site of the protein upon binding of the ligand, which are termed induced fit, (iii) producing an accurate estimate of the differential energy of binding or at least ranking different conformations of the same ligand and conformations of different ligands correctly by their differential energy of binding. Methods tackling problem (ii) can also be used to rectify smaller errors in structural models of proteins whose structure has not been resolved experimentally. The solution of problem (iii) provides the essential selection criterion for preferred ligands and binding modes, namely those with lowest differential energy of binding.

Challenge (i) has basically been conquered in the last decade as a number of docking programs have been developed that can efficiently sample the conformational space of

the ligand and produce correct binding modes of the ligand within the protein, assuming that the protein is given in the correct structure for binding the ligand. Several methods are applied here. The most brute-force method is to just try different (rigid) conformations of the ligand one after the other. If the program is fast enough one can run through a sizeable number of conformations per ligand (McGann et al. 2003). A more algorithmic and quite successful method is to build up the ligand from its molecular fragments inside the binding pocket of the protein (Rarey et al. 1996). Yet another class of methods sample ligand conformations inside the protein binding pocket by methods such as local search heuristics, Monte Carlo sampling or genetic algorithms (Abagyan et al. 1994; Jones et al. 1997; Morris et al. 1998). There are also programs exercising combinations of different methods (Friesner et al. 2004). The reported methods usually can compute the binding mode of a ligand inside a protein within fractions of a minute to several minutes. The resulting programs can be applied to screening through large databases of ligands involving hundreds of thousands to millions of compounds and are routinely used in pharmaceutical industry in the early stages of drug design and selection. They are also repeatedly compared on benchmark datasets (Kellenberger et al. 2004; Chen et al. 2006; Englebienne et al. 2007). More complex methods from computational biophysics, such as molecular dynamics (MD) simulations that compute a trajectory of the molecular movement based on the forces exerted on the molecules take hours on a single problem instance and can only be used for final refinement of the complex.

Challenges (ii) and (iii) have not been solved yet. Concerning problem (ii), structural changes in the protein can involve redirections of side chains in or close to the binding pocket and more substantial changes involving backbone movement. While recently methods have been developed to optimize side-chain placement upon ligand binding (Claußen et al. 2001; Sherman et al. 2006), the problem of finding the correct structural change upon binding involving backbone and side-chain movement is open (Carlson 2002). Concerning problem (iii), there are no scoring functions to date that are able to sufficiently accurately estimate the differential energy of binding on a diverse set of protein-ligand complexes (Wang et al. 2003; Huang and Zou 2006). This is especially unfortunate as an inaccurate estimate of the binding energy causes the docking program to disregard correct complex structures even though they have been sampled by the docking program because they are labeled with incorrect energies. This is the major problem in docking which limits the accuracy of the predictions. Recent reviews on protein-ligand docking have been published in Sousa et al. (2006) and Rarey et al. (2007).

One restriction with protein-ligand docking as it applies to drug design and selection is that the three-dimensional structure of the target protein needs to be known. Many pharmaceutical targets are membrane-standing proteins for which we do not have the three-dimensional structure. For such proteins there is a version of drug screening that can be viewed as the negative imprint of docking: Instead of docking the

drug candidate into the binding site of the protein – which is not available – we superpose the drug candidate (which is here called the test molecule) onto another small molecule which is known to bind to the binding site of the protein. Such a molecule can be the natural substrate for the target protein or another drug targeting that protein. Let us call this small molecule the reference molecule. The suitability of the new drug candidate is then assessed on the basis of its structural and chemical similarity with the reference molecule. One problem is that now both the test molecule and the reference molecule can be highly flexible. But in many cases largely rigid reference molecules can be found, and in other cases it suffices to superpose the test molecule onto any low-energy conformation of the reference molecule. There are several classes of drug screening programs based on this molecular comparison, ranging from (i) programs that perform a detailed analysis of the three-dimensional structures of the molecules to be compared (e.g. (Lemmen et al. 1998; Krämer et al. 2003)) across (ii) programs that perform a topological analysis of the two molecules (Rarey and Dixon 1998; Gillet et al. 2003) to (iii) programs that represent both molecules by binary or numerical property vectors which are compared with string methods (McGregor and Muskal 1999; Xue et al. 2000). The first class of programs require fractions of seconds to fractions of a minute for a single comparison, the second can perform hundreds comparisons per second, the third up to several ten thousand comparisons per second. Reviews of methods for drug screening based on ligand comparison are given in (Lengauer et al. 2004; Kämper et al. 2007).

15.2 Protein–protein docking

Here both binding partners are proteins. Since drugs tend to be small molecules this version of the docking problem is not of prime interest in drug design. Also, the energy balance of protein–protein binding is much more involved than for protein–ligand binding. Optimal binding modes tend not to form troughs in the energy landscape that are as pronounced as for protein–ligand docking. The binding mode is determined by subtle side-chain rearrangements of both binding partners that implement the induced fit along typically quite large binding interfaces. The energy balance is dominated by difficult to analyze entropic terms involving the desolvation of water within the binding interface. For these reasons, the software landscape for protein–protein docking is not as well developed as for protein–ligand docking and there is no commercial market for protein–protein docking software.

Protein–protein docking approaches are based either on conformational sampling and MD – which can naturally incorporate molecular flexibility but suffers from very high computing demands – or on combinatorial sampling with both proteins considered rigid in which case handling of protein flexibility has to be incorporated with methodical extensions. For space reasons we do not detail methods for protein–protein docking. A recent review on the subject can be found in Hildebrandt et al. (2007).

A variant of protein–protein docking is protein–DNA docking. This problem shares with protein–protein docking the character that both binding partners are macromolecules. However, entropic aspects of the energy balance are even more dominant in protein–DNA docking than in protein–protein docking. Furthermore DNA can assume nonstandard shapes when binding to proteins which deviate much more from the known double helix than we are used to when considering induced fit phenomena.

References

- Abagyan R, Totrov M, Kuznetsov D (1994) ICM-a method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J Comput Chem* 15: 488–506
- Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol* 24: 537–544
- Albert R (2005) Scale-free networks in cell biology. *J Cell Sci* 118: 4947–4957
- Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. *J Mol Biol* 332: 989–998
- Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. *Nat Biotechnol* 22: 1317–1321
- Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol* 7: 188–197
- Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32: D226–D229
- Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 310: 311–325
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29
- Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M (2008) Computing topological parameters of biological networks. *Bioinformatics* 24: 282–284
- Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular interaction network database. *Nucleic Acids Res* 31: 248–250
- Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5: 101–113
- Barker D, Meade A, Pagel M (2007) Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* 23: 14–20
- Barker D, Pagel M (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol* 1: e3
- 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
- Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 35: D301–D303

- Bock JR, Gough DA (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics* 17: 455–460
- Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32: D267–D270
- Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S, Hopf C, Huhse B, Mangano R, Michon AM, Schirle M, Schlegl J, Schwab M, Stein MA, Bauer A, Casari G, Drewes G, Gavin AC, Jackson DB, Joberty G, Neubauer G, Rick J, Kuster B, Superti-Furga G (2004) A physical and functional map of the human TNF- α /NF- κ B signal transduction pathway. *Nat Cell Biol* 6: 97–105
- Bowers PM, Cokus SJ, Eisenberg D, Yeates TO (2004) Use of logic relationships to decipher protein network organization. *Science* 306: 2246–2249
- Breitkreutz BJ, Stark C, Tyers M (2003) The GRID: the General Repository for Interaction Datasets. *Genome Biol* 4: R23
- 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
- Calderwood MA, Venkatesan K, Xing L, Chase MR, Vazquez A, Holthaus AM, Ewence AE, Li N, Hirozane-Kishikawa T, Hill DE, Vidal M, Kieff E, Johannsen E (2007) Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci USA* 104: 7606–7611
- Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, Miller-Graziano C, Moldawer LL, Mindrinos MN, Davis RW, Tompkins RG, Lowry SF (2005) A network-based analysis of systemic inflammation in humans. *Nature* 437: 1032–1037
- Camargo LM, Collura V, Rain JC, Mizuguchi K, Hermjakob H, Kerrien S, Bonnert TP, Whiting PJ, Brandon NJ (2006) Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia. *Mol Psychiatry* 12: 74–86
- Carlson HA (2002) Protein flexibility is an important component of structure-based drug discovery. *Curr Pharm Des* 8: 1571–1578
- Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the Molecular INteraction database. *Nucleic Acids Res* 35: D572–D574
- Chen H, Lyne PD, Giordanetto F, Lovell T, Li J (2006) On evaluating molecular-docking methods for pose prediction and enrichment factors. *J Chem Inf Model* 46: 401–415
- Chen XW, Liu M (2005) Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics* 21: 4394–4400
- Chen Y, Dokholyan NV (2006) The coordinated evolution of yeast proteins is constrained by functional modularity. *Trends Genet* 22: 416–419
- Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3: 140
- Claußen H, Buning C, Rarey M, Lengauer T (2001) FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol* 308: 377–395
- Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2: 2366–2382
- Colland F, Daviet L (2004) Integrating a functional proteomic approach into the target discovery process. *Biochimie* 86: 625–632
- Colland F, Rain JC, Gounon P, Labigne A, Legrain P, De Reuse H (2001) Identification of the *Helicobacter pylori* anti- σ 28 factor. *Mol Microbiol* 41: 477–487
- Cui Q, Ma Y, Jaramillo M, Bari H, Awan A, Yang S, Zhang S, Liu L, Lu M, O'Connor-McCourt M, Purisima EO, Wang E (2007) A map of human cancer signaling. *Mol Syst Biol* 3: 152

- Cusick ME, Klitgord N, Vidal M, Hill DE (2005) Interactome: gateway into systems biology. *Hum Mol Genet* 14 Spec No. 2: R171–R181
- Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23: 324–328
- Date SV, Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* 21: 1055–1062
- Deng M, Mehta S, Sun F, Chen T (2002) Inferring domain–domain interactions from protein–protein interactions. *Genome Res* 12: 1540–1548
- Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *Febs J* 272: 5129–5148
- Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14: 755–763
- Englebienne P, Fiaux H, Kuntz DA, Corbeil CR, Gerber-Lemaire S, Rose DR, Moitessier N (2007) Evaluation of docking programs for predicting binding of Golgi alpha-mannosidase II inhibitors: a comparison with crystallography. *Proteins* 69: 160–176
- Enright AJ, Iliopoulos I, Kyripides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402: 86–90
- Fields S, Song O (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340: 245–246
- Finn RD, Marshall M, Bateman A (2005) iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 21: 410–412
- Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34: D247–D251
- Fishman MC, Porter JA (2005) Pharmaceuticals: a new grammar for drug discovery. *Nature* 437: 491–493
- Flajolet M, Rotondo G, Daviet L, Bergametti F, Inchauspe G, Tiollais P, Transy C, Legrain P (2000) A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* 242: 369–379
- Franke L, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 78: 1011–1025
- Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* 47: 1739–1749
- Fryxell KJ (1996) The coevolution of gene family trees. *Trends Genet* 12: 364–369
- Gaasterland T, Ragan MA (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics* 3: 199–217
- Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* 38: 285–293
- Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440: 631–636
- 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
- Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, Osterman AL (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol* 185: 5673–5684
- Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelm J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418: 387–391
- Giallourakis C, Henson C, Reich M, Xie X, Mootha VK (2005) Disease gene discovery through integrative genomics. *Annu Rev Genomics Hum Genet* 6: 381–406
- Gillet VJ, Willett P, Bradshaw J (2003) Similarity searching using reduced graphs. *J Chem Inf Comput Sci* 43: 338–345
- Giot L (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302: 1727–1736
- Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, Herbst M, Suopanki J, Scherzinger E, Abraham C, Bauer B, Hasenbank R, Fritzsche A, Ludewig AH, Bussow K, Coleman SH, Gutekunst CA, Landwehrmeyer BG, Lehrach H, Wanker EE (2004) A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell* 15: 853–865
- Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) Co-evolution of proteins with their interaction partners. *J Mol Biol* 299: 283–293
- Goll J, Uetz P (2007) Analyzing Protein Interaction Networks. In: Lengauer T (ed) *Bioinformatics – from genomes to therapies*. Wiley-VCH, Weinheim, pp 1121–1179
- Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. *Proc Natl Acad Sci USA* 104: 8685–8690
- Graham DL, Lowe PN, Grime GW, Marsh M, Rittinger K, Smerdon SJ, Gamblin SJ, Eccleston JF (2002) MgF(3)(-) as a transition state analog of phosphoryl transfer. *Chem Biol* 9: 375–381
- Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 35: D291–D297
- Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res* 34: D436–D441
- Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33: D514–D517
- Han DS, Kim HS, Jang WH, Lee SD, Suh JK (2004) PreSPI: a domain combination based prediction system for protein–protein interaction. *Nucleic Acids Res* 32: 6312–6320
- 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 (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32 (Database issue): D258–D261
- Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* 402: C47–C52
- He X, Zhang J (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet* 2: e88
- Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R (2004) The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. *Nat Biotechnol* 22: 177–183
- Hildebrandt A, Kohlbacher O, Lenhof H-P (2007) Modeling protein–protein and protein–DNA docking. In: Lengauer T (ed) *Bioinformatics – from genomes to therapies*. Wiley-VCH, Weinheim, pp 601–650
- Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415: 180–183
- Hu Z, Mellor J, Wu J, Kanehisa M, Stuart JM, DeLisi C (2007) Towards zoomable multidimensional maps of the cell. *Nat Biotechnol* 25: 547–554
- Hu Z, Mellor J, Wu J, Yamada T, Holloway D, Delisi C (2005) VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res* 33: W352–W357
- Huang SY, Zou X (2006) An iterative knowledge-based scoring function to predict protein–ligand interactions: II. Validation of the scoring function. *J Comput Chem* 27: 1876–1882
- Huynen MA, Bork P (1998) Measuring genome evolution. *Proc Natl Acad Sci USA* 95: 5849–5856
- Itzhaki Z, Akiva E, Altuvia Y, Margalit H (2006) Evolutionary conservation of domain–domain interactions. *Genome Biol* 7: R125
- Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* 302: 449–453
- Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411: 41–42
- Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 267: 727–748
- Jones RB, Gordus A, Krall JA, MacBeath G (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature* 439: 168–174
- Jonsson PF, Bates PA (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22: 2291–2297
- Jothi R, Cherukuri PF, Tasneem A, Przytycka TM (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. *J Mol Biol* 362: 861–875
- Juan D, Pazos F, Valencia A (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci USA* 105: 934–939

- Kaltenbach LS, Romero E, Becklin RR, Chettier R, Bell R, Phansalkar A, Strand A, Torcassi C, Savage J, Hurlburt A, Cha GH, Ukani L, Chepanoske CL, Zhen Y, Sahasrabudhe S, Olson J, Kurschner C, Ellerby LM, Peltier JM, Botas J, Hughes RE (2007) Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genet* 3: e82
- Kämper A, Rognan D, Lengauer T (2007) Lead Identification by virtual screening. In: Lengauer T (ed) *Bioinformatics – from genomes to therapies*. Wiley-VCH, Weinheim, pp 651–704
- Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36: D480–D484
- Kann MG (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform* 8: 333–346
- Kann MG, Jothi R, Cherukuri PF, Przytycka TM (2007) Predicting protein domain interactions from coevolution of conserved regions. *Proteins* 67: 811–820
- Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 25: 197–206
- Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* 57: 225–242
- Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thornecroft D, Zhang Y, Apweiler R, Hermjakob H (2007a) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res* 35: D561–D565
- Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-Aryamontri A, Oesterheld M, Stumpflen V, Salwinski L, Nerothin J, Cerami E, Cusick ME, Vidal M, Gilson M, Armstrong J, Woollard P, Hogue C, Eisenberg D, Cesareni G, Apweiler R, Hermjakob H (2007b) Broadening the Horizon – Level 2.5 of the HUPO-PSI Format for Molecular Interactions. *BMC Biol* 5: 44
- Kitano H (2007) A robustness-based approach to systems-oriented drug design. *Nat Rev Drug Discov* 6: 202–210
- Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* 3: e134
- Korbel JO, Jensen LJ, von Mering C, Bork P (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol* 22: 911–917
- Krämer A, Horn HW, Rice JE (2003) Fast 3D molecular superposition and similarity search in databases of flexible molecules. *J Comput Aided Mol Des* 17: 13–18
- Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O’Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440: 637–643
- Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res* 36: D684–D688
- LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes RE (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* 438: 103–107

- Lage K, Karlberg EO, Stirling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 25: 309–316
- Legrain P, Selig L (2000) Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett* 480: 32–36
- Lemmen C, Lengauer T, Klebe G (1998) FLEXS: a method for fast flexible ligand superposition. *J Med Chem* 41: 4502–4520
- Lengauer T, Lemmen C, Rarey M, Zimmermann M (2004) Novel technologies for virtual screening. *Drug Discov Today* 9: 27–34
- Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34: D257–D260
- Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, Barabasi AL, Vidal M, Zoghbi HY (2006) A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 125: 801–814
- Lin J, Gan CM, Zhang X, Jones S, Sjoblom T, Wood LD, Parsons DW, Papadopoulos N, Kinzler KW, Vogelstein B, Parmigiani G, Velculescu VE (2007) A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res* 17: 1304–1318
- Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, Kasif S (2007) Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet* 3: e96
- Loscalzo J, Kohane I, Barabasi AL (2007) Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol* 3: 124
- Lu X, Jain VV, Finn PW, Perkins DL (2007) Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Mol Syst Biol* 3: 98
- Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 35: D237–D240
- Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999a) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285: 751–753
- Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999b) A combined algorithm for genome-wide prediction of protein function. *Nature* 402: 83–86
- Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999c) A combined algorithm for genome-wide prediction of protein function [see comments]. *Nature* 402: 83–86
- McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Gaussian docking functions. *Biopolymers* 68: 76–90
- McGregor MJ, Muskal SM (1999) Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J Chem Inf Comput Sci* 39: 569–574
- Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc Natl Acad Sci USA* 102: 10930–10935
- Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A (2006) Human protein reference database–2006 update. *Nucleic Acids Res* 34: D411–D414
- Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B, Bork P (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat Biotechnol* 21: 790–795

- Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Compu Chem* 19: 1639–1662
- Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database. *Nucleic Acids Res* 35: D224–D228
- Nedeva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol* 3: e405
- Ng SK, Zhang Z, Tan SH (2003a) Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* 19: 923–929
- Ng SK, Zhang Z, Tan SH, Lin K (2003b) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res* 31: 251–254
- Nicastro G, Menon RP, Masino L, Knowles PP, McDonald NQ, Pastore A (2005) The solution structure of the Josephin domain of ataxin-3: structural determinants for molecular recognition. *Proc Natl Acad Sci USA* 102: 10493–10498
- Noiroit P, Noiroit-Gros MF (2004) Protein interaction networks in bacteria. *Curr Opin Microbiol* 7: 505–512
- Nooren IM, Thornton JM (2003) Diversity of protein–protein interactions. *EMBO J* 22: 3486–3492
- Oda K, Kitano H (2006) A comprehensive map of the toll-like receptor signaling network. *Mol Syst Biol* 2: 2006 0015
- Oda K, Matsuoka Y, Funahashi A, Kitano H (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol* 1: 2005 0010
- Orchard S, Kerrien S, Jones P, Ceol A, Chatr-Aryamontri A, Salwinski L, Nerothin J, Hermjakob H (2007a) Submit Your Interaction Data the IMEx Way: a Step by Step Guide to Trouble-free Deposition. *Proteomics*: 28–34
- Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, De Las Rivas J, Prieto C, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H (2007b) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat Biotechnol* 25: 894–898
- Orengo CA, Thornton JM (2005) Protein families and their evolution—a structural perspective. *Annu Rev Biochem* 74: 867–900
- Oti M, Brunner HG (2007) The modular nature of genetic diseases. *Clin Genet* 71: 1–11
- Overbeek R, Fonstein M, D’Souza M, Pusch GD, Maltsev N (1999) Use of contiguity on the chromosome to predict functional coupling. In *Silico Biol* 1: 93–108
- Pacifico S, Liu G, Guest S, Parrish JR, Fotouhi F, Finley RL Jr (2006) A database and tool, IM Browser, for exploring and integrating emerging gene and protein interaction data for *Drosophila*. *BMC Bioinformatics* 7: 195
- Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, Ruepp A, Frishman D (2005) The MIPS mammalian protein–protein interaction database. *Bioinformatics* 21: 832–834
- Pagel P, Oesterheld M, Tovstukhina O, Strack N, Stumpflen V, Frishman D (2007) DIMA 2.0 predicted and known domain interactions. *Nucleic Acids Res* 36: D651–D655

- Pagel P, Wong P, Frishman D (2004) A domain interaction map based on phylogenetic profiling. *J Mol Biol* 344: 1331–1346
- Pages S, Belaich A, Belaich JP, Morag E, Lamed R, Shoham Y, Bayer EA (1997) Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: prediction of specificity determinants of the dockerin domain. *Proteins* 29: 517–527
- Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. *Nat Biotechnol* 24: 805–815
- Park J, Lappe M, Teichmann SA (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol* 307: 929–938
- Parrish JR, Yu J, Liu G, Hines JA, Chan JE, Mangiola BA, Zhang H, Pacifico S, Fotouhi F, Dirita VJ, Ideker T, Andrews P, Finley RL Jr (2007) A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol* 8: R130
- Pasek S, Bergeron A, Risler JL, Louis A, Ollivier E, Raffinot M (2005) Identification of genomic features using microsynteny of domains: domain teams. *Genome Res* 15: 867–874
- Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. *Science* 300: 445–52
- Pazos F, Ranea JA, Juan D, Sternberg MJ (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol* 352: 1002–1015
- Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng* 14: 609–614
- Pearlman RS (1987) Rapid generation of high quality approximate 2-dimension molecular structures. *Chem Des Auto News* 2: 1–6
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96: 4285–4288
- Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, Elshamy WM, Rual JF, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Sole X, Hernandez P, Lazaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, Livingston DM, Gruber SB, Parvin JD, Vidal M (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet* 39: 1338–1349
- Puntrevoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 31: 3625–3630
- Raghavachari B, Tasneem A, Przytycka TM, Jothi R (2008) DOMINE: a database of protein domain interactions. *Nucleic Acids Res* 36: D656–D661
- Ralser M, Albrecht M, Nonhoff U, Lengauer T, Lehrach H, Krobitsch S (2005) An integrative approach to gain insights into the cellular function of human ataxin-2. *J Mol Biol* 346: 203–214
- Ramírez F, Schlicker A, Assenov Y, Lengauer T, Albrecht M (2007) Computational analysis of human protein interaction networks. *Proteomics* 7: 2541–2552
- Rarey M, Degen J, Reulecke I (2007) Docking and scoring for structure-based drug design. In: Lengauer T (ed) *Bioinformatics – from genomes to therapies*. Wiley-VCH, Weinheim, pp 541–600
- Rarey M, Dixon JS (1998) Feature trees: a new molecular similarity measure based on tree matching. *J Comput Aided Mol Des* 12: 471–490
- Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 261: 470–489

- 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
- Riley R, Lee C, Sabatti C, Eisenberg D (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol* 6: R89
- Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, Waegele B, Schmidt T, Doudieu ON, Stumpflen V, Mewes HW (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res* 36: D646–D650
- Ruffner H, Bauer A, Bouwmeester T (2007) Human protein–protein interaction networks and the value for drug discovery. *Drug Discov Today* 12: 709–716
- Sadowski J, Gasteiger J, Klebe G (1994) Comparison of automatic three-dimensional models builders using 639 X-ray structures. *J Chem Inf Comput Sci* 34: 1000–1008
- Saito R, Suzuki H, Hayashizaki Y (2002) Interaction generality, a measurement to assess the reliability of a protein–protein interaction. *Nucleic Acids Res* 30: 1163–1168
- Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32: D449–D451
- Santonico E, Castagnoli L, Cesareni G (2005) Methods to reveal domain networks. *Drug Discov Today* 10: 1111–1117
- Sato T, Yamanishi Y, Horimoto K, Kanehisa M, Toh H (2006) Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions. *Bioinformatics* 22: 2488–2492
- Sato T, Yamanishi Y, Kanehisa M, Toh H (2005) The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* 21: 3482–3489
- Schlicker A, Huthmacher C, Ramirez F, Lengauer T, Albrecht M (2007) Functional evaluation of domain–domain interactions and human protein interaction networks. *Bioinformatics* 23: 859–865
- Schuster-Bockler B, Bateman A (2007) Reuse of structural domain–domain interactions in protein networks. *BMC Bioinformatics* 8: 259
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498–2504
- Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* 102: 1974–1979
- Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. *Mol Syst Biol* 3: 88
- Sherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem* 49: 534–553
- Snel B, Huynen MA (2004) Quantifying modularity in the evolution of biomolecular systems. *Genome Res* 14: 391–397
- Sousa SF, Fernandes PA, Ramos MJ (2006) Protein–ligand docking: Current status and future challenges. *Proteins* 65: 15–26
- Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 100: 12123–12128
- Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein–protein interaction. *J Mol Biol* 311: 681–692
- Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34: D535–D539
- Stein A, Russell RB, Aloy P (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res* 33: D413–D417
- Stelzl U, Wanker EE (2006) The value of high quality protein–protein interaction networks for systems biology. *Curr Opin Chem Biol* 10: 551–558

- Suderman M, Hallett M (2007) Tools for visually exploring biological networks. *Bioinformatics* 23: 2651–2659
- 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- β signaling network. *Mol Cell* 13: 469–482
- Uetz P (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403: 623–627
- Uetz P, Dong YA, Zeretzke C, Atzler C, Baiker A, Berger B, Rajagopala SV, Roupelieva M, Rose D, Fossum E, Haas J (2006) Herpesviral protein networks and their interaction with the human proteome. *Science* 311: 239–242
- Uetz P, Rajagopala SV, Dong YA, Haas J (2004) From ORFeomes to protein interaction maps in viruses. *Genome Res* 14: 2029–2033
- Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 8: R39
- Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA (2004) Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 14: 208–216
- von Brunn A, Teepe C, Simpson JC, Pepperkok R, Friedel CC, Zimmer R, Roberts R, Baric R, Haas J (2007) Analysis of intraviral protein–protein interactions of the SARS coronavirus ORFeome. *PLoS ONE* 2: e459
- von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 35: D358–D362
- 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
- Wang H, Segal E, Ben-Hur A, Li QR, Vidal M, Koller D (2007a) InSite: a computational method for identifying protein–protein interaction binding sites on a proteome-wide scale. *Genome Biol* 8: R192
- Wang R, Lu Y, Wang S (2003) Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem* 46: 2287–2303
- Wang RS, Wang Y, Wu LY, Zhang XS, Chen L (2007b) Analysis on multi-domain cooperation for predicting protein–protein interactions. *BMC Bioinformatics* 8: 391
- Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction and encoding rules. *J Chem Inf Comput Sci* 28: 31–36
- Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 450: 1001–1009
- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 35: D5–D12
- Winzler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW (1999) Functional char-

- acterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285: 901–906
- Wojcik J, Schachter V (2001) Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17(Suppl 1): S296–S305
- Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H, Gerstein M (2004) Analyzing cellular biochemistry in terms of molecular networks. *Annu Rev Biochem* 73: 1051–1087
- Xu J, Li Y (2006) Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 22: 2800–2805
- Xue H, Xian B, Dong D, Xia K, Zhu S, Zhang Z, Hou L, Zhang Q, Zhang Y, Han JD (2007) A modular network model of aging. *Mol Syst Biol* 3: 147
- Xue L, Godden JW, Bajorath J (2000) Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J Chem Inf Comput Sci* 40: 1227–1234
- Yaffe MB (2006) “Bits” and pieces. *Sci STKE* 2006: pe28
- Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M (2007) Drug-target network. *Nat Biotechnol* 25: 1119–1126
- Yook SH, Oltvai ZN, Barabasi AL (2004) Functional and topological characterization of protein interaction networks. *Proteomics* 4: 928–942
- Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 3: e59
- Zarrinpar A, Bhattacharyya RP, Lim WA (2003) The structure and function of proline recognition domains. *Sci STKE* 2003: RE8

SECTION 7

Infrastructure for distributed protein annotation

CHAPTER 7

Infrastructure for distributed protein annotation

G. A. Reeves^{1,*}, A. Prlic^{2,*}, R. C. Jimenez^{1,3}, E. Kulesha¹ and H. Hermjakob¹

¹European Molecular Biology Laboratory Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

²The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

³Bioinformatics and Genomics Department, Centro de Investigación Principe Felipe (CIPF), Valencia, Spain

1 Introduction

Understanding human variation and disease often requires knowledge of a broad array of biomolecular data items, down to the role of an individual amino acid in a protein, and how mutations or alternative splicing events can change function and phenotype. There are a number of key databases that collect biomolecular information; the EMBL DNA database (Cochrane et al. 2006) and Ensembl (Flicek et al. 2007) collect annotations on genomic sequence features, the UniProt knowledge base (Bairoch et al. 2005) provides detailed annotation on protein sequences, and the Worldwide PDB member databases (Berman et al. 2007) provide protein structural information. Whilst these databases house a great deal of information on sequences and structures, the advent of high throughput methods in genome sequencing and structural genomics initiatives has produced an explosion in the quantity of uncharacterised data. As a result, the development of tools which annotate these sequences and structures by prediction or transfer of information from homologous relatives has also increased in number and diversity. These methods are crucial in order to fill in the functional space between characterised and uncharacterised protein sequences and structures.

Many computational biology laboratories specialise in different aspects of proteome annotation for a range of features and processes (Table 1). However, these tools are numerous, ever changing and located all over the world, often with more than one method annotating a similar feature. It has become important to provide ways in which

*These authors contributed equally.

Corresponding author: Henning Hermjakob, European Molecular Biology Laboratory Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK (e-mail: hhe@ebi.ac.uk)

Table 1 Selection of tools and resources which display their feature annotations using DAS

Method	Description
Protein Function	
FunCut (Abascal and Valencia 2003)	Automatic annotation of protein function based on family identification
Catalytic Site Atlas (Torrance et al. 2005)	A database documenting enzyme active sites and catalytic residues in enzymes with known 3D structure
Domain Annotations	
CATH (Pearl et al. 2005)	A database of protein domains in a hierarchical classification: Class, Architecture, Topology and Homology
InterPro (Mulder et al. 2005)	A database of protein families, domains and functional sites
Pfam (Finn et al. 2006)	Multiple sequence alignments and hidden Markov models covering many common protein domains and families
SMART (Letunic et al. 2006)	Simple Modular Architecture Research Tool
Prosite (Hulo et al. 2006)	A database of protein families and domains
Prints (Attwood et al. 2003)	A compendium of protein fingerprints
Protein Structure Prediction and Comparison	
Threader (Jones et al. 2005a)	Fold recognition
PSIpred (Bryson et al. 2005)	Protein structure prediction server
Post-translational Modifications of Proteins	
NetPhos (Blom et al. 1999)	Neural network predictions for serine, threonine and tyrosine phosphorylation sites in eukaryotic proteins
NetOGlyc (Julenius et al. 2005)	Neural network predictions for mucin type GalNAc O-glycosylation sites in mammalian glycoproteins
Protein Sorting	
SignalP (Bendtsen et al. 2004)	Prediction of the presence and location of signal peptide cleavage sites in amino acid sequences
TargetP (Emanuelsson et al. 2000)	Predicts the subcellular location of eukaryotic proteins
Transmembrane Predictions	
TMHMM	Prediction of transmembrane helices in proteins
Memsat (Jones et al. 1994)	Predicts the secondary structure and topology of all-helix integral membrane proteins

these annotations can be collated into a single view, providing as much information as we know about a particular sequence or structure from all disparate methods in one location. In the BioSapiens consortium, 19 groups from Europe and Israel provide annotation of protein features as part of their activities, ranging from highly specialised prediction methods for a specific kind of protein annotation to dozens of automatic methods providing millions of annotations for a large part of the known proteins. Due to continuous updates in methodology, these annotations are also frequently updated. The quality control, integration and continuous maintenance of such a broad array of annotations in a central resource like UniProt is almost impossible.

2 The Distributed Annotation System (DAS)

The Distributed Annotation System (DAS) (Dowell et al. 2001) provides a practical solution to this problem. Originally developed by Lincoln Stein for the collaborative annotation of genome sequences, the DAS protocol has been adapted for protein annotation by the BioSapiens consortium. A central *reference server* provides the reference set of protein sequences, based on the UniProt knowledge base. Each protein is uniquely identified by its UniProtKB accession number. Each participating laboratory, independent of geographical location, provides protein annotation relative to the UniProt reference set, through one or more locally installed *annotation servers*. The reference server and all annotation servers are listed in the *DAS registry* (Prlic et al. 2007, www.dasregistry.org). A *DAS client*, once activated by a user, connects to the DAS registry and retrieves the internet address of the reference server and all known annotation servers. Now, the user enters a specific protein accession number, and the DAS client retrieves the protein sequence from the reference server, and potentially hundreds of annotations for this particular protein from dozens of annotation servers distributed across the internet. Annotation items can be *positional*, for example a functional domain extending from amino acids 52 to 184 of the query protein, as well as *non-positional*, for example a literature reference pertinent to the entire protein. DAS annotations can contain a link, usually back to the original source, leading to more detailed information. The DAS client displays the retrieved annotations relative to the protein sequence, usually in a graphically attractive manner.

Using the Distributed Annotation System, information from dozens of different, geographically disparate sources can be centrally displayed using a technically simple protocol and minimal central infrastructure. Centralised databases do not need to invest time and resources to resolve contradictions between different third-party annotations as all information are reported, allowing the user to interpret the results independently. DAS allows annotations to be viewed in a central location while the control remains with the data provider. The DAS protocol has been chosen as the central data integration strategy by the BioSapiens Network of Excellence. Within the BioSapiens project, the scope of the DAS protocol has been extended from the original genomic sequence annotation to protein sequence and structure annotation as well as alignments. As of December 2007, annotation servers from 19 partner sites provide 69 different distributed annotation sources. This comprises information for genomic sequences and protein sequences as well as for protein structures.

3 DAS infrastructure

The independence of DAS servers and clients, only linked through a well defined protocol, allows the development of independent, specialised server and client software

SEARCH
 Protein ID: Registry label:

Examples: P05067, P03973, P13569, MDM2_MOUSE, BRCA1_HUMAN, ...

CHECKING
 Annotation servers loaded: **System information:**
 ... loading features from sdm

SEQUENCE
MANIPULATION OPTIONS
GRAPHIC

TYPES	FEATURE ANNOTATIONS	DAS SOURCES
SIGNAL		signalp
PHOSPHORYLATION (S)		netphos
PHOSPHORYLATION (T)		netphos
PHOSPHORYLATION (Y)		netphos
signal peptide (SO:0000418)		uniprot
mature protein region		uniprot
mature protein region		uniprot
mature protein region		uniprot
mature protein region		uniprot
mature protein region		uniprot
mature protein region		uniprot
mature protein region		uniprot
mature protein region		uniprot
active peptide (SO:0001064)		uniprot
extramembrane (SO:0001072)		uniprot
transmembrane (SO:0001077)		uniprot
polypeptide domain		uniprot
polypeptide region		uniprot

Fig. 1 Dasty2 loading data from annotation servers. The section “CHECKING” shows the progress bar, while the section “GRAPHIC” already shows retrieved data. Note the two “SIGNAL” features (feature 1 and 4)

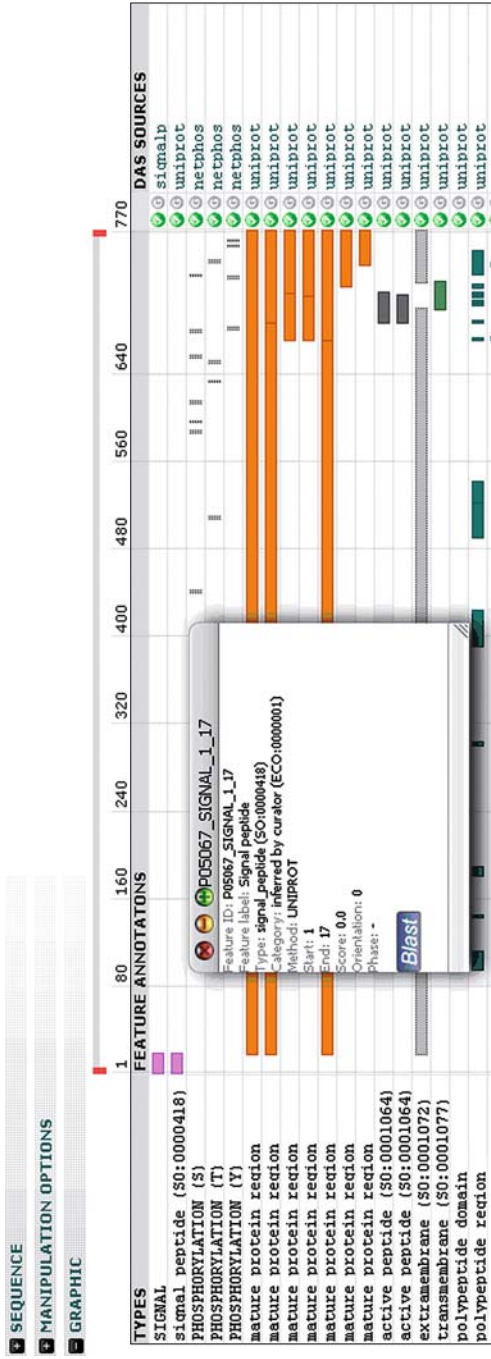


Fig. 2 The second signal feature annotation has been moved below the first one by drag-and-drop. Note that the first feature is provided by “signalp”, a prediction server, while the second feature is provided by UniProt, based on experimental evidence

packages, often distributed as open source software. In this section, we will briefly present key elements of the DAS infrastructure, with a focus on DAS clients as the end user interface of DAS.

3.1 DASTY2 – a protein sequence-oriented DAS client

DASTY2 (<http://www.ebi.ac.uk/dasty>) is a highly interactive protein DAS client. For a given protein sequence accession number, it retrieves the reference sequence and annotations provided by servers listed in the DAS registry. In contrast to its predecessor, DASTY1 (Jones et al. 2005b), DASTY2 displays features dynamically, as soon as the first annotation servers provide data, the features are displayed, rather than waiting until all servers have provided data. Figure 1 shows DASTY2 in the process of loading data from annotation servers. All positional features are aligned to the reference sequence, thus allowing to visually compare annotations. Annotations can be reordered, either in alphanumeric order for the different columns, or manually through drag-and-drop. Figure 2 shows the same sequence as the previous figure, but with the second signal feature annotation moved directly below the first signal annotation. In addition, the popup window for the detailed description of the second signal feature has been activated.

3.2 SPICE – a protein structure-oriented DAS client

SPICE (Prlic et al. 2005) (<http://www.efamily.org.uk/software/dasclients/spice/>) is a DAS client that can visualize protein sequence and structure information. It provides a 3D viewer that allows investigating annotations mapped onto the 3D protein structure, as well as several sequence panels that display the sequences and annotations for UniProt, matching Ensembl proteins and the sequence of the currently displayed chain of the PDB protein structure. The data is integrated, so whenever a sequence region is selected, it is projected onto the other sequences and can also be viewed in the 3D structure. In Fig. 3 two non-synonymous Single Nucleotide Polymorphisms (SNPs) have been selected on the Ensembl protein sequence. The position is projected through UniProt onto the PDB.

Besides this single-object display, SPICE can also be used to display 3D protein structure alignments. This feature has been used to display the results from the Critical Assessment of Techniques for Protein Structure Prediction (CASP-7) experiment (Moult et al. 2007). In this experiment the protein structure prediction community attempts to predict the three dimensional conformation of experimentally obtained protein structures, without knowing the structure at that point in time. SPICE can show the results of these predictions and allows to compare them with the solved protein structure. In Fig. 4 the experimentally obtained structure is shown in white and 2 predictions from different groups are shown in yellow and green. SPICE

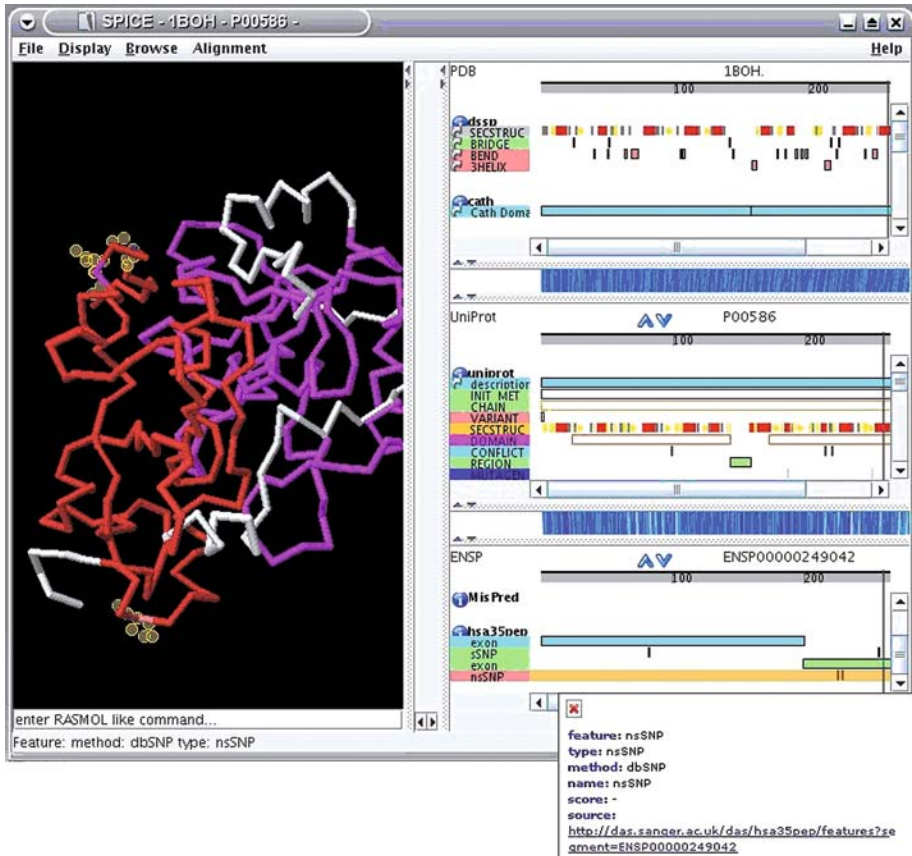


Fig. 3 SPICE projection of protein sequence features onto the three-dimensional protein structure. Two non-synonymous Single Nucleotide Polymorphisms (SNPs) have been selected on the Ensembl protein sequence. The position is projected through UniProt onto PDB

supports switching between 3 different alignment algorithms, which have been used to evaluate the predictions. This data is made available in a precalculated way *via* DAS from <http://www.predictioncenter.org> and SPICE obtains all the data from there.

Another application of SPICE is to visualize the 3D protein structure alignments as they are provided by the SISYPHUS database (Andreeva et al. 2007). SISYPHUS provides a set of manually curated alignments of proteins that have non-trivial relationships and that pose problems for most of the current standard alignment algorithms. SPICE can visualize these manually curated alignments, with the alignment data being obtained from <http://sisyphus.mrc-cpe.cam.ac.uk>.

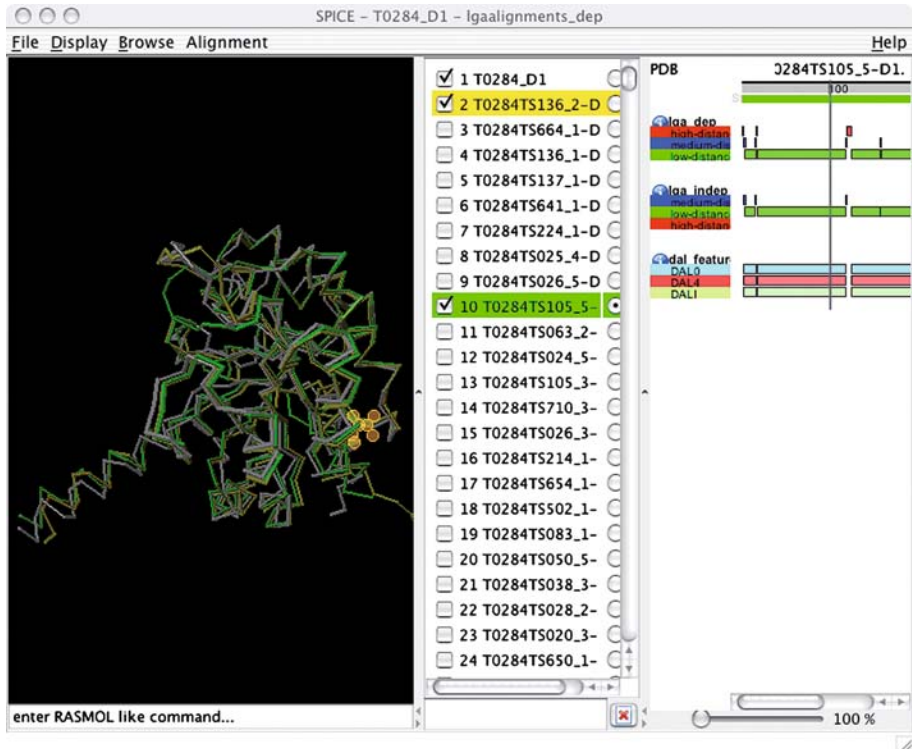


Fig. 4 SPICE as a viewer for results of the CASP contest. The experimentally obtained structure is shown in white and 2 predictions from different groups are shown in yellow and green

3.3 Ensembl

Ensembl (Flicek et al. 2007) (<http://www.ensembl.org>) is a comprehensive genome information system that is an example of a web server that itself is a DAS client and a DAS server. While Ensembl provides a very large database at its core, this data can be enriched with data obtained *via* DAS from external sources. In many Ensembl views it is possible to integrate the external data together with data obtained from the Ensembl database thus enabling the analysis of user provided data in context of the Ensembl genome information as well as information from other laboratories.

In addition to a list of predefined DAS data sources, the user can also add custom DAS data sources, based on the well-defined DAS reference coordinate systems, for

Fig. 5 Using the DAS STYLESHEET command it is possible to configure how a track will be displayed in a DAS client. Here several examples are shown how this can be used in Ensembl to display color gradients, histograms and line-plots

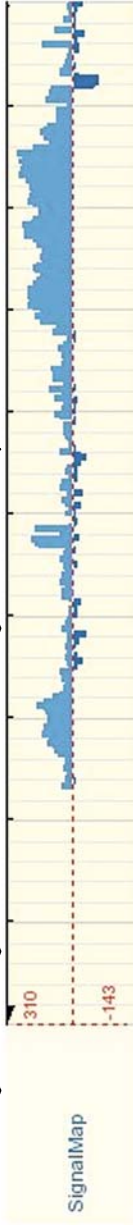
- **GRADIENT** - a color gradient where each sequence position has the same height.



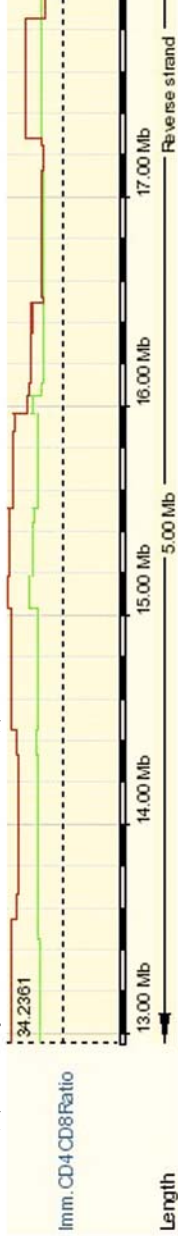
- **HISTOGRAM** - Each sequence position has a height depending on the feature score. An example where the minimum score has a height of 0:



Here a histogram with negative values. Note how the negative scores are plotted as bars below 0.



- **LINEPLOT** - displaying a function line. Note how this example displays an overlay of 2 functions. This can be done by providing features of two different types.



- **TILING** - display of tiling array data. This is essentially the same as a histogram with negative values, but there is some performance gains to be made in ensembl by specifying this as a separate type.



example Ensembl gene ID or UniProt accession number. It is even possible to upload user-provided data to the site and display this in the appropriate context on the Ensembl web page.

Also a growing number of datasets from the Ensembl database are being presented as DAS sources, which makes Ensembl also a DAS server. It makes it possible for other developers to create their own DAS clients displaying Ensembl data without the need to install the full Ensembl dataset.

One of the DAS extensions that was introduced as part of the BioSapiens project was a convention for how to deal with big quantitative data sets *via* DAS. DAS provides a *STYLESHEET* command that can configure how the data from a DAS server should be displayed in a DAS client. This can be used to display data as histograms, data plots, color gradient and any function both in the Ensembl and SPICE DAS clients (Fig. 5).

3.4 DAS servers

Providing biological sequence annotation to the scientific community through DAS can be a surprisingly easy task. As described above, the Ensembl web site allows direct upload of user data. In addition, a number of DAS servers are available as open source software, and installing them is often easier than installing a standard web server. Examples of freely available, well-tested servers are:

- ProServer (<http://www.sanger.ac.uk/Software/analysis/proserver/>) (Finn et al. 2007) (Perl-based)
- Dazzle (<http://www.derkholm.net/thomas/dazzle/>) (Java-based)
- myDAS (<http://code.google.com/p/mydas/>) (Java-based)

Once installed, these servers can be configured to serve local data either through database connectors from local databases, or even from tab-delimited files. To make the data widely available, new servers should be registered with the DAS registry at <http://www.dasregistry.org>.

4 The protein feature ontology

Whilst the independence of each annotation server is a major strength of the DAS protocol, it has also resulted in a major weakness. Each annotation type provided by a server is characterised by a *feature type*, defining the kind of information described in the feature, for example a glycosylation site. However, the DAS protocol does not define the possible feature types. Thus, a specific type of glycosylation might be predicted by two independent servers, once named “N-glycosylation”, once named “Glycosylation (N)”. While it is desirable to be able to compare annotation of the same functional

property from independent sources, it would be technically difficult even to display the two features next to each other, due to the different naming. With the advent of dozens of annotation servers, such differences became a major obstacle to the efficient analysis of DAS data. Efficient DAS data analysis, combining and comparing annotations from many different sources, inherently relies on the consistent organisation and presentation of the data displayed.

To address this challenge, the BioSapiens consortium has developed a hierarchical controlled vocabulary or ontology¹ of terms to clearly designate protein annotation types. The protein feature ontology is a composite ontology comprising selected terms from the Sequence Ontology (Eilbeck et al. 2005), the MOD protein modification ontology and some BioSapiens specific terms. The terms describe the features which make up protein function and form, from chemical modifications of amino acids *via* structural motifs such as helix-turn-helix to overall tertiary structure marking a globular domain. It is divided into two parts: *Positional* terms which refer to a specific residue or range of residues in the protein and *non-positional* terms which refer to the whole protein sequence or structure. The positional terms are features that are located on the sequence. These terms can also be found in the Sequence Ontology. The main categories of this section are:

- *polypeptide_region* describing a continuous sequence or single residue in a reference/mature protein sequence. Within this category lies the term *polypeptide_domain*, describing a structurally or functionally defined protein region which has been shown to recur throughout evolution.
- *biochemical_region* including *post_translational_modifications* (linking to the MOD ontology), *catalytic_residues* which are involved in the catalytic mechanism of enzymes and *molecular_contact_regions* indicating those residues which help to bind ligands or metal ions.
- *mature_protein_product* and *immature_peptide_region* categories distinguish the final folded peptide from regions which are cleaved during the mature protein folding process.
- *structural_region* which describes the backbone conformation of the polypeptide and includes child terms to describe both secondary structure and membrane structure.
- *polypeptide_variation_site* indicates alternative sequence due to naturally occurring events such as polymorphisms and alternative splicing or experimental methods such as site directed mutagenesis.

¹ The protein feature ontology has been developed to clearly designate annotation types and facilitate display and analysis of protein features. It is not an ontology as used in computer science, allowing automated reasoning. As often practised in molecular biology, we are using the term “ontology” as a synonym for “hierarchical controlled vocabulary”.

- *polypeptide_sequencing_information* clusters annotations which report differing results in experimental sequence determination.

Non-positional terms are not located to a particular region of the sequence, instead these annotations provide a description of the properties of the whole protein such as *publication* or provide links to other sources of non positional information such as *GO_term_annotation* or *EC_annotation*. The protein feature ontology currently comprises approximately 140 terms: 100 positional and 40 non positional terms.

Through the introduction of systematic naming of feature types according to the protein feature ontology, the annotations provided by participating annotation servers can be systematically compared and displayed, and analyzed in a user-friendly way, grouping related features provided by many servers close to each other in the display on the client, as shown in Fig. 6.

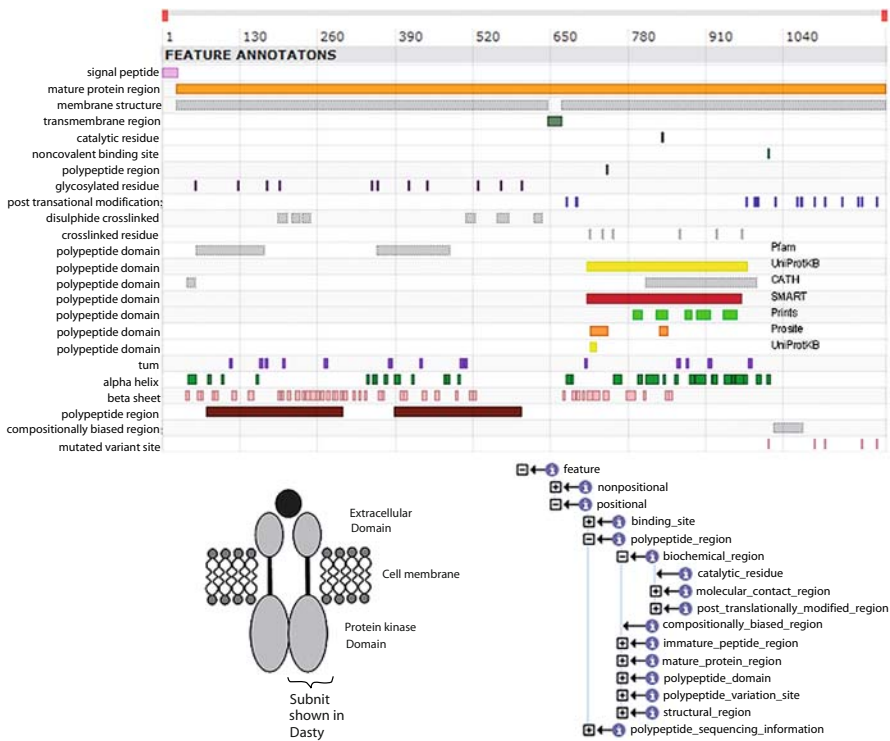


Fig. 6 Illustrating the BioSapiens protein feature ontology. The diagram depicts selected features for the human epidermal growth factor receptor (UniProtKB/SwissProt accession P00533) annotated by members of the BioSapiens Consortium and displayed using Dasty2. The top level terms in the ontology are shown using OBO edit (Day-Richter et al. 2007) in the bottom right hand of the diagram

5 Conclusion

DAS is a powerful system for data exchange between remote sites over the internet. It separates visualization from the actual data, thus making it much easier to show data distributed over multiple sites. It can be used to access the latest versions of data, without the need for local installations. It can also be used to integrate local data into popular bioinformatics resources.

References

- Abascal F, Valencia A (2003) Automatic annotation of protein function based on family identification. *Proteins* 53: 683–692
- Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32: D226–D229
- Andreeva A, Prlic A, Hubbard TJ, Murzin AG (2007) SISYPHUS – structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res* 35: D253–D259
- Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 31: 400–402
- 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(Database Issue): D154–D159
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: signal P 3.0. *J Mol Biol* 340: 783–795
- Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 35: D301–D303
- Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 294: 1351–1362
- Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res* 33: W36–W38
- Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, Baldwin A, Bates K, Bhattacharyya S, Browne P, van den Broek A, Castro M, Duggan K, Eberhardt R, Faruque N, Gamble J, Kanz C, Kulikova T, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, McHale M, McWilliam H, Mukherjee G, Nardone F, Pastor MP, Sobhany S, Stoehr P, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R (2006) EMBL nucleotide sequence database: developments in 2005. *Nucleic Acids Res* 34: D10–D15
- Day-Richter J, Harris MA, Haendel M, Lewis S (2007) OBO-Edit – an ontology editor for biologists. *Bioinformatics* 23: 2198–2200
- Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L (2001) The distributed annotation system. *BMC Bioinformatics* 2: 7
- Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M (2005) The sequence ontology: a tool for the unification of genome annotations. *Genome Biol* 6: R44
- 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
- Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34: D247–D251
- Finn RD, Stalker JW, Jackson DK, Kulesha E, Clements J, Pettett R (2007) ProServer: a simple, extensible Perl DAS server. *Bioinformatics* 23: 1568–1570

- Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S (2007) Ensembl 2008. *Nucleic Acids Res* 35: D707–D714
- Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ (2006) The PROSITE database. *Nucleic Acids Res* 34: D227–D230
- Jones DT, Bryson K, Coleman A, McGuffin LJ, Sadowski MI, Sodhi JS, Ward JJ (2005a) Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins* 61(Suppl 7): 143–151
- Jones DT, Taylor WR, Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 33: 3038–3049
- Jones P, Vinod N, Down T, Hackmann A, Kahari A, Kretschmann E, Quinn A, Wieser D, Hermjakob H, Apweiler R (2005b) Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics* 21: 3198–3199
- Julienius K, Molgaard A, Gupta R, Brunak S (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* 15: 153–164
- Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34: D257–D260
- Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A (2007) Critical assessment of methods of protein structure prediction-Round VII. *Proteins* 69(Suppl 8): 3–9
- Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH (2005) InterPro, progress and status in 2005. *Nucleic Acids Res* 33: D201–D205
- Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res* 33: D247–D251
- Prlic A, Down TA, Hubbard TJ (2005) Adding some SPICE to DAS. *Bioinformatics* 21(Suppl 2): ii40–ii41
- Prlic A, Down TA, Kulesha E, Finn RD, Kahari A, Hubbard TJ (2007) Integrating sequence and structural biology with DAS. *BMC Bioinformatics* 8: 333
- Torrance JW, Bartlett GJ, Porter CT, Thornton JM (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J Mol Biol* 347: 565–581

SECTION 8

Applications

CHAPTER 8.1

Viral bioinformatics

B. Adams¹, A. Carolyn McHardy¹, C. Lundegaard² and T. Lengauer¹

¹Max-Planck-Institut für Informatik, Saarbrücken, Germany

²Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Kongens Lyngby, Denmark

1 Introduction

Pathogens have presented a major challenge to individuals and populations of living organisms, probably as long as there has been life on earth. They are a prime object of study for at least three reasons: (1) Understanding the way of pathogens affords the basis for preventing and treating the diseases they cause. (2) The interactions of pathogens with their hosts afford valuable insights into the working of the hosts' cells, in general, and of the host's immune system, in particular. (3) The co-evolution of pathogens and their hosts allows for transferring knowledge across the two interacting species and affords valuable insights into how evolution works, in general. In the past decade computational biology has started to contribute to the understanding of host-pathogen interaction in at least three ways which are summarized in the subsequent sections of this chapter.

Taking influenza as an example the computational analysis of viral evolution within the human population is discussed in Sect. 2. This evolutionary process takes place in the time frame of years to decades as the virus is continuously changing to evade the human immune system. Understanding the mechanisms of this evolutionary process is key to predicting the risk of emergence of new highly pathogenic viral variants and can aid the design of effective vaccines for variants currently in circulation.

Section 3 addresses the molecular basis of how such vaccines can be developed. Vaccines present the human immune system molecular with determinants of viral strains that elicit an immune response against the virus and activate the buildup of molecular immune memory without being pathogenic. That section also gives a succinct introduction to the workings of the human immune system.

Section 4 addresses the issue of highly dynamic viral evolution inside a single patient. Some viruses have the capability of this kind of evolution in order to evade the

Corresponding author: Thomas Lengauer, Max-Planck-Institut für Informatik, Campus E1 4, 66123 Saarbrücken, Germany (e-mail: lengauer@mpi-sb.mpg.de)

immune response of the host or the effects of a drug therapy. HIV is the example discussed here. Drug therapies against HIV become ineffective due to the virus evolving to a variant that evades the therapy. If this happens the therapy has to be replaced with another therapy that effectively targets the viral variant now present inside the patient.

2 Viral evolution in the human population

Influenza is a classic example of a pathogen that evades immunity at the population level. Due to a strong immune response in the host, which clears the virus within a few days, the virus can only survive by moving on quickly. Following an infection, hosts retain strong immunity to a particular antigenic type. As immunity accumulates in the population, there is increasing selection for pathogens with altered antigenic types that are less effectively recognized and thus have a higher probability of finding a susceptible host. By rapid evolution influenza is able to persist at relatively high prevalence in the human population. Consequently, vaccines must be frequently updated to ensure a good match with the circulating strain. However, even with current vaccination programs, endemic influenza remains a significant burden and is associated with an estimated 37,000 deaths in the U.S. alone.

In addition to the endemic activity, influenza pandemics occasionally occur when avian forms of the virus adapt to humans or provide genetic material that is incorporated into existing human forms. The antigenic novelty of these variants allows them to sweep through the global population, often causing severe disease. There were three such pandemics in the twentieth century. The most severe of them, the ‘Spanish Flu’ of 1918, resulted in 30 to 50 million deaths.

Thus, two key goals of influenza research are predicting viral evolution in the human population to determine optimum vaccine configurations and the early recognition of potential pandemic strains circulating in, or emerging from, the avian population. Large-scale genome sequencing and high-throughput experimental studies of influenza isolates from various sources have a central role in both of these endeavors.

2.1 Biology and genetics

Influenza viruses are single-stranded, negative sense RNA viruses of the family *Orthomyxoviridae* (Webster et al. 1992). Three phylogenetically and antigenically distinct types currently circulate, referred to as influenza A, B and C. All types infect humans and some other mammals. Influenza A also infects birds. This section will focus on influenza A, because of its high prevalence and increased virulence in humans, compared to types B and C.

The influenza A genome is composed of eight RNA segments totaling approximately 14 kb of sequence. The segments encode eleven proteins that are required for the

replication and infection cycle of the virus. The two major determinants recognized by the human immune system are the surface glycoproteins hemagglutinin (HA) and neuraminidase (NA). Hemagglutinin is responsible for binding to sugar structures on the epithelial cells lining the respiratory tract and entry into the cell during the first stage of infection. Neuraminidase plays a part in releasing assembled viral particles from an infected cell by cleaving terminal sugar structures from neighboring glycoproteins and glycolipids on the cell surface. Several subtypes of influenza A are distinguished on the basis of the antigenic properties of the HA and NA proteins. There are 16 known subtypes for HA and 9 for NA, all of which occur in birds. In humans, subtypes H2N2 and H3N8 have circulated in the past but currently only H3N2 and H1N1 are endemic. Of these H3N2 is more virulent and evolves more rapidly.

There are two distinct mechanisms by which the influenza genome evolves. One is the acquisition of mutations, deletions or insertions during the replication process. This occurs at a higher rate than for DNA-based viruses, as RNA polymerases do not possess a proof-reading mechanism. Some of these changes subsequently become fixed in the viral sequence, either through the random fixation process of genetic drift or because they confer a selective advantage. This gradual change and its impact on the phenotype level is referred to as antigenic drift. The second mechanism of evolution is reassortment (see Fig. 1). If two different strains simultaneously infect the same host, a novel strain may arise with a combination of segments from the two. The phenotypic change associated with the emergence of such a viral variant is referred to as antigenic shift.

2.2 Vaccine strain selection for endemic influenza

The human immune system primarily targets the hemagglutinin surface protein of the influenza virus. Whether primed by infection or vaccination, antibodies provide long lasting immunity to that particular HA configuration. However, due to antigenic drift

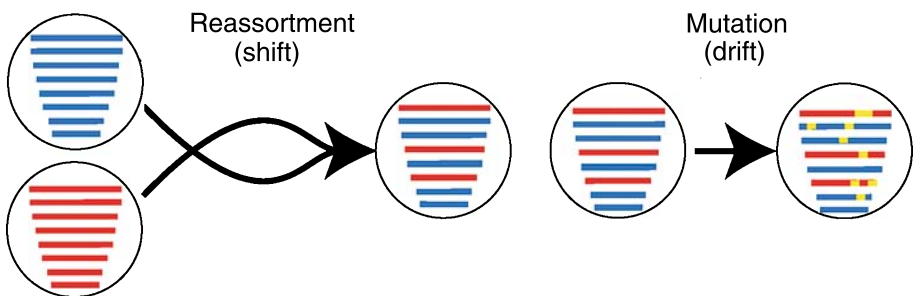


Fig. 1 Schematic representation of influenza evolution by reassortment (left) and mutation (right). Each viral genome is composed of 8 RNA segments. Reassortment of the 8 segments from two distinct viruses can result a new viable form of the virus. Drift occurs when errors during viral replication produce novel variants with small changes, i.e. insertions, deletions or mutations in the sequence segments

within just a few years those antibodies do not efficiently recognize the circulating HA. Influenza vaccines must thus be regularly updated and re-administered. The WHO makes vaccine recommendations based on the prevalence of recently circulating strains. If a new genotype, based on the HA segment, appears to be increasing in prevalence, then hemagglutination-inhibition (HI) assays using post infection ferret sera are carried out to determine whether this is associated with phenotypic change in terms of the antigenicity. If there is significant phenotypic change, the current vaccine is unlikely to be effective against the proposed emergent strain and must be updated. The genotype-phenotype map for influenza virus is unclear and genotyping is only used to choose candidate strains for HI assays. However, recent advances have indicated several ways in which genome-based methods may improve vaccine selection.

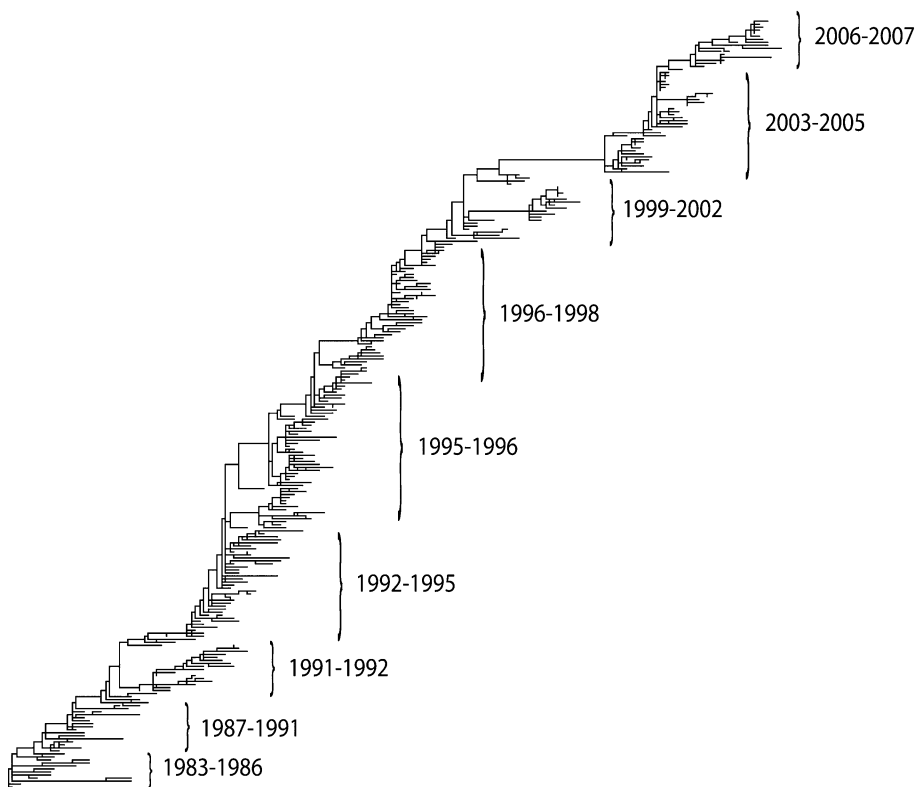


Fig. 2 Phylogenetic tree for the influenza HA coding sequences constructed by maximum parsimony using the software PAUP (<http://paup.csit.fsu.edu/>) from the sequences of 507 viruses isolated between 1983 and 2007. Dates to the right of the tree indicate the year that the majority of sequences contributing to that section were isolated. The tree has a distinctive cactus like shape characterized by constant turnover and limited diversity at any point in time

Bioinformatic analyses of the hemagglutinin encoding sequences have revealed characteristics of the evolutionary process and also determined relevant properties with respect to viral fitness. Phylogenetic trees of these sequences have a cactus-like topology (see Fig. 2). A diverse strain repertoire is periodically replaced by just a single strain, which constitutes the progenitor for all future lineages (Fitch et al. 1997). Population genetic theory states that such trees can be derived by random genetic drift if population size becomes very small or by selection if fitter variants emerge and periodically replace all others.

Further analyses of such trees led to the identification of a set of rapidly evolving codons in the antibody-binding and receptor-binding sites of the protein (Bush et al. 1999). These codons show a significantly higher ratio of synonymous to nonsynonymous substitutions than expected by chance, indicating that the driving force in the evolution of the HA gene is selection for variants that are fitter in terms of the evasion of host immunity acquired from previous infections. These positively selected for codons also possess predictive value with respect to the future fitness of a set of viral strains.

The relationship between the influenza genotype and phenotype has been elucidated by the application of multidimensional scaling to create a low dimensional representation of antigen-antibody distances measured with hemagglutinin inhibition assays (Smith et al. 2004). This showed that genotypes isolated over the same 2–5 year period cluster in phenotype space. Significant differences between clusters mostly localize to antibody-binding sites, the receptor-binding site and positively selected codons of the HA sequence. As more data become available, the combined analysis of genotypes and their relationship to the antigenic phenotype will enhance our capability to predict dominant circulating strains and estimate the efficacy of proposed vaccines.

2.3 Pandemic influenza

Antigenic drift allows partial immune evasion, but the host population, on average, always has some degree of immunity. Occasionally however, novel strains with no antigenic history cause global pandemics. In the twentieth century this happened in 1918, 1957 and 1968. Further pandemics are considered inevitable unless their origin can be rapidly detected or, better still, predicted (Taubenberger et al. 2007). Whole genome analysis has shown that the 1968 and 1957 pandemic strains were reassortants that introduced avian HA, PB1 and, in 1957 NA, segments into viruses already circulating in, and adapted to, the human population. The antigenic novelty of the 1918 pandemic strain also stems from its introduction from an avian source. Whether it crossed to humans directly from birds, circulated in swine first, or was a reassortment of existing avian and human strains remains a matter of debate.

Since 1997, the avian H5N1 subtype has been considered a serious candidate for a novel pandemic, due to a small but increasing number of human cases. This requires the avian HA protein to undergo adaptation to bind to human receptors. Analysis of the

viral genotypes responsible for the human H5N1 cases has identified several common amino acids changes in and around the binding region. It has also shown that the virus is repeatedly crossing directly from birds, without reassortment or sustained human to human transmission (2005). So far, an H5N1 strain with pandemic potential has not emerged, but continual surveillance is vital. Early detection of the accumulation of mutations that may facilitate a host switch, the mixing of genetic material from human and avian forms or evidence of human to human transmission will be critical for containment strategies

The efficiency of such surveillance measures may also be improved by targeting particular geographic regions. Based on a phylogenetic tree of avian H5N1 sequences, a phylogeography of significant migratory trajectories has been constructed for Eurasia by minimizing the number of migration events necessary to keep the phylogeny geographically consistent (Wallace et al. 2007). These data indicated that Indochina is a largely isolated subsystem in terms of H5N1 evolution and Guangdong in China is the main source of diversity and diffusion throughout Eurasia. It may therefore be practical to invest more of the surveillance effort into this region.

2.4 Conclusion

Even with modern medicine the burden of annual influenza is significant and the threat of a pandemic constantly hangs over the world. Vaccines, chemo prophylactics, detection and containment strategies are all in use. But the influenza virus, like malaria and HIV, is a constantly moving target and optimizing pharmaceutical design and public health policy is a complex problem requiring an integrated knowledge of, among other things, epidemiology, immunology and molecular biology. Bioinformatics has provided, and will continue to provide, vital insights in all of these areas.

3 Interaction between the virus and the human immune system

3.1 Introduction to the human immune system

The human immune system rests basically on two pillars. One pillar is solely genetically determined and remains unchanged throughout the life of an individual. This so-called innate immune system basically provides physical protection barriers and registers if generally recognizable foreign substances are entering the organism. If such substances are detected a fast and general protection mechanism sets in whose nature is determined by the type of substance registered. The innate protection mechanisms also include an activation of the other pillar of the immune system, the adaptive immune system. This part evolves during the life, and its present state is highly

dependent of the infection history of the individual. The adaptive immune system is itself basically split up in two parts. First the humoral immunity, which happens outside cells within the body liquids and is antibody-driven. Special immunoglobulin molecules (antibodies) mediate the humoral response. Antibodies are produced by B lymphocytes that bind to antigens by their immunoglobulin receptors, which is a membrane bound form of the antibodies. When the B lymphocytes become activated, they start to secrete the soluble form of the receptor in large amounts. Antibodies are Y-shaped, and each of the two branches functions independently and can be recombinantly produced and is then known as fragments of antibodies (Fab). The antibody can coat the surface of an antigen such as a virus and generally this will inactivate whatever undesired function the respective object may have, and facilitate the uptake of the antibody-bound object *via* phagocytosis by macrophages, which will then digest the object. Macrophages, B cells, and dendritic cells are all so-called professional antigen presenting cells (APC). They carry a special receptor named the major histocompatibility complex (MHC) class II. This receptor is able to present peptides derived from degraded phagocytosed proteins. Other cells (T cells) carries a receptor, the T cell receptor (TCR), which, if the T cell also carries a so called CD4+ receptor, is able to bind to MHC class II molecules presenting a foreign peptide, e.g. one not originating from the human proteome. Such an interaction will stimulate B cells to divide and further progress to produce more antibodies as well as survive for a long time as memory B cells. The presence of memory B cells enables the immune system to react faster in a subsequent infection by the same pathogen. The CD4+ T cells actually also belong to the second part of the adaptive immune system, which is the cellular immune system. Another important feature of cellular immunity regards T cells with the CD8 coreceptor (CD8+ T cells). The TCR of CD8+ T cells can recognize foreign peptides in complex with membrane bound MHC class I molecules on the outer side of nucleated cells. Such an interaction will activate the T cells to signal and induce cell death of the cell presenting the foreign peptide.

Both antibodies and TCRs are composed of a light and a heavy chain. These chains are translated from genes resulting from a genetic recombination of two and three genes, respectively, during the B-cell development in the bone marrow. These genes exist in several nonidentical duplicates on the chromosome and can be combined into a large number of different rearrangements. However, the molecular processes linking the genes are imprecise and involve generation of P (palindromic) nucleotides, addition of N (non-templated) nucleotides by terminal deoxynucleotidyl transferase (TdT) and trimming of the gene ends and therefore also play a major role in the generation of the huge diversity needed to be able to respond to any given pathogen. The T cells having a mature TCR are being validated in the thymus. The host will eliminate T cells having a TCR that is either unable to bind to an MHC:peptide complex or that will recognize an MHC with a peptide originating from the hosts proteome (self peptides). All the above is highly simplified text book immunology (Janeway 2005).

3.2 Epitopes

To be able to combat an infection the immune system must first recognize the intruder as foreign. The specific parts of the pathogen that is recognized and induces an immune response are called epitopes. Epitopes are often parts of larger macromolecules, which most often happen to be polypeptides and proteins. B-cell epitopes are normally classified into two groups: continuous and discontinuous epitopes. A continuous epitope, (also called a sequential or linear epitope) is a short peptide fragment in an antigen that is recognized by antibodies specific for the given antigen. A discontinuous epitope is composed of residues that are not adjacent in the amino acid sequence, but are brought into proximity by the folding of the polypeptide.

The cellular arm of the immune system consists as described of two parts; the CD8+ cytotoxic T lymphocytes (CTLs), and the CD4+ helper T lymphocytes (HTLs). CTLs destroy cells that present non-self peptides (epitopes). HTLs are needed for B cells activation and proliferation to produce antibodies against a given antigen. CTLs on the other hand perform surveillance of the host cells, and recognize and kill infected cells. Both CTLs and HTLs are raised against peptides that are presented to the immune cells by major histocompatibility complex (MHC) molecules, which are encoded in the most polymorphic mammalian genes. The human versions of MHCs are referred to as the human leucocyte antigens (HLA). The cells of an individual are constantly screened for presented peptides by the cellular arm of the immune system. In the MHC class I pathway, class I MHCs presents endogenous peptides to T cells carrying the CD8 receptor (CD8+ T cells). To be presented, a precursor peptide is normally first generated by cutting endogenous produced proteins inside the proteasome, a cytosolic protease complex. Generally, resulting peptides should bind to the TAP complex for translocation into the endoplasmic reticulum (ER). During or after the transport into the ER the peptide must be able to bind to the MHC class I molecule to invoke folding of the MHC before the complex can be transported to the cell surface. When the peptide:MHC complex is presented on the surface of the cell, it might bind to a CD8+ T cell with a fitting TCR. If such a TCR clone exists a CTL response will be induced and the peptide is considered an epitope. The most selective step in this pathway is binding of a peptide to the MHC class I molecule. As mentioned above, the MHC is the most polymorphic gene system known. The huge variety of protein variants brought forth by this polymorphism is a big challenge for T-cell epitope discoveries, enhancing the need for bioinformatical analysis and resources. It also highly complicates immunological bioinformatics, as predictive methods for peptide MHC binding have to deal with the diverse genetic background of different populations and individuals. On a population basis, hundreds of alleles (gene variants) have been found for most of the HLA encoding loci (1839 in release 2.17.0 of the IMGT/HLA Database, <http://www.ebi.ac.uk/imgt/hla/>). In a given individual either one or two different alleles are expressed per locus depending on whether the same (in homozygous individuals) or two different (in heterozygous individuals) alleles are present on the two

different chromosomes. Each MHC allele binds a very restricted set of peptides and the polymorphism affects the peptide binding specificity of the MHC; one MHC will recognize one part of the peptide space, whereas another MHC will recognize a different part of this space. The very large number of different MHC alleles makes reliable identification of potential epitope candidates an immense task if all alleles are to be included in the search. Many MHC alleles, however, share a large fraction of their peptide-binding repertoire and it is often possible to find promiscuous peptides, which bind to a number of different HLA alleles. The problem can thus be largely reduced by grouping all the different alleles into supertypes in a manner were all the alleles within a given supertype have roughly the same peptide specificity. This grouping generally requires some knowledge regarding the binding repertoire of either the specific allele or an allele with a very similar amino acid sequence.

The peptides recognized by the CD4+ T cells are called helper epitopes. These are presented by the MHC class II molecule, and peptide presentation on this MHC follow a different path than the MHC class I presentation pathway: MHC class II molecules associate with a nonpolymorphic polypeptide referred to as the invariant chain (Ii) in the ER. The Ii chain is a type II membrane protein, and unlike MHC molecules the C-terminal part of the molecule extends into the lumen of the ER. The MHC:Ii complex accumulates in endosomal compartments and here, Ii is degraded, while another MHC-like molecule, called HLA-DM in humans, loads the MHC class II molecules with the best available ligands originating from endocytosed antigens. The peptide:MHC class II complexes are subsequently transported to the cell surface for presentation to the CD4+ T helper cells. The helper T cells will bind the complex and be activated if they have an appropriate TCR.

3.3 Prediction of epitopes

A major task in vaccine design is to select and design proteins containing epitopes able to induce an efficient immune response. The selection can be aided by epitope prediction in whole genomes, relevant proteins, or regions of proteins. In addition, prediction of epitopes may help to identify the individual epitopes in proteins that have been analyzed and proven to be antigens using experimental techniques based on, e.g., Western blotting, immunohistochemistry, radioimmunoassay (RIA), or enzyme-linked immunosorbent assays (ELISA).

Today, the state-of-the-art class I T-cell epitope prediction methods are of a quality that makes these highly useful as an initial filtering technique in epitope discovery. Studies have demonstrated that it is possible to rapidly identify and verify MHC binders from upcoming possible threats with high reliability, and take such predictions a step further and validate the immunogenicity of peptides with limited efforts, as has been shown with the influenza A virus (see next subsection). It is also possible to identify the vast majority of the relevant epitopes in rather complex organisms using class I MHC

binding predictions and only have to test a very minor fraction of the possible peptides in the virus proteome itself. MHC class II predictions can be made fairly reliably for certain alleles. B-cell epitopes are still the most complicated task. However, some consistency between predicted and verified epitopes is starting to emerge using the newest prediction methods (Lundegaard et al. 2007).

B-cell epitope prediction is a highly challenging field due to the fact that the vast majority of antibodies raised against a specific protein interact with parts of the antigen that are discontinuous in the polypeptide sequence. The prediction of continuous, or linear, epitopes, however, is a somewhat simpler problem, and may be still useful for synthetic vaccines or as diagnostic tools. Moreover, the determination of continuous epitopes can be integrated into determination of discontinuous epitopes, as these often contain linear stretches. More successful methods combine scores from the Parker hydrophilicity scale and a position specific scoring matrix (PSSM) trained on linear epitopes. Different experimental techniques can be used to define conformational epitopes. Probably the most accurate and easily defined is using the solved structures of antibody–antigen complexes. Unfortunately, the amount of this kind of data is still scarce, compared to linear epitopes. Furthermore, for very few antigens all possible epitopes have been identified. The simplest way to predict the possible epitopes in a protein of known 3D structure is to use the knowledge of surface accessibility and newer methods using protein structure and surface exposure for prediction of B-cell epitopes have been developed. The CEP method calculates the relative accessible surface area (RSA) for each residue in the structure. The RSA is defined as the fraction of solvent exposed surface of a given amino acid in the native structure relative to the exposed surface the same amino acid placed centrally in a tri-peptide, usually flanked by glycines or alanines. It is then determined which areas of the protein are exposed enough to be antigenic determinants. Regions that are distant in the primary sequence, but close in three-dimensional space will be considered as a single epitope. DiscoTope (www.cbs.dtu.dk/services/DiscoTope) uses a combination of amino acid statistics, spatial information and surface exposure. The system is trained on a compiled dataset of discontinuous epitopes from 76 X-ray structures of antibody–antigen protein complexes. (Haste Andersen et al. 2006). B-cell epitope mapping can be performed experimentally by other methods than structure determination, e.g., by phage display. The low sequence similarity between the mimotope (i.e. a macromolecule, often a peptide, which mimics the structure of an epitope) identified through phage display and the antigen complicates the mapping back onto the native structure of the antigen, however, a number of methods have been developed that facilitate this.

A number of methods for predicting the binding of peptides to MHC molecules have been developed. The majority of peptides binding to MHC class I molecules have a length of 8–10 amino acids. Position 2 and the C-terminal position have turned out generally to be very important for the binding to most class I MHCs and these positions are referred to as anchor positions (Fig. 3). For some alleles, the binding motifs

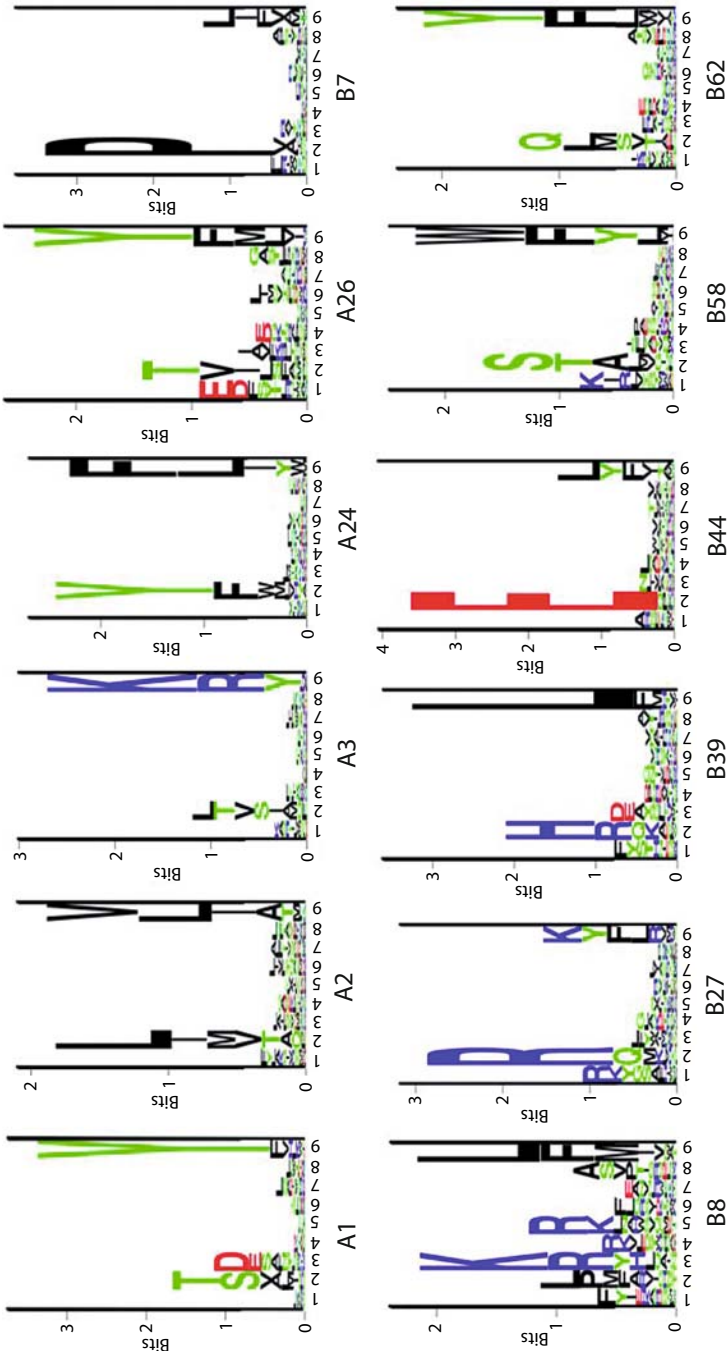


Fig. 3 Sequence Logos of raw count statistics of peptides measured to bind with a $I_{C_{50}}$ stronger than 500 nM. Peptides were mainly selected from the ImmuneEpitope database

have additional anchor positions. E.g., epitopes binding to the human HLA-A*0101 allele have positions 2, 3 and 9 as anchors (Rammensee et al. 1999) (Fig. 1A). The discovery of such allele-specific motifs led to the development of the first reasonably accurate algorithms. In these prediction tools, it is assumed that the amino acids at each position along the peptide sequence contribute a given binding energy, which can be added up to yield the overall binding energy of the peptide. Several of these matrix methods are trained on exclusively positive examples like peptides eluted from MHCs on living cells, peptides that have been shown to induce significant interferon gamma responses in CTL assays, or peptides that bind the MHC more strongly than a certain binding affinity value (usually below 500 nM). Other matrix methods, like the SMM method, aim at predicting an actual affinity and thus use exclusively affinity data. However, matrix-based methods cannot take correlated effects into account (when the binding affinity of peptide with a given amino acid at one position depends on amino acids that are present at other positions in the peptide). Higher-order methods like ANNs and SVMs are ideally suited for taking such correlations into account and can be trained with data either in the format of binder/non-binder classification, or with real affinity data. Some of the recent methods combine the two types of data and prediction methods. The different types of predictors are reviewed in (Lundegaard et al. 2007) and an extensive benchmark of the performances of the different algorithms have been published by (Peters et al. 2006).

Representing a supertype by a well-studied allele risks the confinement to selecting epitopes that are restricted to this allele, excluding other alleles within the supertype. Thus another, and potentially more rational approach, would be to select a limited set of peptides restricted to as many alleles as possible. This should be within reach with new methods that directly predict epitopes that can bind to different alleles (Brusic et al. 2002), or pan-specific approaches that can make predictions for all alleles, even those whose sequences are not yet known (Heckerman et al. 2007; Nielsen et al. 2007a). Finally, even though MHC binding is the most limiting step in the class I pathway the cleavage and transporting events are not insignificant. Several tools have been developed that integrate predictions of the different steps, and this has been shown to improve the predictions of actual CTL epitopes (Larsen et al. 2007).

Unlike the MHC class I molecules, the binding cleft of MHC class II molecules is open at both ends, which allows for the bound peptide to have significant overhangs in both ends. As a result MHC class II binding peptides have a broader length distribution even though the part of the binding peptide that interacts with the MHC molecule (the binding core) still includes only 9 amino acid residues. This complicates binding predictions as the identification of the correct alignment of the binding core is a crucial part of identifying the MHC class II binding motif. The MHC class II binding motifs have relatively weak and often degenerate sequence signals. While some alleles like HLA-DRB1*0405 show a strong preference for certain amino acids at the anchor positions, other alleles like HLA-DRB1*0401 allow basically all amino acids at all

positions. In addition, there are other issues affecting the predictive performance of most MHC class II binding prediction methods. The majority of these methods take as a fundamental assumption that the peptide:MHC binding affinity is determined solely by the nine amino acids in binding core motif. This is clearly a large oversimplification since it is known that peptide flanking residues (PFR) on both sides of the binding core may contribute to the binding affinity and stability. Some methods for MHC class II binding have attempted to include PFRs indirectly, in terms of the peptide length, in the prediction of binding affinities. It has been demonstrated that these PFRs indeed improve the prediction accuracy (Nielsen et al. 2007b).

3.4 Epitope prediction in viral pathogens in a vaccine perspective

As described in Sect. 2 some of the important B cell antigens vary significantly between different influenza A viral strains. Current influenza vaccines are based on inactivated influenza virus and thus mimic only the B cell response obtained by a fully infection competent strain. This has the drawback that only closely related strains will be covered by this response and new vaccines have to be produced annually as a result of the antigenic drift (see Sect. 2). Thus the ideal influenza vaccine will raise an immune response against parts of the pathogen that are conserved between as many strains as possible. To identify these parts the described prediction tools will be an invaluable help. Initial *in silico* scans of the viral genome for potential immunogenic parts will reduce the potential epitope space, and thus make experimental validations feasible. In a published example all genomic sequenced strains of H1N1 were scanned for CTL epitopes. Only 9-mer peptides in the influenza proteome that were at least 70% conserved in all strains were considered. The top 15 predicted epitopes for each of the 12 supertypes were subsequently selected to be synthesized for further validation. Because of the limited size of the influenza genome and the high variability of some of the proteins the conservation criteria resulted in relatively low prediction scores of some of the chosen peptides. 180 peptides were selected and 167 were synthesized and further validated for MHC binding and CTL response. The fraction of validated MHC binding peptides (with a binding affinity of below 500 nM) was relatively low (about 50%) compared to some other studies (60–75%) (Sundar et al. 2007; Sylvester-Hvid et al. 2004), but 13 of the 89 binding peptides, or 15%, gave a positive output in a CTL recall assay. Obviously, the conserved epitopes were found in the less variable proteins, but the large majority of the validated epitopes (85%) turned out to be 100% conserved not only in H1N1 strains but also in the H5N1 avian strains that in the last few years have infected humans resulting in severe symptoms and high mortality (Wang et al. 2007). Such epitopes can be highly valuable starting points for vaccine development. Even though cellular immunity does not protect against infections it might protect against a fatal outcome of an infection with a new aggressive strain.

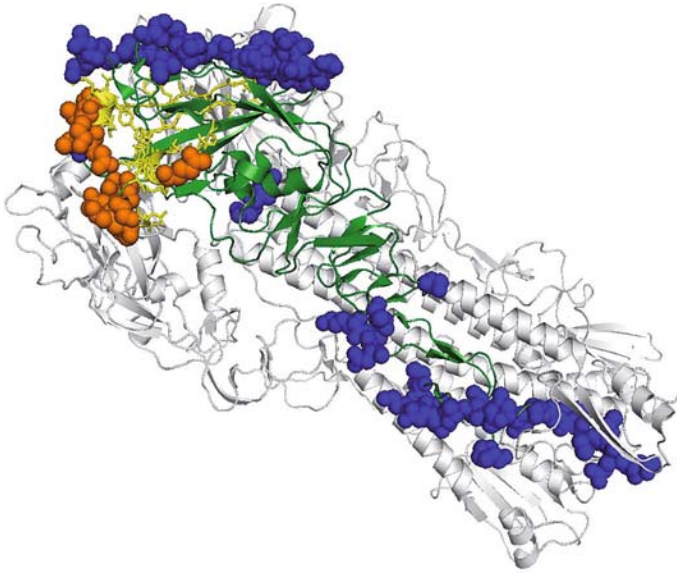


Fig. 4 3D structure of hemagglutinin with highlighted epitope predictions using chain A from the pdb entry 2IBX. White cartoon: Other chains in multimer not used for predictions. Green cartoon: Part of chain where no class II or B cell epitopes are predicted. Yellow sticks: Predicted helper epitopes (NetMHCII predictions) considering the DRB*0101 allele. Blue spheres: Predicted B cell epitopes (DiscoTope). Orange spheres: Residues predicted to be in both B cell and helper epitopes. The tools FeatureMap3D (www.cbs.dtu.dk/services/FeatureMap3D/) and PyMol (pymol.sourceforge.net/) were used to generate the drawing

To find conserved B cell and helper epitopes a similar approach could be used even though conserved conformational epitopes might be hard to find and even harder to direct a response to. Figure 2 displays a three-dimensional protein structure model of the variable surface protein hemagglutinin from a H5N1 strain. Predicted B cell epitopes are mapped on the structure, as well as helper epitopes restricted to the relatively common HLA-DRB*0101 allele.

4 Viral evolution in the human host

4.1 Introduction

The previous section has discussed the evolution which a pathogen population undergoes within the human population over a time span of years or longer. Some pathogens but not all, by any means, play a more dynamic evolutionary game inside the host by which they try to evade the host's immune system or the drug therapy that is applied to combat the disease. We observe this kind of process both with unicellular pathogens and

viruses. An example of the former is *Plasmodium falciparum* which causes Malaria. Here the pathogen evolves new suites of surface epitopes repeatedly to evade the adaptive immune response of the host, and the immune system of the host responds to the new populations of modified pathogens with recurrent fever bouts that manifest the periodic amplifications of the immune system activity.

This section will present a viral example, namely the case of Human Immunodeficiency Virus (HIV), which causes AIDS.

4.2 Replication cycle of HIV

HIV is a single-stranded RNA virus with two copies of the genome per virus particle. The replication cycle of the virus is schematically illustrated in Fig. 5.

The virus enters the human cell by attaching with its surface protein gp120 to the cellular receptor CD4. It needs one of the two cellular coreceptors CCR5 or CXCR4 to facilitate cell entry. After fusion with the cell membrane it releases its content and uses one of the viral enzymes, namely the *Reverse Transcriptase* (RT) to transcribe its genome back to DNA. Another viral enzyme, the *Integrase* (IN) splices the DNA version of the viral genome, the so-called *provirus*, into the genome of the infected host cell. This cell is often a T-helper cell of the host's immune system. Once this cell starts dividing, i.e., as part of the immune response to the HIV infection, the cell starts producing the building blocks of the virus. New virus particles assemble at the cell surface and segregate. During a final virus maturation phase, a third viral enzyme, the *Protease* (PR) cleaves the viral polyproteins into their active constituents. The dynamic evolution of HIV is manifested by the fact that RT lacks a proof-reading mechanism and introduces genomic variants during the copying process. The high turnover of over a billion virus particles per host and day during periods of high-activity immune response affords a sufficient genomic diversity for a selective evolutionary process that lends an advantage to forms of the virus that are resistant to the immune system and drug therapy with which they are confronted.

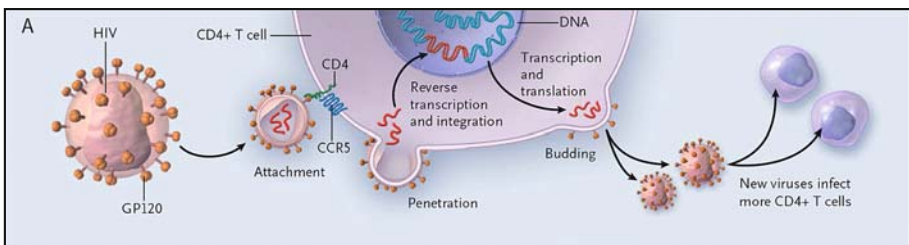


Fig. 5 Replication cycle of HIV (from (Markel 2005))

4.3 Targets for antiviral drug therapy

Antiviral drugs target one of several of the proteins involved in the viral replication cycle. The historically earliest drugs target RT and block it by providing “fake” nucleotides for the DNA assembly that act as terminators for the chain elongation process. These drugs are called *nucleoside analog RT inhibitors* (NRTIs). Another class of drugs targeting RT, the *non-nucleoside analogs* (NNRTIs), facilitate inhibition of the enzyme by binding to a specific part of its binding pocket. Since the mid 90s, inhibitors of PR (PIs) that substitute for the peptides to be cleaved by the enzyme have entered the market place. Inhibitors of integrase are just about to come to market. Finally, in recent years, several drugs have been developed that target the blockage of the process of viral cell entry, by blocking one of the involved proteins, either the viral surface protein gp41, or one of the cellular proteins, CD4, CCR5 or CXCR4. Within the older classes of drugs there are up to about a dozen different compounds in each class. The justification for so many compounds is that there are many different variants of HIV that have different resistance profiles. This is also the reason why, for over ten years, the so-called *highly antiretroviral therapy* (HAART) approach administers several drugs from several drug classes to the patient simultaneously, in order to present a high barrier for the virus on its evolutionary path to resistance. Still, after a time of several weeks up to about a year or two, the virus succeeds in evolving a variant that is resistant against the given therapy regimen. At this point, a new drug combination has to be selected to combat the new viral variant.

4.4 Manual selection of antiretroviral combination drug therapies

Even before the use of computers, doctors have selected drug therapies based on the genome of the viral variant prevalent inside the patient which, in developed countries, is routinely determined from virus in the patient’s blood serum via sequencing methods. The basis for the selection is a set of *mutation tables*. There is one such table for each molecular target. The table lists, for each drug, the observed and acknowledged set of mutations (on the protein level) that have been observed to confer resistance against that drug. The offered tables are updated regularly by international societies such as the International AIDS Society (Johnson et al. 2007).

There are two problems with the mutation tables. (1) They regard different mutations as independent from each other. Any one of the mutations listed in the table is considered to confer resistance on its own. However, in some cases, mutations at different positions have been observed to interact in complex ways. For instance, a mutation can resensitize a virus to a drug to which an earlier mutation has rendered it resistant. (2) Mutations are selected to enter the mutation table by a consensus process among experts that cannot claim to be objective and

reproducible. Problem 1 has been countered by the introduction of rule-based expert systems that can implement complex resistance rules involving several mutations (Schmidt et al. 2002). Problem 2 has been approached by introducing bioinformatics methods for predicting resistance from the viral genotype. Such methods derive statistical models directly from clinical data that comprise experience on viral resistance development. We now survey the methods by which such statistical models are derived and applied.

4.5 Data sets for learning viral resistance

First we need data sets for deriving the statistical models. The availability of data in sufficient volume and quality is a major hurdle for bioinformatical approaches to resistance analysis. Data have been collected in several parts of the world, e.g., in the USA (Stanford HIV Database (Rhee et al. 2003)), over Germany (Arevir Database (Roomp et al. 2006)) and, more recently, over Europe (Euresist Database¹). These databases contain two types of data.

1. *Genotypic data* list viral variants sampled from patients together with clinical information about the patient, including their viral load (the amount of free virus) and counts of immune cells in the blood serum. This allows for correlating the viral genotype with the virologic and immunological status of the patient.
2. *Phenotypic data* report results from laboratory experiments, in which virus containing the resistance mutations observed in the patient is subjected to different concentrations of single antiretroviral drugs and the replication fitness of the virus is measured. This results in a quantitative measure of viral resistance, the so-called *resistance factor*. Briefly, a virus with a resistance factor of 10 against some drug requires ten times the concentration of that drug in comparison to the wild-type virus in order to reduce the replication fitness of both viruses to the same extent.

In developed countries, genotypic data are collected routinely in clinical practice. Thus they are available in high volume (tens of thousands of data points). The viral genotypes are usually restricted to the genes of the target molecules (here RT and PR). Phenotypic data require high-effort laboratory procedures and cannot be collected routinely. Thus they are available in lower quantities (thousands of data points). While phenotypic data represent viral resistance in an artificial environment, they provide a highly informative quantitative value for resistance. Thus can are of substantial value for learning statistical models with high predictive power.

¹ <http://www.euresist.org>

4.6 Computational procedures for predicting resistance

We will survey approaches to solving three problems in resistance prediction:

1. *Quantifying the information that a mutation carries with respect to the resistance of any viral variant with that mutation against a given drug.* Any method solving this problem can be used to generate mutation tables such as the one derived by hand through expert panels.
2. *Predicting the resistance of a given genotypic variant against a given drug.* Any method solving this problem can also take complex interactions between different mutations into account and thus competes with the rule-based expert systems mentioned before.
3. *Assessing the effectiveness of a combination drug regimen against a given genotypic variant.* Methods solving this problem can take the future viral evolution into account. Thus, in effect, they can attempt to answer the question how effective the virus will be in evading the present combination drug therapy. Thus they go further methodically than any competing method.

We will now summarize the methods that are used to solve the above problems.

Several methods are available for solving Problem 1. Computing the mutual information content of a viral mutation with respect to the wild type is one alternative (Beerenwinkel et al. 2001). Another is to generate a support-vector machine model for predicting resistance against the drug and deriving the desired information from it (Sing et al. 2005). The resulting methods yield suggestions for new resistance mutations that are highly desired by the medical community.

Problem 2 can be solved with classical supervised learning techniques such as decision trees or support vector machines (Beerenwinkel et al. 2002). These methods provide classification of viral variants into *resistant* or *susceptible*, or regression of the measured resistance factor or the viral load observed in a clinical setting. The models incur error rates of about 10–15% against measured phenotypic data and the resulting web-based prediction servers² are very popular with practicing physicians and laboratories evaluating patient data. Figure 6 shows an excerpt of a respective patient report that presents an intuitive display of the level of the virus against each drug.

The solution of Problem 3 is somewhat more complicated. We need several ingredients for a respective method. First, we need a notion of success and failure, respectively, of a combination drug therapy that incurs more than a moment's observation of the patient. One way is to assess the effectiveness of a therapy after

² E.g. <http://www.geno2pheno.org>

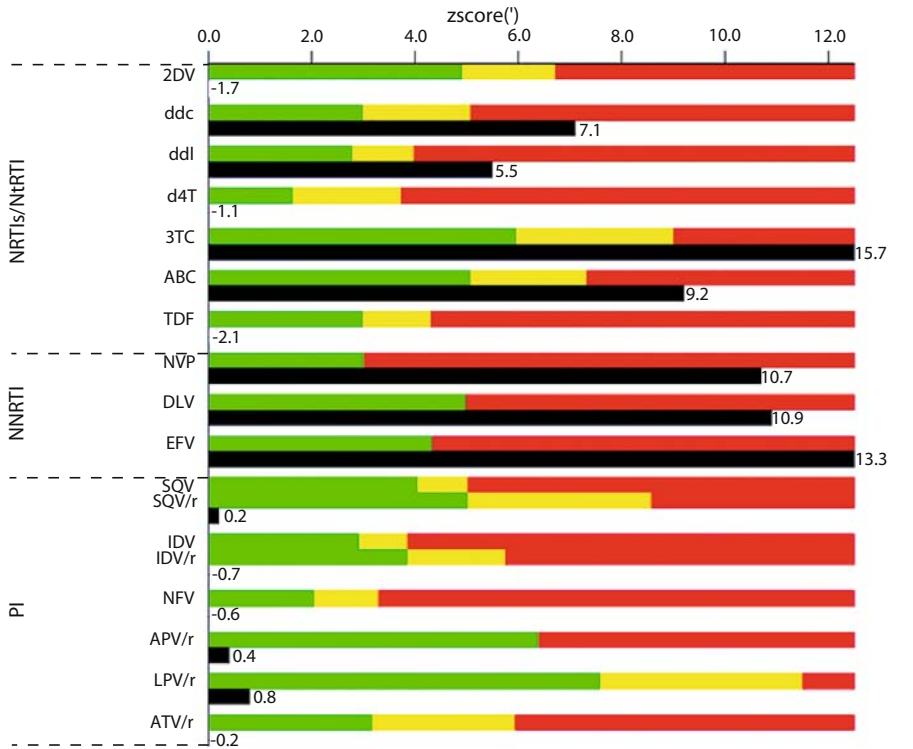


Fig. 6 Patient report by the geno2pheno resistance prediction server. There is one line for each drug. The level of resistance of the virus against the drug is represented by the length of the black bar. The colored bar above indicates the region of resistance (green – susceptible, yellow – intermediate, red- resistant)

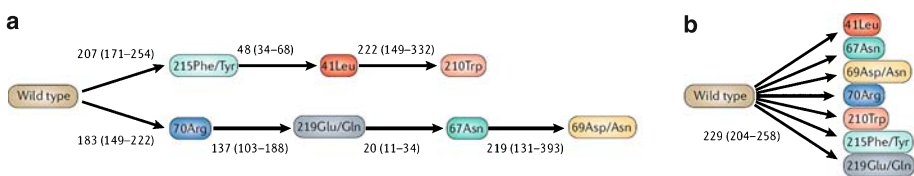


Fig. 7 Tree model of the evolutionary development of viral resistance against the NRTI zidovudine. The model consists of two trees. Tree (a) displays two clinically observed nontrivial paths to resistance, indicated by mutations that accumulate from the wild type from left to right. Tree (b) represents unstructured noise in the data. The method also return quantitative estimates for how much of the data is explained by what tree. In this case the left tree explains about 78% of the data

some time since onset, say eight weeks. The second is a model of viral evolution under drug therapy. We have developed a statistical model that represents the paths of the virus to resistance by a set of trees ((Beerenwinkel et al. 2005b), see Fig. 7)

Given such a tree model, we can derive a quantitative value for the probability of a virus to become resistant against a certain drug after a given amount of time, given a specific combination drug therapy. This value is called *the genetic barrier to drug resistance* (Beerenwinkel et al. 2005a). Finally we use multivariate statistical learning

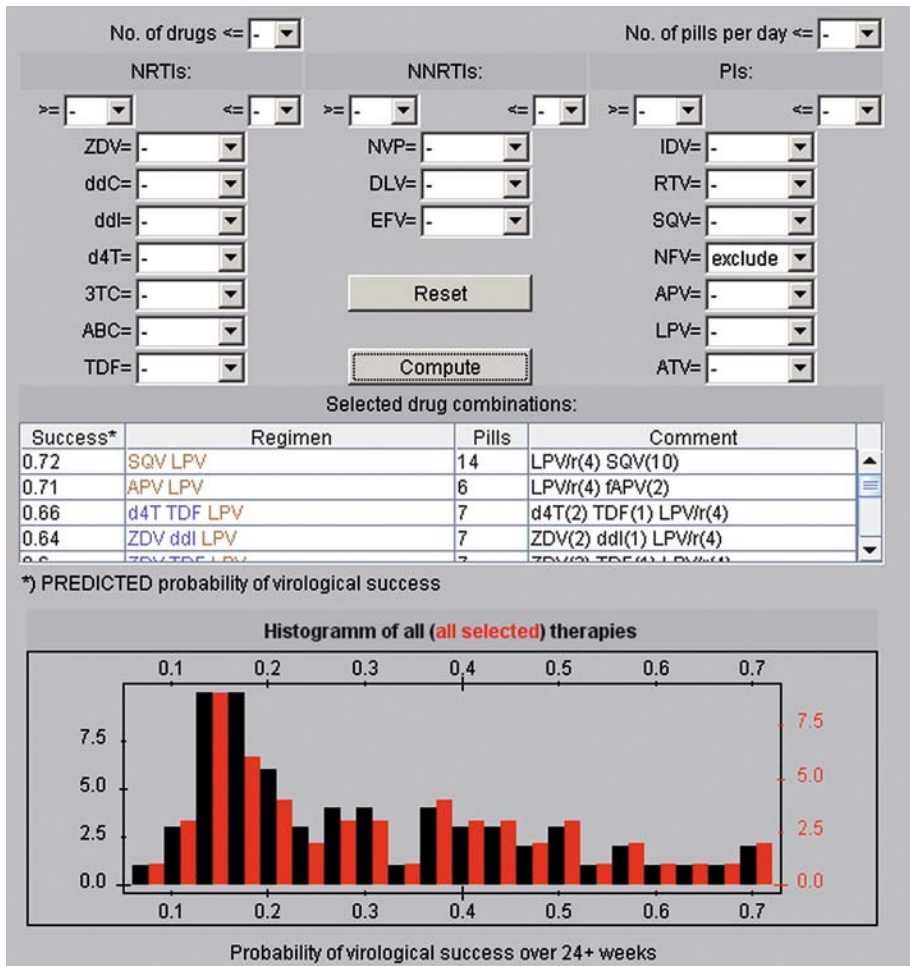


Fig. 8 Results of prediction of therapy effectiveness on the same sample as used for generating Fig. 2. At the top, the user can preselect, here, by excluding the use of the protease inhibitor NFV. In the middle a list of ranked therapies is given. The two top-ranking therapies involve two protease inhibitors, which is not surprising since, by inspection of Fig. 2 the viral variant displays few resistances against protease inhibitors. The distribution of therapy effectiveness with (red) and without (black) preselection is displayed at the bottom

techniques to generate models that classify the therapies into successes and failures based on input comprising the viral variant, the applied therapy, phenotypic resistance prediction (Problem 2), and the predicted genetic barrier to resistance. The results of the implementation of this method called THEO are displayed as illustrated in Fig. 8.

The resulting method reduces the error of therapy classification from about 24% (without any use of software) to under 15% (Altmann et al. 2007). While this is a substantial improvement in accuracy, doctors are still hesitant to use the method in clinical practice, for two reasons: (1) They would like more information on why the method arrives at its results, i.e., they ask for the results to be more interpretable. (2) They question the “objectiveness” of the data. In some sense the subjectivity of the expert decision is replaced with the arbitrariness of how the dataset is collected, from which the model are built.

Addressing both issues is possible but requires additional research which is currently under way.

4.7 Clinical impact of bioinformatical resistance testing

The methods described here are applied within clinical practice in the context of research projects and clinical studies. They improve the rate of selection of adequate drug combination therapies significantly. Besides the statistical evaluation by cross validation, there has been a retrospective study, in which previously applied therapies have been rechecked with the *geno2pheno* software (Problem 2 above) among other prediction systems, and the software has proven to pick successful therapies statistically significantly more often than therapies that turned out not to be successful. Among the single cases that can be reported is a patient who had been receiving HAART for 16 years within several therapy changes but without ever having virus cleared from his blood serum. After the mutation tables offered no more option for therapy, the bioinformatics software made a suggestion that was amended by the doctor. The resulting therapy was the first to clear the patient’s blood of virus and held for at least 2.5 years. Thus, while the software does not make flawless suggestions it advances the state of therapy selection significantly.

Bioinformatics solutions to Problem 3 have yet to win acceptance with the practicing physicians.

The methods described here can be transferred to other diseases for which viral evolution to resistance inside the patient can be observed and for which the relevant genotypic and phenotypic data are available. Transferring the methods to Hepatitis B and C is in preparation.

A recent review on bioinformatical resistance testing is provided in (Lengauer and Sing 2006).

4.8 Bioinformatical support for applying coreceptor inhibitors

As the new coreceptor inhibitors are entering the marketplace and affording a completely new approach to AIDS therapy, there are also new problems that have to be dealt with and that can be supported with bioinformatics methods. We mentioned above that CCR5 and CXCR4 are the two coreceptors that are used alternatively by HIV to enter the infected cell. The clinical picture manifests that, almost exclusively, CCR5 is required for primary infection (R5 virus). As the disease progresses, the virus often switches to using CXCR4 (X4 virus). Some viral variants can use both coreceptors (R5X4 virus). The use of CXCR4 is often associated with enhanced disease symptoms and accelerated disease progression. Thus, preventing the virus from evolving to an X4 variant is a therapy goal. CCR5 seems to be inessential, as humans with an ineffective CCR5 gene shows no disease phenotype, but are highly resistant to developing AIDS. Thus CCR5 is an attractive target for inhibiting drugs. The first CCR5 blocker Maraviroc (Pfizer) has just entered the marketplace. Regulatory agencies, as they were admitting the drug for clinical use, prescribed accompanying tests of the virus for coreceptor usage, as it is ineffective to treat X4 viruses with CCR5 blockers.

For testing of coreceptor usage we have a similar picture as for resistance testing. Coreceptor usage is determined based on the viral genotype. There are laboratory assays for measuring coreceptor usage. They are a little bit closer to clinical routine than phenotypic resistance tests, but they still suffer from limited accessibility, long times (weeks) to receive the results and high cost.

Using genotypic and phenotypic data, one can develop statistical models for viral coreceptor usage based on the viral genotype. Supervised learning models such as support vector machines or position-specific scoring matrices are used for this purpose. The methods are based mainly on the viral genotype (this time restricted to the hypervariable V3 loop of the viral gp120 gene that binds to the coreceptor). Prediction accuracy can be enhanced by including clinical parameters, such as patient immune status, in the model or by specifically offering 3D-structural information on the V3 loop in the form of a structural descriptor that is based on mapping the viral variant under investigation onto the x-ray model of a reference V3 loop. Reviews on bioinformatical prediction of coreceptor usage can be found in (Jensen and van 't Wout 2003; Lengauer et al. 2007).

5 Perspectives

In the last decade, computational biology has embarked on the analysis of host-pathogen interactions. However, the field is still in an early stage. The analysis of viral evolution inside the human population is currently targeting genetic drift but does not yet have a handle on analyzing and predicting genetic shift. The analysis of interactions between viral epitopes and molecules of the human immune system has brought forth effective methods for analyzing and predicting the strength of MHC-binding but has yet

to develop models that adequately represent the many stages of molecular interactions and molecular transport that lead to eliciting an immune response. And the support of the selection of new antiviral therapies in the face of emerging resistant strains inside a patient is still mainly based on statistical analysis of previously applied therapies (to many different patients) rather than on a mechanistic understanding of the molecular interaction networks manifesting the disease. In all fields we would greatly benefit from dynamic simulatable models of the molecular processes manifesting the disease and of the way in which molecular determinants of the virus, the immune system of the host and the applied drugs influence them. Basic research in the field of computational modeling of virus-host interactions will be directed towards generating this network-based understanding of the involved processes. Towards this end we need not only develop new computational models but also generate the relevant experimental data for calibrating the models and for identifying the molecular determinants involved.

References

- The World Health Organization Global Influenza Program Surveillance Network (2005) Evolution of H5N1 avian influenza viruses in Asia. *Emerg Infect Dis* 11: 1515–1521
- Altmann A, Beerenwinkel N, Sing T, Savenkov I, Däumer M, Kaiser R, Rhee SY, Fessel WJ, Shafer RW, Lengauer T (2007) Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antivir Ther* 12: 169–178
- Beerenwinkel N, Lengauer T, Selbig J, Schmidt B, Walter H, Korn K, Kaiser R, Hoffmann D (2001) Geno2pheno: interpreting genotypic HIV drug resistance tests. *IEEE Intel Syst* 16: 35–41
- Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, Korn K, Selbig J (2002) Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proc Natl Acad Sci USA* 99: 8271–8276
- Beerenwinkel N, Däumer M, Sing T, Rahnenführer J, Lengauer T, Selbig J, Hoffmann D, Kaiser R (2005a) Estimating HIV Evolutionary Pathways and the Genetic Barrier to Drug Resistance. *J Infect Dis* 191: 1953–1960
- Beerenwinkel N, Rahnenführer J, Däumer M, Hoffmann D, Kaiser R, Selbig J, Lengauer T (2005b) Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol* 12: 584–598
- Brusic V, Petrovsky N, Zhang G, Bajic VB (2002) Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol Cell Biol* 80: 280–285
- Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. *Science* 286: 1921–1925
- Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci USA* 94: 7712–7718
- Haste Andersen P, Nielsen M, Lund O (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci* 15: 2558–2567
- Heckerman D, Kadie C, Listgarten J (2007) Leveraging information across HLA alleles/supertypes improves epitope prediction. *J Comput Biol* 14: 736–746
- Janeway C (2005) *Immunobiology: the immune system in health and disease*. Garland Science, New York
- Jensen MA, van 't Wout AB (2003) Predicting HIV-1 coreceptor usage with sequence analysis. *AIDS Rev* 5: 104–112
- Johnson VA, Brun-Vezinet F, Clotet B, Gunthard HF, Kuritzkes DR, Pillay D, Schapiro JM, Richman DD (2007) Update of the Drug Resistance Mutations in HIV-1: 2007. *Top HIV Med* 15: 119–125

- Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M (2007) Large-Scale Validation of Methods for Cytotoxic T-Lymphocyte Epitope Prediction. *BMC Bioinformatics* 8: 424
- Lengauer T, Sander O, Sierra S, Thielen A, Kaiser R (2007) Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol* 25: 1407–1410
- Lengauer T, Sing T (2006) Bioinformatics-assisted anti-HIV therapy. *Nat Rev Microbiol* 4: 790–797
- Lundegaard C, Lund O, Kesmir C, Brunak S, Nielsen M (2007) Modeling the adaptive immune system: predictions and simulations. *Bioinformatics* 23: 3265–3275
- Markel H (2005) The search for effective HIV vaccines. *N Engl J Med* 353: 753–757
- Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Roder G, Peters B, Sette A, Lund O, Buus S (2007a) Quantitative, pan-specific predictions of peptide binding to HLA-A and -B locus molecules. *PLoS-One* 2: e796
- Nielsen M, Lundegaard C, Lund O (2007b) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* 8: 238
- Peters B, Bui HH, Frankild S, Nielson M, Lundegaard C, Kostem E, Basch D, Lamberth K, Harndahl M, Fleri W, Wilson SS, Sidney J, Lund O, Buus S, Sette A (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol* 2: e65
- Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50: 213–219
- Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res* 31: 298–303
- Roomp K, Beerenwinkel N, Sing T, Schülter E, Büch J, Sierra-Aragon S, Däumer M, Hoffmann D, Kaiser R, Lengauer T, Selbig J (2006) Arevir: A secure platform for designing personalized antiretroviral therapies against HIV. In: Leser U, Naumann F, Eckman B (eds) *Third International Workshop on Data Integration in the Life Sciences (DILS 2006)*. Springer Verlag, Hinxton, U.K. 4075, pp 185–194
- Schmidt B, Walter H, Zeitler N, Korn K (2002) Genotypic drug resistance interpretation systems – the cutting edge of antiretroviral therapy. *AIDS Rev* 4: 148–156
- Sing T, Svicher V, Beerenwinkel N, Ceccherini-Silberstein F, Däumer M, Kaiser R, Walter H, Korn K, Hoffmann D, Oette M, Rockstroh J, Fätkenheuer G, Perno C-F, Lengauer T (2005) Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling and SVM-Based feature ranking. In: Alipio MJ, Torgo L, Bradzil PB, Camacho R, Gama J (eds) *Knowledge discovery in databases: PKDD 2005*. Lecture notes in computer science No. 3721, Springer Verlag, Berlin/Heidelberg, pp 285–296
- Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA (2004) Mapping the antigenic and genetic evolution of influenza virus. *Science* 305: 371–376
- Sundar K, Boesen A, Coico R (2007) Computational prediction and identification of HLA-A2.1-specific Ebola virus CTL epitopes. *Virology* 360: 257–263
- Sylvester-Hvid C, Nielsen M, Lamberth K, Roder G, Justesen S, Lundegaard C, Worning P, Thomadsen H, Lund O, Brunak S, Buus S (2004) SARS CTL vaccine candidates; HLA supertype-, genome-wide scanning and biochemical validation. *Tissue Antigens* 63: 395–400
- Taubenberger JK, Morens DM, Fauci AS (2007) The next influenza pandemic: can it be predicted? *JAMA* 297: 2025–2027
- Wallace RG, Hodac H, Lathrop RH, Fitch WM (2007) A statistical phylogeography of influenza A H5N1. *Proc Natl Acad Sci USA* 104: 4473–4478
- Wang M, Lamberth K, Harndahl M, Roder G, Stryhn A, Larsen MV, Nielsen M, Lundegaard C, Tang ST, Dziegiel MH, Rosenkvist J, Pedersen AE, Buus S, Claesson MH, Lund O (2007) CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. *Vaccine* 25: 2823–2831
- Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y (1992) Evolution and ecology of influenza A viruses. *Microbiol Rev* 56: 152–179

CHAPTER 8.2

Alternative splicing in the ENCODE protein complement

M. L. Tress¹, R. Casadio², A. Giorgetti⁵, P. F. Hallin⁶, A. S. Juncker⁶,
E. Kulberkyte⁶, P. Martelli², D. Raimondo⁴, G. A. Reeves³, J. M. Thornton³,
A. Tramontano⁴, K. Wang⁶, J.-J. Wesselink¹ and A. Valencia¹

¹Computational and Structural Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

²Bologna Biocomputing Unit, CIRB/Department of Biology, University of Bologna, Bologna, Italy

³European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

⁴Department of Biochemical Science, University of Rome “La Sapienza”, Rome, Italy

⁵Faculty of Mathematical, Physical and Natural Sciences, University of Verona, Verona, Italy

⁶Center for Biological Sequence Analysis BioCentrum-DTU, Technical University of Denmark, Lyngby, Denmark

1 Introduction

In eukaryotic cells transcribed precursor messenger RNA (mRNA) contains introns and exons. Precursor mRNA is converted into mature mRNA by removal of introns and the splicing together of the remaining exons by the spliceosome. Alternative splicing is the process whereby the splicing process can generate a diverse range of mature RNA transcripts from different combinations of exons and allows the cell to generate a series of distinct protein isoforms. Alternative splicing events that occur within exons that are protein coding are likely to alter the structure and biological function of the expressed protein isoform and may even create new protein functions. This has led to suggestions that alternative splicing has the potential to expand the cellular protein repertoire (Lopez 1998; Black 2000; Modrek and Lee 2002).

Recent studies have estimated that at least 60% of multi-exon human genes can produce differently spliced mRNAs (Harrow et al. 2006; Scherer et al. 2006) and that alternative splicing has the potential to more than double the number of different proteins in the cell.

Alternative splicing has long been linked to processes such as development (Wojtowicz et al. 2004) and has been implicated in a number of cellular pathways

Corresponding author: Michael L. Tress and Alfonso Valencia, Computational and Structural Biology Group, Spanish National Cancer Research Centre (CNIO): Madrid, Spain (e-mails: mtress@cnio.es, valencia@cnio.es)

(Ermak et al. 1995; Matsushita et al. 2006; Wells et al. 2006). It has been suggested that the purpose of alternative splicing is to expand the range of functions in the cell (Graveley 2001; Hui and Bindereif 2005) and the extent of alternative splicing in eukaryotic genomes has even lead to suggestions that alternative splicing is key to explaining the discrepancy between the number of human genes and functional complexity (Pennisi 2005). In order to generate proteins with new functions at different stages of development and in different tissues some external regulation is required and it has been suggested that the splicing process is controlled by a sophisticated regulation system (Smith and Valcarcel 2000; Florea 2006).

A highly curated reference set of human spliced variants has been annotated as part of the pilot project of the Encyclopedia of DNA Elements (ENCODE) project (The ENCODE Project Consortium 2004, 2007). The ENCODE project seeks to identify all the functional elements in the human genome sequence; the pilot project concentrated on 44 selected regions that added up to 1% of the human genome.

The reference set of splice variants was annotated by the HAVANA team as part of GENCODE consortium (Harrow et al. 2006) and served as the starting point for the detailed study of alternatively spliced gene products that was carried out by members of the BioSapiens Network of Excellence. The study was carried out on the October 2005 freeze of the HAVANA/GENCODE set that contained a total of 434 protein-coding genes. There were a total of 1097 annotated splice variants for these 434 genes, with on average 2.53 protein coding variants per locus; 182 loci were annotated with just one variant while one locus, RP1-309K20.2 (CPNE1) had 17 coding variants.

A total of 57.8% of the loci were annotated with alternatively spliced transcripts, although this is probably an underestimate of the true number of splice variants – several of the ENCODE regions were selected for biological interest (The ENCODE Project Consortium 2004) and alternative splicing was less frequent in these regions, in part because of gene clusters such as the cluster of 31 olfactory receptor genes from chromosome 11 (Taylor et al. 2006). These olfactory receptors are recent in evolutionary origin, have a single large coding exon and code for a single isoform. In the 0.5% of the human genome that was selected by the stratified random-sampling process (The ENCODE Project Consortium 2004), 68.7% of the loci had multiple variants. This number is towards the higher end of previous estimates, but in line with the most recent reports (Nusbaum et al. 2005).

While a full understanding of the functional implications of alternative splicing is still a long way off, the HAVANA set provided us with the material to make the first assessment of a systematically collected reference set of splice variants. The BioSapiens study looked at the distinct protein isoforms generated by alternative splicing and attempted to predict differences between splice isoforms at the level of protein structure, location and function. As a result of the study we were able to highlight a number of interesting splice isoforms and we have begun experimental work to investigate the expression as proteins of many of these splice variants.

In Sects. 2–5 we look in detail at the results of the BioSapiens study. As a result of this pilot study we are in the process of developing a pipeline for the automatic annotation of the structure, function and location of splice variants at a genomic level. The pipeline has been developed in the context of the ENCODE scale up and will work in collaboration with the HAVANA annotation team in the Sanger Centre. Aspects of the pipeline are described in Sect. 6 and the pipeline itself is described in Sect. 7.

2 Prediction of variant location

Machine learning tools have had to be adapted in order to annotate of the structure and function of the 1097 transcripts selected by the GENCODE consortium (Harrow et al. 2006), in particular when investigating differences between variants of the same gene. Not all transcripts can be annotated by searches for homologous proteins. In fact in many cases differences are masked by the similarity between splice isoforms so methods that annotate by homology can be easily fooled.

As a first characterisation of the transcripts we discriminated between membrane-bound and globular proteins. Various state-of-art tools were used for predicting the presence of trans-membrane domains: ENSEMBLE (Martelli et al. 2003), PRODIV (Viklund and Elofsson 2004) and PHOBIUS (Kall et al. 2004). The first two methods need to be coupled with a predictor for the presence of N-terminal signal peptides (SignalP, Bendtsen et al. 2004) because signal peptides are easily confused with transmembrane alpha helices owing to their hydrophobic composition. PHOBIUS integrates the prediction of signal peptides and membrane-bound domains in a single hidden Markov Model.

Three prediction methods means there will be three different sets of predictions since the tools are based on different methodologies and were trained on distinct datasets. For this reason a CONSENSUS prediction was also computed. The CONSENSUS method required at least two out of the three methods to agree. PRODIV, ENSEMBLE and PHOBIUS predicted transmembrane helices (TMH) for 271 (24.7%), 331 (30.2%) and 324 (29.5%) out of the 1097 proteins. A CONSENSUS prediction could be made for 1026 variants; 225 of them were predicted as membrane proteins. A signal peptide was predicted for 180 sequences with PHOBIUS and for 229 sequences with both PRODIV and ENSEMBLE (using SignalP). The CONSENSUS method predicted a signal peptide in 204 transcripts.

Figure 1 shows the distribution of the proteins with respect to the number of predicted TMH for all methods. Evidently, the two most abundant classes are the proteins predicted with one and seven transmembrane segments. Almost all the proteins endowed with seven transmembrane helices derive from the single olfactory receptor cluster in chromosome 11. Nevertheless, based on our prediction we were able to annotate a transcript from RHBDF1, previously annotated as hypothetical protein, as a new 7-helix transmembrane receptor.

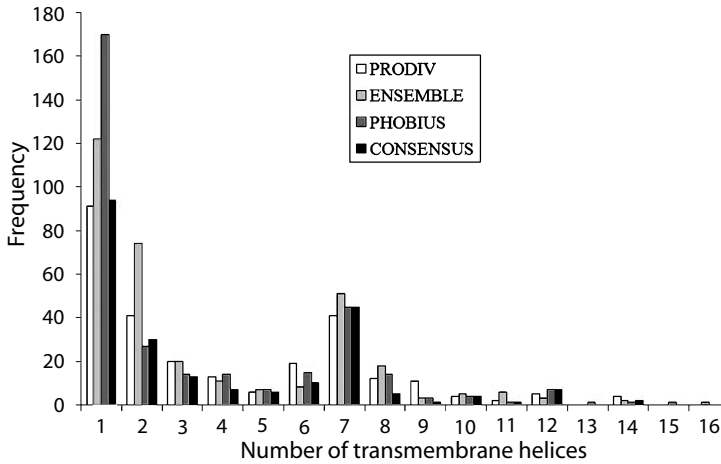


Fig. 1 The predictions made by three different transmembrane helix prediction methods for the 1097 protein sequences in the GENCODE annotated set. The methods used are PRODIV, ENSEMBLE and PHOBIUS. The CONSENSUS prediction requires 2 of the 3 methods to agree. This figure is copyright of the National Academy of Sciences, USA

Table 1 Predictions for transmembrane helices made for the 548 complete sequences from loci with multiple variants. This table is copyright of the National Academy of Sciences, USA

Genes encode:	SP-PRO ¹	SP-ENS ²	PHOBIUS	CONSENSUS
Globular variants only	184	163	166	185
Variants with identical TM structure	26	26	34	28
Both globular and TM variants	19	37	26	16
Variants with varying no. of TMH	17	22	18	13
TM variants with inverted orientation	6	4	8	2

¹ SP-PRO: SignalP coupled with PRODIV; ² SP-ENS: SignalP coupled with ENSEMBLE

In order to understand the functional implications of alternative splicing we compared predictions for the variants within the same locus (Table 1). When the CONSENSUS prediction is considered, 185 loci encode only for globular isoforms, 28 loci encode for transmembrane variants with the same structure (signal peptide, number of TMH and orientation with respect to the membrane plane), and 31 loci encode for variants with different transmembrane structure.

It is notable that 16 of the genes have both globular and transmembrane variants. Some of these are detailed below:

- UGT1A10 encodes for 2 variants of UDP-glucuronosyltransferase 1A10. One lacks an 89-residue long C-terminal segment that contains a predicted transmembrane helix. If expressed, this would be the first soluble UDP-glucuronosyltransferase naturally encoded. However, an engineered water-soluble form is reported (Kurkela et al. 2004).
- KIR2DL4 encodes six different variants of a killer cell immunoglobulin-like receptor. All but one share the same 217 residue long N-terminal domain but they all differ in the C-terminal domain. A transmembrane alpha helix is predicted in this C-terminal portion for just three variants. Genetic polymorphism of KIR2DL4 results in alleles with either 9 or 10 consecutive adenines in exon 6, which encodes the transmembrane domain. The loss of a single adenine leads to a frame shift and the expressed protein is truncated at the C-terminal and soluble (Goodridge et al. 2007).
- IFNAR2 encodes five isoforms of the beta chain of the interferon alpha/beta receptor. Four isoforms share a segment predicted as transmembrane that is missing in the two remaining variants. Evidence for both soluble and transmembrane forms of the interferon receptor have previously been reported (Novick et al. 1995).

Thirteen other loci code for variants that are predicted to contain differing numbers of transmembrane regions. Some examples of loci encoding for variation in numbers of TMH are detailed below:

- AVPR2 encodes two variants of the type two vasopressin receptor. One isoform is predicted to span the membrane with 7 helices, as with other well-known G protein coupled receptors. The other variant lacks a 67-residue long C-terminal segment and is endowed with only 6 helices. Recently another case of a six transmembrane splice variant of a G-protein coupled receptor has been reported, although its function, if any, is still unknown (Vielhauer et al. 2004).
- TTYH1 encodes four variants of a protein similar to TWEETY, a putative cation transporter in *Drosophila*. Three of the variants are predicted to contain five transmembrane segments. The remaining variant lacks a 210-residue long N-terminal domain that contains the two first helices and its loss is predicted to convert the third transmembrane helix into a signal peptide. In short, this variant is predicted as a monotopic transmembrane protein and, if that is confirmed, the function of this variant would be dramatically altered.

The localisation of globular proteins can be predicted using specific tools such as BaCelLo (Pierleoni et al. 2006), which predicts the localisation to one of four compartments in the eukaryotic cell (nucleus, cytoplasm, mitochondria and secretory pathway). Out of 830 transcripts for which the CONSENSUS method predicted no transmembrane helix, 313 were predicted as nuclear, 281 as cytoplasmic, 95 as mitochondrial and 141 as secreted.

Also in this case, the comparison between the variants coded by the same locus reveals differences that are likely to be important from a functional point of view in 12 cases. Among them one locus (LILRA3) codes for two secreted and one cytoplasmic variant. A single variant is predicted as cytoplasmic because it has a 17-residue N-terminal insertion ahead of a predicted signal peptide. If the cytoplasmic variant is expressed, there is some evidence that it might be related to leukaemia.

3 Prediction of variant function – analysis of the role of alternative splicing in changing function by modulation of functional residues

3.1 Functions associated with alternative splicing

Alternative splicing events have been associated with a variety of functional variations, the biomedical literature includes examples in growth factor receptor ligand binding, cell adhesion, cytoskeletal differences, cell growth, cell death, differentiation and development, neuronal connectivity, cell excitation and cell contraction. Splicing events have also been shown to effect sub-cellular localisation, phosphorylation by protein kinases and the binding of an enzyme by its allosteric effector. They have even been shown to alter the activation of transcription factor domains. However, functional explanation for alternative splicing is not well understood.

3.2 Functional adaptation through alternative splicing

The most documented case of alternative splicing and function is that of the *Drosophila* Dscam gene. Dscam codes for an axon guidance receptor that is involved in the formation of complex branching patterns of synaptic connections crucial for the formation of distinct neural circuits and that is responsible for directing growth cones to their proper target in Bolwig's nerve of the fly (Schmucker et al. 2000). The protein comprises 10 immunoglobulin-like domains, 6 fibronectin type III domains, one transmembrane and one cytoplasmic domain. Dscam has four sets of interchangeable alternative exons that may be spliced in a mutually exclusive manner. Alternative splicing allows permutations of 3 of the immunoglobulin-like domains and the trans-membrane domain. Exons 4, 6, 9, and 17 are each encoded by an array of potential alternative exons (the four exons have 12, 48, 33 and 2 alternative exons respectively) that would theoretically allow the gene to express 38,016 different proteins. It has been shown that the expression of different Dscam isoforms has an effect on neuron-target recognition of mechanosensory neurons that connect *via* multiple axonal branches (Chen et al. 2006). Two of the different isoforms have recently been crystallised showing that the 2nd and 3rd immunoglobulin domains undergo homo-dimerisation (Meijers et al. 2007). The mutually exclusive splicing of

whole immunoglobulin-like domains by Dscam has an interesting corollary in the insertion/deletion of whole immunoglobulin domains seen in the immunoglobulin-based receptor clusters in the HAVANA set.

Other examples, such as the potential functional mediation of the potassium ion channels in the inner of lower vertebrates seem less clear. Cells within the inner ear of chicks and turtles were shown to have differentially expressed transcripts in different cells within the inner ear (Navaratnam et al. 1997; Rosenblatt et al. 1997). This led to suggestions that these isoforms might be expressed in a gradient along the 10,000 sensory-receptor cells present in the inner ear and that this might enable the perception of different sound frequencies (Ramanathan et al. 1999). However, no such gradient was found in a study of the rat inner ear (Langer et al. 2003), suggesting that if alternative splicing is crucial in lower vertebrate sound perception, higher vertebrates must have developed a different system.

There are a great many papers relating splice variants to functional roles, which gives the impression that alternative splicing has a direct functional relevance in the cell. However, there are also a great many examples of splice variants that appear not to have an obvious function and the role of these variants is not yet understood. The reference set of splice variants from the ENCODE regions (The ENCODE Project Consortium 2004) manually annotated by the HAVANA team at the Sanger Institute (Harrow et al. 2006) provides an excellent dataset from which we can draw general conclusions about the proportion of splicing events that appear to have functional roles. Two particularly interesting examples from the reference set and are detailed below:

3.2.1 Tafazzin

Six splice variants were present in the HAVANA dataset. Although there are 10 isoforms reported in SwissProt (Bairoch et al. 2004), only a subset of these was expressed in a range of mouse and human tissues (Lu et al. 2004). The highest levels of the protein are found in cardiac and skeletal muscle cells, which have large numbers of mitochondria. The precise biochemical function of tafazzin in humans remains unknown, but it shows a distant sequence homology to the glycerolipid acyltransferase family of enzymes and is thought to function as an acyltransferase in the remodelling of cardiolipin in the inner mitochondrial membrane (Vaz et al. 2003). In *Drosophila* tafazzin has been shown to function as a CoA-independent, acylspecific phospholipid transacylase, converting cardiolipin to phosphatidylcholine and *vice versa* (Xu et al. 2006).

In order to assess the functional importance of the alternative splicing, we located the structural positions of the variable exons. No structure of tafazzin exists therefore models were obtained from the ModBase server (<http://modbase.compbio.ucsf.edu>), in order to map the splice variants onto the structure of the protein. Structures were based on the PDB (Berman et al. 2000) entry 1k30 (plant glycerol-3-phosphate acyltransferase, GPAT), which is also a member of the glycerolipid acyltransferase family. The

catalytic residues for GPAT are annotated in the Catalytic Site Atlas (CSA, Porter et al. 2004) as Asp144 and His139. There was a high level of similarity between the GPAT and tafazzin sequences over these and neighbouring residues. Splice variant 006 is missing the exon that codes for the two catalytic residues and without an intact active site this isoform is unlikely to function as an enzyme. Variations in the other splice variants are mostly located at the entrance of the binding site and might be responsible for conferring specificity for different substrates to the different isoforms (Fig. 2). As splice variants are often expressed differentially in different tissues, these different loops are perhaps required for the recognition of different substrates in these various environments.

It has been shown that only isoform 002 has full cardiolipin metabolism activity and that the variant regarded as the main isoform by SwissProt (isoform 001) can only



Fig. 2 A superposition of the 3D models of the 6 ENCODE splice variants of tafazzin. The models are represented by a backbone trace running along the chain of the polypeptide. Exon 2, which appears to be the crucial functional exon and is present in all but one variant, is coloured yellow in all structures. The parts of the protein encoded by exons unique to one or two of the variants are shown by the thicker backbone trace. The view represents the view into the active site from outside. Thus it can be seen that some of the unique exons are located at the entrance of the binding site and might be responsible for the specificity of the different isoforms for different substrates. The inferred catalytic residues (His69 and Asp74) are indicated by the stick bonds at the centre of the picture

partially metabolise cardiolipin (Bione et al. 1996). Isoform 002 has a 31 residue deletion that is the result of skipping the fifth exon. We were also able to show that the exonic structure of isoform 002 was conserved between human and mouse, but that the exonic structure of the main isoform was not conserved. The fact that there was a difference of opinion between experimentalists and SwissProt annotators over this and other genes lead us to propose a pipeline for the detection of principal isoforms (see Sect. 6, this chapter).

3.2.2 Phosphoribosylglycinamide formyltransferase (GARS-AIRS-GART)

Although there are many examples of short indels (less than 20 residues) both in this dataset and in the literature, indels can potentially also correspond to the removal or insertion of a whole domain or domains. The best example of this in the HAVANA set is the immunoglobulin domains in the clusters immunoglobulin-based receptors on chromosome 19 (Tress et al. 2007).

Another example is phosphoribosylglycinamide formyltransferase, or GARS-AIRS-GART protein, is a trifunctional polypeptide, highly conserved in vertebrates, which is involved in *de novo* purine biosynthesis. It has phosphoribosylglycinamide formyltransferase (E.C.2.1.2.2), phosphoribosylglycinamide synthetase (E.C.6.3.4.13) and phosphoribosylaminoimidazole synthetase (E.C.6.3.3.1) activity, which are required for steps 2, 3 and 5 of the purine biosynthesis pathway (Henikoff et al. 1986). The HAVANA dataset and the literature (Brodsky et al. 1997) show that alternative splicing of this gene results in two variant transcripts encoding two different isoforms. These are differentially expressed during human brain development and temporally over-expressed in the cerebellum of individuals with Down's syndrome (Brodsky et al. 1997). These two splice variants are shown in Fig. 3, a schematic diagram of their Pfam

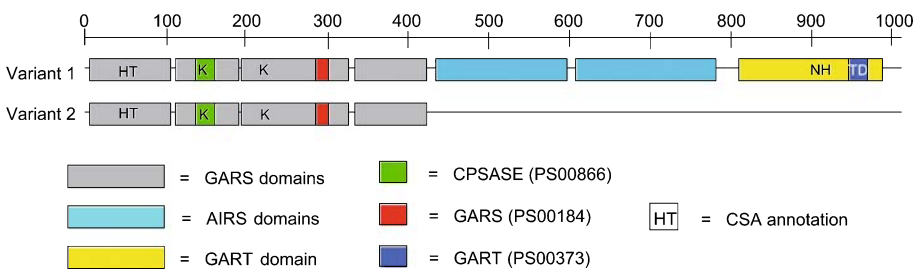


Fig. 3 A schematic diagram showing the Pfam domains, CSA catalytic residues and PROSITE patterns identified in the HAVANA variants of the trifunctional GARS-AIRS-GART protein. The scale indicates the residue positions along the sequence. The three Pfam domains, GARS, AIRS and GART, are shown in grey, blue and yellow. The PROSITE patterns for CPSASE, GARS and GART are indicated in red, green and dark blue. The letters on the bars identify catalytic residues suggested by homology to CSA-annotated proteins

(Finn et al. 2006) domains, CSA catalytic residues and PROSITE (Hulo et al. 2008) patterns identified in the annotated variants. The first variant is complete, with all three functional regions present. The second is missing the AIRS and GART regions. This splice isoform could serve to up-regulate the phosphoribosylglycinamide formyltransferase (GARS) function of this trifunctional peptide by providing more copies.

3.3 Analysis across the ENCODE dataset

Whilst there are examples in which function has been altered by alternative splicing, is it a mechanism that is commonly used for function modulation or a just phenomenon that occurs in a limited number of cases in nature? We carried out a global analysis aimed to provide information on the types of changes occurring to the functionally important residues. How often are exons containing such residues retained and how often are they spliced out or replaced?

We mapped the variants onto homologous structures to allow the transfer of functional information: catalytic residues from the CSA, contact to ligands, metal and DNA from PDBsum (Laskowski 2007), PDB 'site' records from the PDB files and PROSITE patterns from the PROSITE database. Residues annotated with functional information were labelled as functionally important. Each functionally important residue was analysed within all splice variants. Was the residue part of an exon conserved in all variants? How often was the residue left out of splice variants? Were the exons with functional residues substituted for other exons and was the substituted residue the same, similar or completely different? The results can be seen in Fig. 4. It can be seen that the majority of functionally important residues are conserved, suggesting a mechanism in which changes involving the substitution or removal of conserved functionally important residues are more the exception than the rule.

Splicing events occur within Pfam-A hand-curated functional domains in 46.5% of sequence-distinct isoforms. The definition of Pfam-A functional domains is based on expert knowledge and sensitive alignment tools. Therefore if a Pfam-A domain is broken by a splicing event we would expect to see an effect on the function of the domain. Although 46.5% is a surprisingly high figure, it is less than would be expected: if the same number of splicing events were to happen randomly at exon boundaries, splicing events would be expected to fall inside Pfam-A domains in 59.8% of isoforms. Since Pfam-A domains are broken less often than would be anticipated from a purely random process, it seems that there is some (weak) selection against splicing events that affect functional domains.

A similar study carried out by Talavera et al. (manuscript in preparation) on a data set from SwissProt showed that there were no examples in which alternative splicing occurred exclusively in conjunction with domain boundaries. These findings suggest that alternative splicing is rarely a precise functional mechanism for the removal of a particular domain or domains. In addition to this, the data shows that alternative

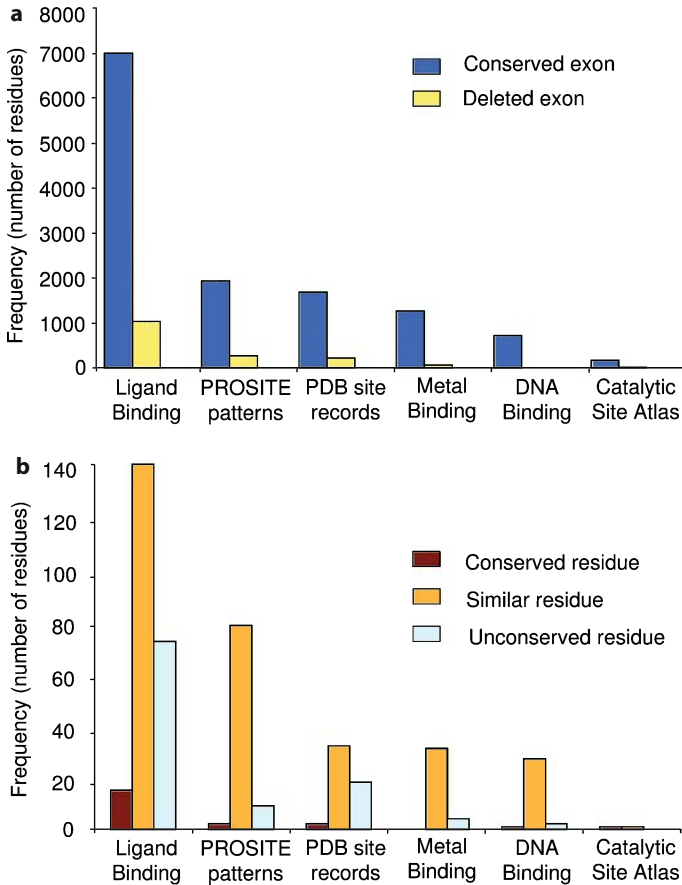


Fig. 4 Histogram showing how often a residue annotated by one of the six functional categories listed at the bottom of the graph is preserved, substituted or lost in the other variants of the same protein. The bars are coloured according to the key on the right and show numbers of residues. This figure is copyright of the National Academy of Sciences, USA

splicing events often leave behind a fragment of a domain. Examples of this can be seen throughout the dataset. Is the purpose of this splicing mechanism to knock out the core of the domain and prevent it from folding (thus removing its function) or is it possible that some of these domains refold?

4 Prediction of variant structure

Before one can ask questions about the effects of alternative splicing on protein structure it must first be possible to match the putative coding sequences to a structural fold or

fold. Identifying template with known structures is a crucial initial step in the protein structure prediction process. In the case of the splice variants from the HAVANA set we were particularly interested in identifying multidomain proteins with known structural domains. One important issue is whether alternative splicing acts at the level of structural domains and whether splice isoforms of multidomain proteins differ from each other by the addition, deletion or substitution of complete domains or sub-domains.

Because of the plasticity of protein structure, the precise assignment of structural domains is somewhat subjective and the assignments from different methods and databases can differ, sometimes even substantially. Our strategy was to combine several methodologies and analyse their results. Without entering into technical details, we used Hidden Markov Models (HMMs, Karplus et al. 1997), statistical models that describe protein family alignments, that were derived from the structural classification databases CATH (Pearl et al. 2005) and scop (Murzin et al. 1995), and from Pfam domain families (Finn et al. 2006), and also the mGenThreader (McGuffin and Jones 2003) fold recognition method. The assignments from mGenThreader are compared with those derived from the CATH-HMM method (Lee et al. 2005) and with the combination of both in Fig. 5.

The conclusions from this partially automatic and partially manual analysis demonstrated that complete domain splicing is not very frequent in the HAVANA dataset (30 out of the 688 sequences for which a match with a CATH-HMM could be found). All types of expected variations (truncations, insertions and substitutions) can be observed within the splice variants in the ENCODE regions, but they are fairly infrequent. If the ENCODE pilot project loci are, as expected, representative of the human genome, only a small percentage of the alternative splicing events affect complete structural domains of multidomain proteins.

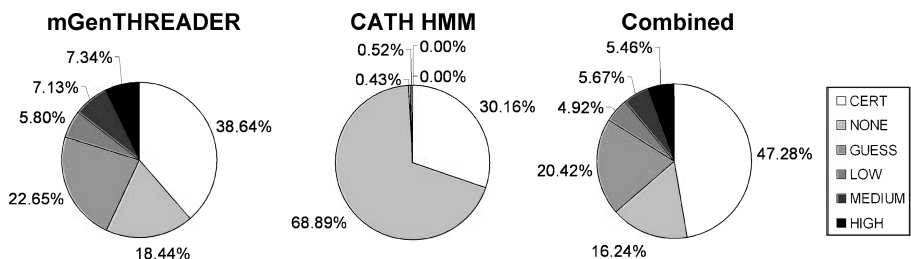


Fig. 5 Residue coverage of ENCODE dataset at five confidence levels. Upper bounds on residue coverage for the mGenTHREADER predictions, CATH HMM domain matches and combined predictions are shown. The confidence levels correspond to expectation (E) values as follows: CERT: $E < 0.0001$, HIGH: $E < 0.001$, MEDIUM: $E < 0.01$, LOW: $E < 0.1$, GUESS: $E > 0.1$. This figure is copyright of the National Academy of Sciences, USA

It follows that most of the alternative splicing events must affect the three-dimensional structure within a domain and we decided to analyse the structure of each of the domains that undergo alternative splicing events. Clearly this analysis had to be limited to domains of known structure and to those for which a reliable comparative model (Tramontano and Morea 2003) can be built.

Perhaps surprisingly, structures had been experimentally solved for a high proportion of the genes in the ENCODE region. As of June 2006 it was possible to find experimentally resolved structures in the Protein Data Bank (Berman et al. 2000) for all or part of 53 genes, 39 of which were annotated as having the potential to generate a total of 98 protein sequence distinct splice variants. The entire structure or practically the entire structure was solved for 21 loci, of which 10 could generate splice variants.

The remaining genes were analysed to assess whether a reliable model could be built using the steps described in the comparative modelling chapter. Namely, we used BLAST (Altschul et al. 1997) to search for proteins sufficiently similar (*E*-value of less than 10^{-5}) to a protein of known structure, discarding cases where the sequence alignment missed more than forty amino acids in a continuous region. This was the case for one hundred isoforms, and twenty of these had an alternatively spliced isoform.

Comparative models of the isoforms for which there was the highest template coverage (that we will call, perhaps improperly, the “main” isoform) were then built using the HHpred server (Soeding et al. 2005). This is an interactive server for protein homology detection and structure prediction based on the pairwise comparison of profile HMMs.

The main isoform models were checked by visual inspection of several features. From this manual inspection we were able to estimate the effect of each alternative splicing event on the structure of the main isoforms. If a splice isoform had a deleted region with respect to the main isoform we tried to assess whether the deletion would affect the packed hydrophobic core of the protein and whether the residues flanking the deletion were so distant in the structure that major rearrangements would be required in the spliced isoform to preserve chain connectivity. If the alternative isoform contained an insertion with respect to the main isoform, we inspected the main isoform model to see whether the location of the insertion was on the surface or the core of the protein and whether or not it would interrupt any of the main secondary structural elements. The effect of substitutions by alternative splicing events was more difficult to assess. If the difference in length between the two alternative exons was sufficiently large, they could be thought of as insertions or deletions. A few examples of alternatively spliced proteins are shown in Fig. 6.

More than 70% (19 out of 26) of the splicing events we looked at in detail would give rise to a “problematic” protein structure, i.e. a protein that could not be thought of as the main isoform plus or minus a peripheral part of the structure. Are these putative alternative splicing products functional? It is very hard to say. It is possible that other proteins could interact with the “problematic” isoform and stabilise its structure, or that the splice variant could assume a structure completely different from that of the main

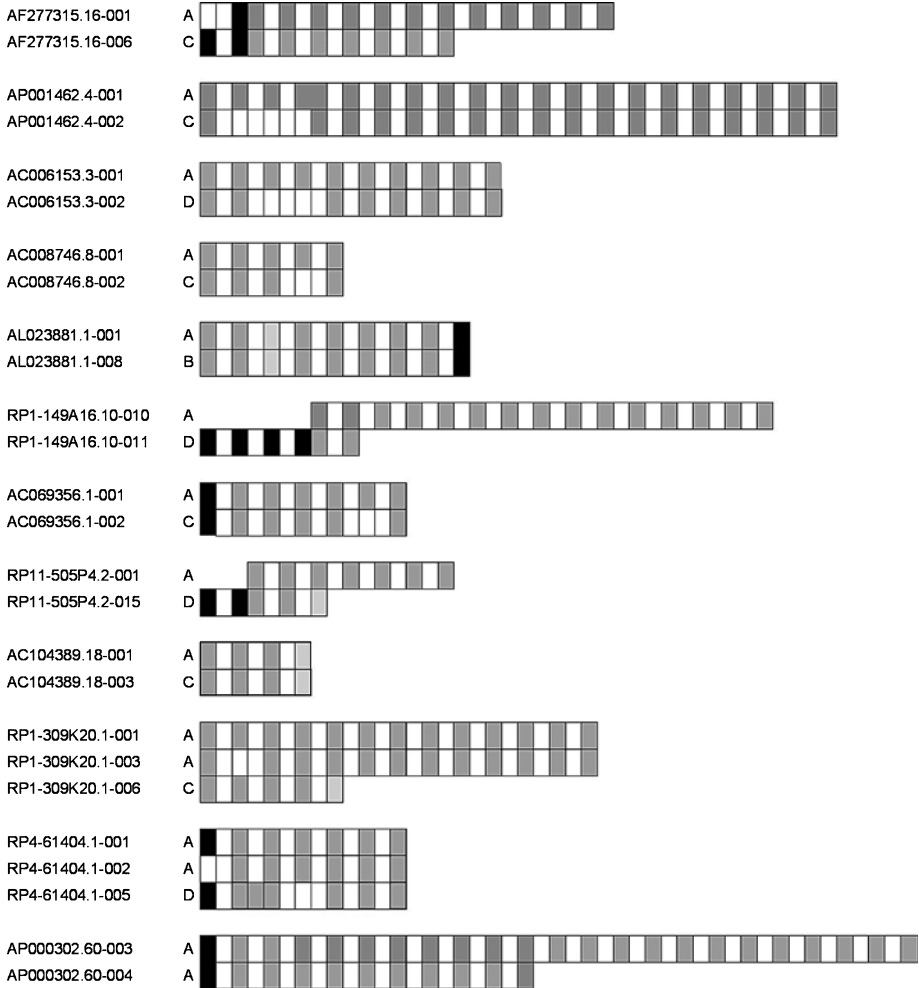


Fig. 6 Results of homology modelling, for clarity, only a few loci are shown. The letter after the ID of the transcript indicates the extent to which a reliable model could be built. We found cases where the different exonic structure of some of the transcripts could not be built by comparative modelling (labelled as D in the figure) as well as cases where some more (C) or less (B) substantial local rearrangement would be required to build the models. Coloured boxes correspond to exons, white boxes to introns (not in scale). Dark gray boxes indicate that the two (or more) transcripts share the same exon; light gray ones are used when the corresponding exons are of different size, but encoded by the same region; untranslated regions are in black. The primary sequences in all of the loci are included in the comparison

one. There are examples in the PDB of proteins with identical sequences that undergo large structural conformational changes on the addition or deletion of residues. For example the tonB transport protein from *Escherichia coli* that undergoes significant

conformational changes when 15 residues are inserted at the N-terminal (Koedding et al. 2005).

These hypotheses are difficult to test not only computationally, but also experimentally. We think, nevertheless, that these phenomena are unlikely to be the explanation in all the cases we have observe. It follows that probably a sizeable fraction of these alternative spliced products are either never translated, translated and degraded immediately after translation because of their inability to form a properly folded stable structure, or translated and somehow tolerated by the cell. Any of these outcomes would be yet another surprise coming from genomic analysis.

5 Summary of effects of alternative splicing

While alternative splicing can produce a range of differently spliced protein isoforms, there is conflicting evidence about their biological relevance. It has been suggested that the purpose of alternative splicing is to expand functional complexity and that the multiple variants are likely to encode functional proteins (Graveley 2001; Hui and Bindereif 2005). In this way several proteins can be encoded in the same DNA sequence, leading to greater efficiency. If alternative splicing can give rise to a range of proteins with functional importance at different stages of development or in different tissues, some external regulation of the pre-mRNA would be needed to decide which protein is produced at which stage and in which location. A sophisticated regulation of the splicing process has been proposed (Smith and Valcarcel 2000; Florea 2006). In many cases these alternative splice variants are hypothesised to function as dominant negative isoforms that can regulate the pathways in which the main functional form is involved (Arinobu et al. 1999; Stojic et al. 2007).

However, evidence for this hypothesis at the protein level is far from clear. The study of the variants in the ENCODE pilot project regions shows that alternative splicing is commonplace and likely to be more frequent than has commonly been suggested (The ENCODE Project Consortium 2004, 2007; Harrow et al. 2006; Tress et al. 2007). The cross-section of alternative splicing events that we have seen at many different loci does point to the possible versatility of alternative splicing in the creation of new functions. However, while alternative splicing has the potential to increase the variety of protein functions, this still has to be demonstrated at the protein level. The BioSapiens project (Tress et al. 2007) was able to show that there is little to indicate this is translated into an increase in the repertoire of protein functions, despite the widespread evidence for the expression of alternative transcripts.

Many of the proteins that result from alternative exon use would almost certainly have substantially rearranged structures with respect to their constitutively spliced counterparts (Talavera et al. 2007; Tress et al. 2007) and these changes are likely to have profound effects on the location and function of the alternative gene products (see

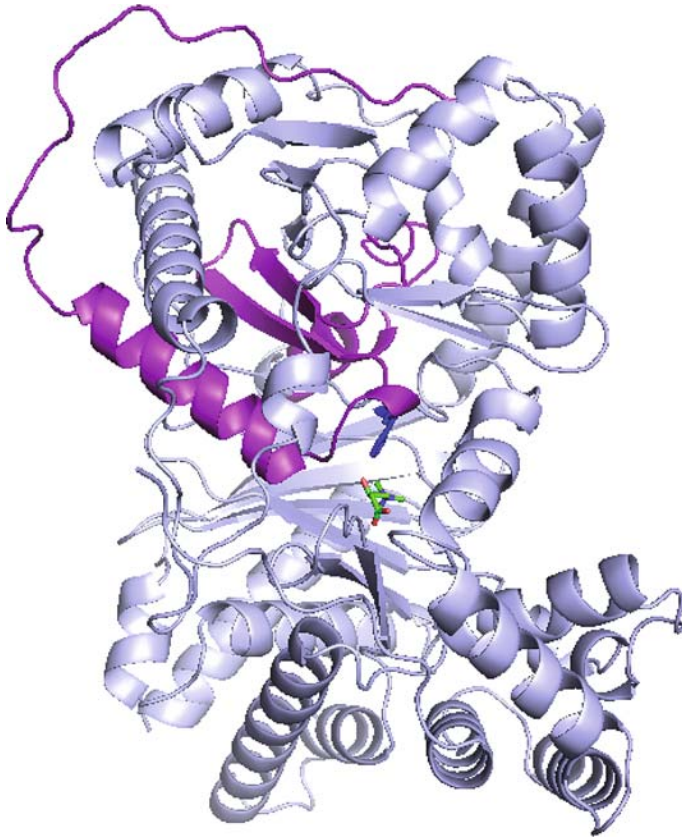


Fig. 7 The effect of splicing on protein structure I – CACP The structure of carnitine *O*-acetyltransferase has been resolved and is deposited in the PDB as structure 1nm8. Here we have mapped the sequence of putative splice isoform 001 onto the structure. The structure is coloured purple where the sequence of the splice isoform is missing. The deletion suggests that the structure isoform 001 would have to undergo substantial reorganisation to fold. The biologically relevant heteroatoms are shown in stick format and the catalytic residue that would be lost on account of the deletion is shown in blue

Figs. 7–9). Although these alternative isoforms must have markedly different structure and function from their constitutively spliced counterparts, exhaustive literature searches on the genes in this data set unearthed very little evidence to suggest they have a role as functional proteins. The effect of splicing on function *in vitro* is known for a few of the alternative isoforms in this set, but even in these cases we are still some way short of knowing their precise role in the cell.

It seems unlikely that the spectrum of conventional enzymatic or structural functions can be substantially extended through alternative splicing, but demonstrating the function of an alternative isoform would require detailed and technically complex

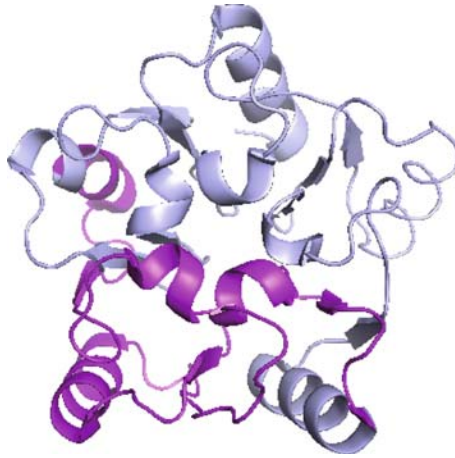


Fig. 8 The effect of splicing on protein structure II – ITGB4BP Splice isoform 005 of eukaryotic initiation factor 6 mapped onto the nearest structural template, yeast ITGB4BP, 1g62A. In this case the residues in purple indicate 89 residues that are missing from the structure. The missing residues remove almost exactly two of the repeats of the propeller structure. The missing residues are replaced by 69 non-homologous amino acids that are predicted to be without secondary structure. Again the protein would require drastic reorganisation in order to fold and will almost certainly lose all or most of its function

experimental approaches. At present most researchers can do little more than hypothesise as to the functional importance of splicing events.

So, if alternative splicing does not meaningfully extend the repertoire of conventional protein functions, what advantage is there to be gained in the cell from alternative splicing? Cells can encode a great many alternative transcripts; even the conservative estimate from this set suggests that alternative splicing can more than double the number of proteins in the cell. So why are there so many transcripts that appear likely to encode proteins that are non-functional, at least in the classical sense?

We cannot rule out the possibility that the expression of alternative transcripts might have implications for the control of gene expression, and indeed there are many transcripts with splicing events in the 3' and 5' untranslated regions. Some alternative splice isoforms may play important roles in development and tissue-specific processes. It seems possible that a number of alternative isoforms may have developed a useful cellular function such as the regulatory role suggested for the isoforms of IRAK1 (Rao et al. 2005) and IL-4 (Arinobu et al. 1999) from the HAVANA set. In addition there are a number of alternative variants with only minor sequence differences with respect to the principal isoform that may have a minor effect on the structure and function of the alternative variant. A small proportion of variants can be formed from homologous, mutually exclusive exons in the same way as in the Dscam gene and a few variants are formed by the addition or deletion of whole functional domains (the immunoglo-

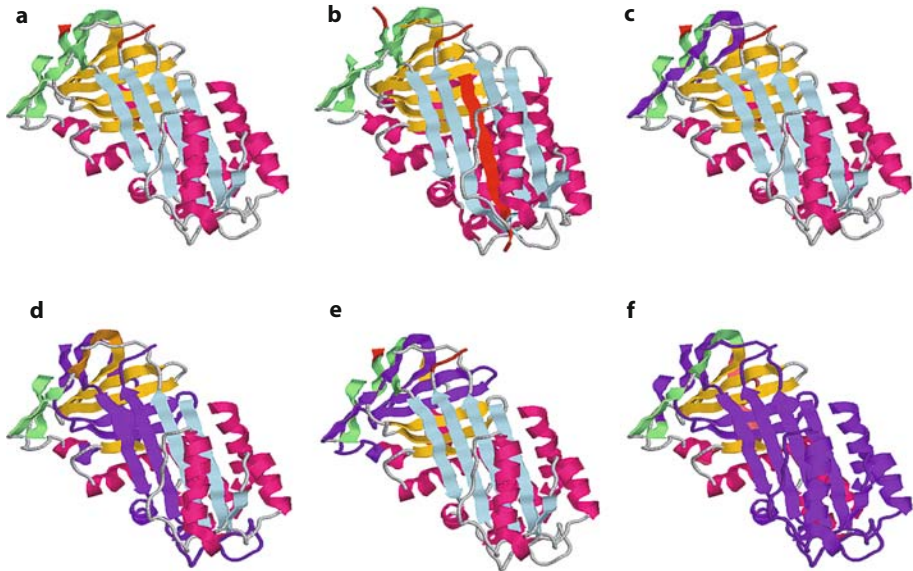


Fig. 9 The effect of splicing on protein structure III – B serpins Serpins are protease inhibitors that inactivate their targets after undergoing an irreversible conformational change. Serpins exist in an inactivated form (a) that is regarded as being “stressed”. Cleavage of the 20-residue RSL region, missing in the structure but with the terminal ends shown in red, causes the RSL region to flip over and fit itself into one of the beta sheets (b, inserted RSL strand in red). This exposes the inhibitory region that inactivates the protease. c–f Show four splice isoforms from different serpin loci mapped onto the structure of serpinB2 (1by7). The sections deleted/substituted in the isoforms are shown in purple. In each case it appears that splicing is likely to cause the structure to fold in a substantially different fashion. Given that the complex structure of the inhibitor is vital to its unique function, it is not clear why so many apparently deleterious isoforms would be necessary. This figure is copyright of the National Academy of Sciences, USA

bulin-based receptors are a clear example of this) and here it is easier to imagine that alternative splicing can be a source of functional modulation. However, there seems to be little to suggest that splicing events are directly related to protein domain boundaries (Kriventseva et al. 2003; Tress et al. 2007) as ought to be expected if splicing were directed by functional constraints.

The standard path of protein evolution is usually conceived as stepwise single base pair mutations. In contrast alternative splicing typically involves large insertions, deletions or substitutions of segments that may or may not correspond to functional domains, sub-cellular sorting signals or trans-membrane regions. The deletion and substitution of multiple exons seen in many of these transcripts suggests that splicing is not always a mechanism for delicate and subtle changes, and as a process may be rather more revolution than evolution. Unless external forces guide alternative splicing, splicing will lead to as many, if not more, evolutionary dead ends as standard evolutionary paths.

In fact alternative splicing can lead to a wide range of outcomes, many of which may be undesirable. There are a large number of splice variants that are likely to code for splice isoforms with dramatic changes in structure and function at the protein level. Changes of this magnitude would not normally be tolerated because of the heavy selection pressure that must oppose such large transformations (Xing and Lee 2006), which suggests that some expressed splice isoforms may have potentially deleterious functions and might only be highly expressed as a result of some disease event. Many of the variants in this set are supported by evidence that comes from diseased cells (Roy et al. 2005) and there is much evidence implicating splice variants in disease (Kishore and Stamm 2006; Ottenheim et al. 2006) and in particular as a secondary symptom in many types of cancer (Brinkman 2004; Pajares et al. 2007).

Despite the implication of splice variants in functional differentiation and diseased states it seems that many variants may have no clear effect on the cell. One possibility is that the cell can tolerate these variants to some extent; alternative isoforms expressed in low numbers may not adversely affect the cell. If this is the case, the selection pressure against exon loss or substitution will be reduced, making large evolutionary changes possible.

It is now important to extend the observations made for the small set of ENCODE pilot project genes to the full genome. The ENCODE scale up project seeks to accurately annotate the rest of the genome. This effort will be accompanied by automated annotation pipelines (see descriptions below) and will provide sufficient evidence to make our observations statistically significant.

The conclusions from this study and the extension to the complete genome have opened up a great opportunity for experimental work to validate the BioSapiens predictions. The following lines of work are already in progress for some of the most interesting variations from the HAVANA set:

- Confirmation of the expression in tissues by RT-PCR
- Confirmation of the expression of the corresponding proteins by raising antibodies against regions specific of splice variants. This obviously requires a substantial amount of work and takes time but can provide critical *in vitro* and *in vivo* evidence about the role of the proteins.
- Reverse proteomics studies looking for evidence of the expression of variants in databases of mass spectrometry (MS) spectra. While many of the proteins have been successfully identified, it has so far been difficult to confirm the presence of pairs of splice variants because the coverage of the MS databases is fairly low and to confirm variants the peptides must cover the splice boundaries for both variants. One alternative would be to carry out specific experiments in the search for peptides characteristic of splicing. This will also be important for addressing issues of quantification, since it is important to know not only if the protein isoforms are expressed but also in what quantities they are expressed.

6 Prediction of principal isoforms

Alternative splicing has the potential to generate a wide range of protein isoforms from the same gene. While many genes have been studied in depth, for others it is far from clear which of the variants retains the core biological function. For those genes where there is little experimental evidence it is important to know which is the principal variant in order to design experiments to determine structure and function. Labelling one of the splice variants as the principal isoform will allow research groups to concentrate their efforts on the main functional isoform. In addition, for many computational applications it is important to know which is the isoform that is likely to have the principal functional activity. Identifying a principal splice isoform for a gene would allow bioinformatics groups to make more reliable predictions of function and structure and would be a vital first step for large-scale studies, such as the ENCODE project (The ENCODE Project Consortium 2004, 2007). Automatic prediction pipelines in particular need reliable input data.

At present the SwissProt database (Bairoch et al. 2004), part of UniProtKB (The UniProt Consortium 2008), provides the best organisation of the complicated web of alternative protein variants. Even if the SwissProt database is only a small part of UniProt, it is the de facto gold standard of protein databases because all entries are manually curated. As part of this manual curation, all UniProt variants from the same gene are merged into a single SwissProt entry. One of the merged sequences is selected as the “display” sequence for the entry; the display sequence is selected after careful inspection and remaining merged sequences are tagged as alternative splice variants for the entry. The display sequence is often the longest variant, because this allows annotators to map more features to the sequence.

At present the selection of the SwissProt display sequence is of huge importance and has implications beyond SwissProt. SwissProt entries brings together experimental and predicted information, including domain definitions, functional annotation, cellular location, post-translational modifications and disease associations. All this information is associated to a single display isoform. The entries are also extensively cross-referenced to a range of external sources and all the functional and structural information is also associated to the display sequence. Those isoforms designated as alternative variants by SwissProt when entries are merged are left out of many versions of UniProt and these alternative isoforms also disappear from external databases.

For example the display sequence of PTPA HUMAN is the only isoform included in the Pfam domain database (Finn et al. 2006) and the ModBase (Pieper et al. 2006) model database. The display sequence is also built into the Pfam seed alignment generated for the PTPA domain. Pfam seed alignments are regarded as the gold standard for alignments and used in turn by many other groups. A structure has already been solved for this gene (by three separate groups), but unfortunately the variant that has been crystallised is missing the fourth protein coding exon of

the display sequence. This crystallised variant is almost certainly the principal isoform for the gene PPP2R4 (it has been shown to fold independently), but it has been left out of SwissProt and therefore also ModBase, Pfam and all related servers and databases.

6.1 A series of automatic methods for predicting the principal isoform

In order to define the principal coding variant for each gene, we had to make two assumptions. The first was that each gene has just a single variant that gives rise to a principal functional isoform. The remaining annotated variants would then be alternatively spliced. This is a general assumption and comparative studies usually suggest that one isoform has the principal function or is expressed in most tissues or in most stages of development. However, while this may be true for most genes, it will not be true for those genes where two (or more) variants might be regarded as equally important.

The second assumption is that this principal variant is evolutionarily conserved between species. Alternative exons tend to be recent evolutionary developments (Alekseyenko et al. 2007), so this is a reasonable assumption. Again this may not always be true for all genes – the principal variant may have evolved (possibly through alternative splicing) towards a function distinct from those performed by the orthologous gene products in neighbouring species. This means that for the purposes of this study we have defined the principal functional isoform as the isoform that performs an orthologous functional role across a wide range of related organisms.

We used the manually curated set of annotated splice variants produced by the HAVANA group (Harrow et al. 2006) for the 44 regions of the human genome analysed in the pilot project of the ENCODE project (The ENCODE Project Consortium 2004, 2007). The set contained 434 protein coding genes, but 181 genes in the set were annotated as having just a single splice variant and a further 38 genes had alternatively spliced transcripts that were protein sequence identical, differing only in the 5' and 3' untranslated regions. We concentrated our analysis on the 215 loci in the set that coded for at least two protein sequence distinct splice isoforms. There were 804 variants in this set, a mean of 3.74 variants per locus.

We used five separate methods to help determine the constitutive isoforms for the genes in the ENCODE project. The methods used in this study were complementary. The majority of methods were conservation-based, requiring evolutionary information in the form of genomic and protein sequences. Two methods (structure mapping and functional residue mapping) also required structural information in the form of homologous proteins with known structure.

As a working hypothesis all of the HAVANA annotated variants in a locus had an equal chance to be the principal variant. Most of our methods were used as a means of rejecting the hypothesis that a given variant could be the principal variant.

6.1.1 Methods

1. *Conservation of Exonic Structure*

Transcripts that do not have conserved exonic structure between species are not likely to code for the principal isoform. Transcripts with exonic structure that was not conserved were rejected as candidates for the principal variant.

2. *Non-neutral Evolution*

Exons with unusual substitution patterns might indicate biological phenomena, such as the generation of a new function in a subset of a species, but transcripts that contain one of these exons are unlikely to be the principal isoform. When one of the transcripts contained an exon with obvious non-standard conservation we did not consider the variant transcript as a candidate for the principal isoform. Non-standard evolution was evaluated using two methods, Prank (Loevelyntoja and Goldman 2005) and SLR (Massingham and Goldman 2005). See Fig. 10.

3. *Protein Structure*

Mapping Variants were also discounted as being principal functional variants if it was not possible to map their amino acid sequence onto a highly similar structural domain without introducing a deletion or insertion event caused by an alternatively spliced exon. Variants that can be mapped to structure without these gaps have more chance of being the functional variant because we know that they are likely to fold properly. As of 2006 there were only five examples of alternative isoforms with resolved protein structures (Romero et al. 2006). See Fig. 11.

4. *Functional Residue Conservation*

Exons that contain conserved functionally important residues are more likely to be part of the principal functional isoform of the protein. We used firestar (Lopez et al. 2007), a method that predicts functionally important residues in protein sequences.

5. *Vertebrate Alignments*

Here we were looking for numbers of species, the more species that had a variant that aligned correctly to each transcript, the better. A good alignment was an alignment without insertions or deletions caused by alternative exons. Good alignments with more distant relatives (Danio, Xenopus, chicken) were regarded as more valuable than alignments with chimpanzee or dog. If the transcript is conserved over a greater evolutionary distance it is more likely to be the constitutive variant. BLAST (Altschul et al. 1997) was used to search a non-redundant database of vertebrate sequences.

6.1.2 Evaluation of pipeline definitions

We were able to determine a principal isoform for 179 genes (83% of the set). While we were able to detect the principal variant for a high proportion of genes, it was impossible to determine a principal isoform for 36 genes. In part this was due to the clusters of

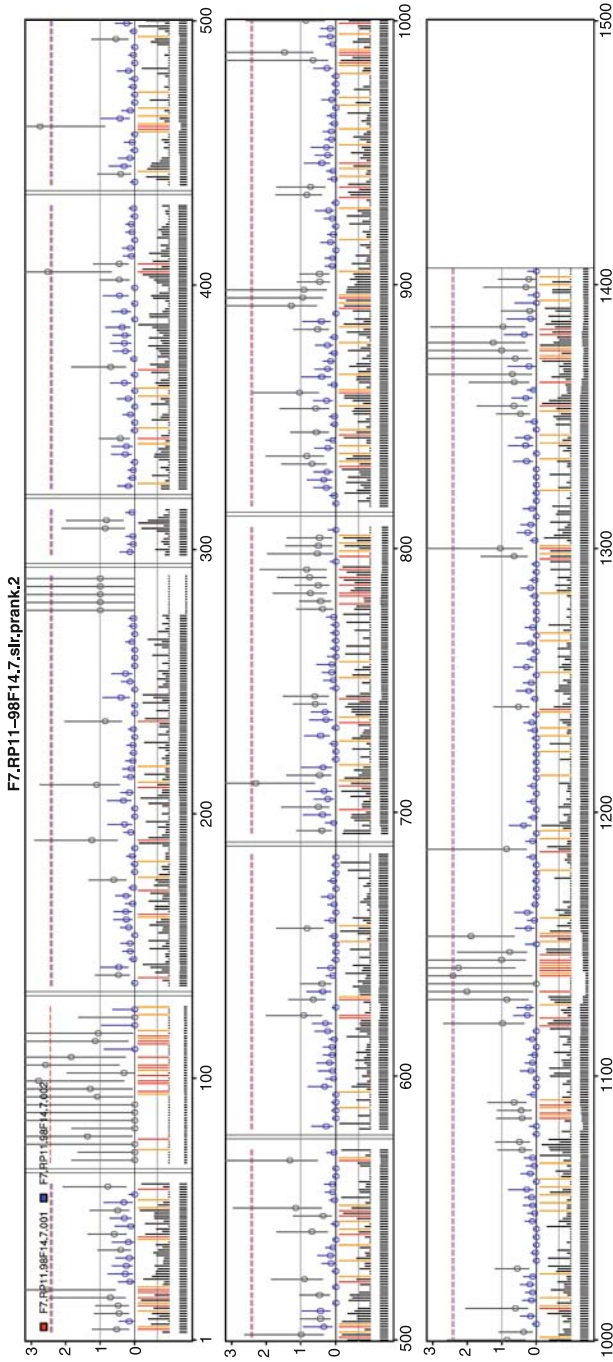


Fig. 10 The visual SLR output for the two variants of the F7 locus (RP11-98F14.7). At the top, the horizontal dashed lines indicate which exon belongs to which transcript (from top to bottom, 001 in red, 002 in blue). The colour of the circles (SLR score, Massingham and Goldman 2005) and bars (confidence intervals) denote selection mode. Below this is a per nucleotide measure of conservation, with abnormally fast sites coloured orange (3rd codon positions) or red (1st or 2nd codon positions). The black hatching at the foot of the record the number of sequences available at each alignment position. Double vertical black lines indicate exon boundaries. The whole of the 2nd coding exon (between approximately positions 75 and 125) is clearly differently conserved to the neighbouring exons, suggesting that it is under unusual selective pressures. This exon is present just in variant 001. On the basis of the SLR output we rejected the hypothesis that variant 001 (the SwissProt display sequence) could give rise to the principal functional isoform

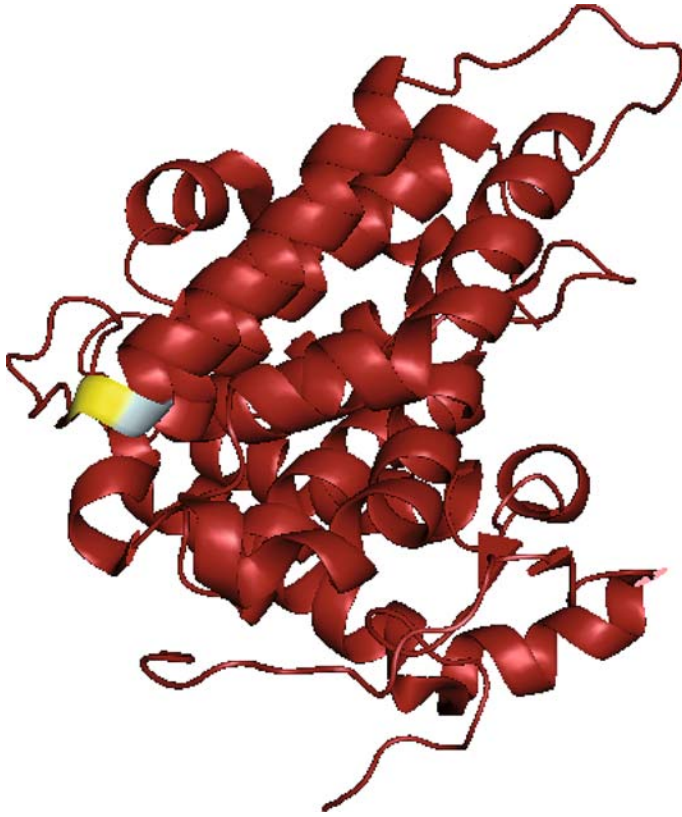


Fig. 11 Sequence to structure mapping. The crystal structure of human protein phosphatase 2A has been solved by three separate groups (for example, Magnusdottir et al. 2006). We were able to map the protein sequence of the 5 variants of the PPP2R4 gene (RP11-247A12.4) onto the protein phosphatase 2A structure (2g62). The SwissProt display sequence (from variant 012) has an insertion of 35 residues relative to the structure of the structure of protein phosphatase 2A. These 35 residues would need to be squeezed in between the residues marked in silver and yellow on the structure, breaking an alpha helix. The SwissProt display sequence is therefore unlikely to be the primary variant

immunoglobulin-based receptors in the ENCODE regions. These receptors are evolving very rapidly, which means that it is difficult to use conservation-based measures. For some genes our methods rejected all the alternative variants. Genes flagged in this way would clearly require further intervention.

A total of 153 of the 179 genes for which we could define the principal isoform had a Swiss-Prot display sequence. Here it was possible to compare our definition of the principal functional splice isoform with the display sequence assigned by Swiss-Prot. The Swiss-Prot display sequence differed from the principal isoform in 37 of the 153 genes.

We analysed the pipeline and SwissProt definitions by inspecting aligned genomic sequences from a wide range of vertebrate species. We also assessed transcription evidence from multiple sources. All the data, including the genomic sequence alignments, are available from the UCSC Genome Browser (<http://genome.ucsc.edu/cgi-bin/hgGateway>). Where possible we also carried out PubMed searches to look for experimental confirmation for the either variant.

For the majority of genes where the principal isoform differed from the SwissProt display sequence, the transcript and experimental evidence backed our selection as the principal isoform. For example the principal isoform we defined for the gene *SNX27* has been shown to be present in lysates from HEK293T cells (Rincon et al. 2007) while the SwissProt display sequence (*SNX27 HUMAN*) was not. For the gene *HYPK* the only transcript that supports the SwissProt display sequence is chimeric, links the *HYPK* locus to the *SERF2* locus upstream of *HYPK* and contains the *SERF2* start codon! Published support for this variant comes from just a single mass-sequencing paper. The start codon for the principal isoform selected by our methods is conserved in primates, rat and mouse.

The methods used to select the principal isoform can easily be automated and together with reliable annotations of splice variants, such as those from the HAVANA group, could form the basis of a pipeline to define principal functional variants for any genome.

7 The ENCODE pipeline – an automated workflow for analysis of human splice isoforms

The BioSapiens network focuses on protein annotation, and in relation to the ENCODE project (The ENCODE Project Consortium 2004, 2007) special attention has been given to alternative splicing and its putative effects on function. In the pilot phase of the BioSapiens project, the properties of the coding sequences within the 44 selected regions of the human genome were analysed separately by the network collaborators (Tress et al. 2007). The next step in the BioSapiens ENCODE project is to establish a scaled-up version of the annotation approach applied to the pilot sequences to cover the 100% of the human genome, including all isoforms. For the scale-up, the ENCODE Pipeline (EPipe) was constructed to allow researchers to compare functional annotations for all the splice variants of a given gene in an automatic way. The pipeline takes a set of protein isoforms as input, sends the sequences to a number of different annotation tools and compiles the predicted features into a graphical representation.

7.1 Behind EPipe

EPipe is currently implemented as a traditional interactive WWW service at: <http://www.cbs.dtu.dk/services/EPipe>. The input sequences are provided in FASTA format

and it is assumed that these sequences represent different splice variants of the same gene. The workflow of EPipe is divided into three major modules: alignment, analysis and presentation. A schematic overview is shown in Fig. 12. The alignment module

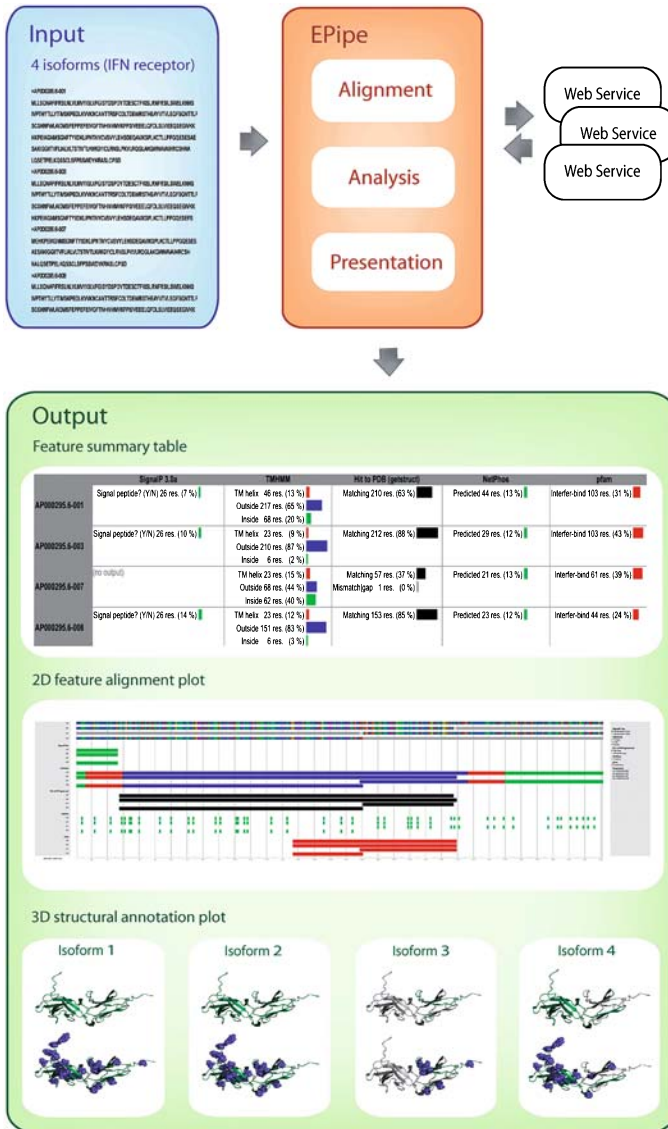


Fig. 12 Overview of the EPipe workflow

provides a multiple alignment of the input amino acid sequences. Four alignment programs are currently implemented: T-Coffee (Notredame et al. 2000), ClustalW (Thompson et al. 1994), Dialign (Morgenstern 2004), and Hmm3align (unpublished). Alternatively, a custom alignment can be provided by the user. The purpose of the analysis module is to create annotation predictions for each of the isoform sequences. In the current version of EPipe, four commonly used analysis methods are implemented; these include signal peptide prediction by SignalP (Bendtsen et al. 2004), transmembrane helix prediction by TMHMM (Krogh et al. 2001), domain identification by PFAM (Finn et al. 2006) and phosphorylation site prediction by NetPhos (Blom et al. 1999, 2004). In addition to these four methods, FeatureMap3D has been implemented (Wernersson et al. 2006). FeatureMap3D finds the best matching structure from the Protein Data Bank (Berman et al. 2000) for each sequence and generates images showing a subset of the predicted features in a structural context. When the results of the alignment and analysis modules are obtained by the parent EPipe process, they are passed on to the presentation module. Here, three output formats are generated: a condensed table providing a summary of predicted features, a graphical 2D alignment representation and 3D plots where the predicted features are mapped onto.

7.2 Example workflow: IFN alpha/beta receptor protein

Let us now take a look at an example workflow for EPipe (see Fig. 12). The input to EPipe is a FASTA file with isoform sequences and EPipe itself consists of three independent modules: alignment, analysis and presentation. As an input example we have used four variants of the human IFN alpha/beta receptor protein (variant 001, 003, 007 and 008) extracted from ENCODE gene AP000295.6 (Tress et al. 2007). The proteins were submitted using T-COFFEE as alignment procedure with standard parameters and the modules SignalP, TMHMM, NetPhos and PFAM were included. The output from EPipe contains three parts, a feature summary table where rows correspond to isoforms and columns to features, and where the functional differences between isoforms are summarised, a 2D feature alignment plot where the predicted features are projected onto the alignment (in the figure the alignment is shown at the top and single features are highlighted with different colours) and 3D structural annotation plots where predictions for each selected feature are projected onto 3D structures by colouring the relevant residues. The output can help identify putative functional changes resulting from alternative splicing events. From the summary table and the feature alignment we can see that isoforms 001, 003 and 008 have putative signal peptides, while isoform 007 does not. The result indicates probable different subcellular localisation for the isoforms. The structural annotation shows the location of a predicted feature, in this case phosphorylation sites, on the 3D structures from PDB.

7.3 Future perspectives

EPIPE is currently based on in-house modules for predicting the various features. Due to the modular workflow design, EPIPE can readily be extended to include more alignment methods and protein annotation tools. It is envisioned that EPIPE and related workflows in the near future will support connections to remote Web Services defined by the users. Such functionality does, however, require stringent standardisation and interoperability such as are currently in development under collaborations such as the EMBRACE Grid1. To facilitate post-processing, the entire EPIPE output can be obtained as a single XML file allowing researchers to systematically scan multiple isoform sets and set up custom criteria for extracting the desired information. For the scanning of several thousands splice variants, EPIPE demands human intervention to draw conclusions – both in terms of structural consequences of alternative splicing as well as for the general properties of the different isoforms. As a natural extension, EPIPE can be used as a tool for investigating functional differences within any set of related protein sequences such as homologues or polymorphic protein sequences.

References

- Alekseyenko AV, Kim N, Lee C J (2007) Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA* 13(5): 661–670
- Altschul S F, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman, DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17): 3389–3402
- Arinobu Y, Atamas SP, Otsuka T, Nihiro H, Yamaoka K, Mitsuyasu H, Niho Y, Hamasaki N, White B, Izuhara K (1999) Antagonistic effects of an alternative splice variant of human IL-4, IL-4delta2, on IL-4 activities in human monocytes and B cells. *Cell Immunol* 191(2): 161–167
- Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-Prot: juggling between evolution and stability. *Brief Bioinform* 5(1): 39–55
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: signalP 3.0. *J Mol Biol* 340(4): 783–795
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucleic Acids Res* 28(1): 235–242
- Bione S, D'Adamo P, Maestrini E, Gedeon AK, Bolhuis PA, Toniolo D (1996) A novel X-linked gene, G4.5, is responsible for Barth syndrome. *Nat Genet* 12(4): 385–389
- Black DL (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* 103(3): 367–370
- Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 294(5): 1351–1362
- Blom, N, Sicheritz-Pontn, T, Gupta, R, Gammeltoft, S, Brunak, S (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4(6): 1633–1649
- Brinkman, BMN (2004) Splice variants as cancer biomarkers. *Clin Biochem* 37(7): 584–594
- Brodsky G, Barnes T, Bleskan J, Becker L, Cox M, Patterson D (1997) The human GARS-AIRS-GART gene encodes two proteins which are differentially expressed during human brain development

- and temporally overexpressed in cerebellum of individuals with Down syndrome. *Hum Mol Genet* 6(12): 2043–2050
- Chen BE, Kondo M, Garnier A, Watson FL, Pettmann-Holgado R, Lamar, DR, Schmucker D (2006) The molecular diversity of Dscam is functionally required for neuronal wiring specificity in *Drosophila*. *Cell* 125(3): 607–620
- Ermak G, Gerasimov G, Troshina K, Jennings T, Robinson L, Ross JS, Figge, J (1995) Deregulated alternative splicing of CD44 messenger RNA transcripts in neoplastic and nonneoplastic lesions of the human thyroid. *Cancer Res* 55(20): 4594–4598
- Finn RD, Mistry, J, Schuster-Boeckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34(Database issue): D247–D251
- Florea L (2006) Bioinformatics of alternative splicing and its regulation. *Brief Bioinform* 7(1): 55–69
- Goodridge JP, Lathbury LJ, Steiner NK, Shulze CN, Pullikotil P, Seidah NG, Hurley CK, Christiansen FT, Witt CS (2007) Three common alleles of KIR2DL4 (CD158d) encode constitutively expressed, inducible and secreted receptors in NK cells. *Eur J Immunol* 37(1): 199–211
- Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* 17(2): 100–107
- Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, Lagarde J, Gilbert JGR, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 7(Suppl 1): S4.1–S4.9
- Henikoff S, Keene MA, Sloan JS, Bleskan J, Hards R, Patterson D (1986) Multiple purine pathway enzyme activities are encoded at a single genetic locus in *Drosophila*. *Proc Natl Acad Sci USA* 83(3): 720–724
- Hui, J, Bindereif, A (2005) Alternative pre-mRNA splicing in the human system: unexpected role of repetitive sequences as regulatory elements. *Biol Chem* 386(12): 1265–1271
- Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuče BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJA (2008) The 20 years of PROSITE. *Nucleic Acids Res* 36(Database issue): D245–D249
- Kall L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338(5): 1027–1036
- Karplus K, Sjolander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C (1997) Predicting protein structure using hidden Markov models. *Proteins (Suppl 1)*: 134–139
- Kishore S, Stamm S (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* 311(5758): 230–232
- Koedding J, Killig F, Polzer P, Howard SP, Diederichs K, Welte W (2005) Crystal structure of a 92-residue C-terminal fragment of TonB from *Escherichia coli* reveals significant conformational changes compared to structures of smaller TonB fragments. *J Biol Chem* 280(4): 3022–3028
- Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S (2003) Increase of functional diversity by alternative splicing. *Trends Genet* 19(3): 124–128
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305(3): 567–580
- Kurkela M, Morsky S, Hirvonen J, Kostianen R, Finel M (2004) An active and water-soluble truncation mutant of the human UDP- glucuronosyltransferase 1A9. *Mol Pharmacol* 65(4): 826–831
- Langer P, Gruender S, Ruesch A (2003) Expression of Ca²⁺-activated BK channel mRNA and its splice variants in the rat cochlea. *J Comp Neurol* 455(2): 198–209
- Laskowski RA (2007) Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. *Bioinformatics* 23(14): 1824–1827
- Lee D, Grant A, Marsden RL, Orengo C (2005) Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins* 59(3): 603–615

- Loeftynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA* 102(30): 10557–10562
- Lopez AJ (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet* 32: 279–305
- Lopez G, Valencia A, Tress ML (2007) firestar–prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res* 35(Web Server issue): W573–W577
- Lu B, Kelher MR, Lee DP, Lewin TM, Coleman RA, Choy PC, Hatch GM (2004) Complex expression pattern of the Barth syndrome gene product tafazzin in human cell lines and murine tissues. *Biochem Cell Biol* 82(5): 569–576
- Magnusdottir A, Stenmark P, Flodin S, Nyman T, Hammarstrom M, Ehn M, H, MAB, Berglund H, Nordlund P (2006) The crystal structure of a human PP2A phosphatase activator reveals a novel fold and highly conserved cleft implicated in protein-protein interactions. *J Biol Chem* 281(32): 22434–22438
- Martelli PL, Fariselli P, and Casadio R (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics* 19(Suppl 1): i205–i211
- Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics* 169(3): 1753–1762
- Matsushita K, Tomonaga T, Shimada H, Shioya A, Higashi M, Matsubara H, Harigaya K, Nomura F, Libutti D, Levens D, Ochiai T (2006) An essential role of alternative splicing of c-myc suppressor FUSE-binding protein-interacting repressor in carcinogenesis. *Cancer Res* 66(3): 1409–1417
- McGuffin LJ, Jones DT (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* 19(7): 874–881
- Meijers R, Puettmann-Holgado R, Skiniotis G, Liu J-H, Walz T, Wang J-H, Schmucker D (2007) Structural basis of Dscam isoform specificity. *Nature* 449(7161): 487–491
- Modrek B, Lee C (2002) A genomic view of alternative splicing. *Nat Genet* 30(1): 13–19
- Morgenstern B (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acids Res* 32(Web Server issue): W33–W36
- Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247(4): 536–540
- Navaratnam DS, Bell TJ, Tu TD, Cohen EL, Oberholtzer JC (1997) Differential distribution of Ca²⁺-activated K⁺ channel splice variants among hair cells along the tonotopic axis of the chick cochlea. *Neuron* 19(5): 1077–1085
- Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1): 205–217
- Novick D, Cohen B, Tal N, Rubinstein M (1995) Soluble and membrane- anchored forms of the human IFN-alpha/beta receptor. *J Leukoc Biol* 57(5): 712–718
- Nusbaum C, Zody MC, Borowsky ML, Kamal M, Kodira CD, Taylor TD, Whittaker CA, Chang JL, Cuomo CA, Dewar K, FitzGerald MG, Yang X, Abouelleil A, Allen NR, Anderson S, Bloom T, Bugalter B, Butler J, Cook A, DeCaprio D, Engels R, Garber M, Gnirke A, Hafez N, Hall JL, Norman CH, Itoh T, Jaffe DB, Kuroki Y, Lehoczky J, Lui A, Macdonald P, Mauceli E, Mikkelsen TS, Naylor JW, Nicol R, Nguyen C, Noguchi H, O’Leary SB, O’Neill K, Piqani B, Smith CL, Talamas JA, Topham K, Totoki Y, Toyoda A, Wain HM, Young SK, Zeng Q, Zimmer AR, Fujiyama A, Hattori M, Birren BW, Sakaki Y, Lander ES (2005) DNA sequence and analysis of human chromosome 18. *Nature* 437(7058): 551–555
- Ottenheijm CAC, Heunks LMA, Hafmans T, van der Ven PFM, Benoist C, Zhou H, Labeit S, Granzier HL, Dekhuijzen PNR (2006) Titin and diaphragm dysfunction in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 173(5): 527–534
- Pajares MJ, Ezponda T, Catena R, Calvo A, Pio R, Montuenga LM (2007) Alternative splicing: an emerging topic in molecular and clinical oncology. *Lancet Oncol* 8(4): 349–357

- Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res* 33(Database issue): D247–D251
- Pennisi E (2005) Why do humans have so few genes? *Science* 309(5731): 80
- Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen M-Y, Kelly L, Melo F, Sali A (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 34 (Database issue): D291–D295
- Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* 22(14): e408–e416
- Porter CT, Bartlett GJ, Thornton JM (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32(Database issue): D129–D133
- Ramanathan K, Michael TH, Jiang GJ, Hiel H, Fuchs PA (1999) A molecular mechanism for electrical tuning of cochlear hair cells. *Science* 283(5399): 215–217
- Rao N, Nguyen S, Ngo K, Fung-Leung W-P (2005) A novel splice variant of interleukin-1 receptor (IL-1R)-associated kinase 1 plays a negative regulatory role in Toll/IL-1R-induced inflammatory signaling. *Mol Cell Biol* 25(15): 6521–6532
- Rincon E, Santos T, Avila-Flores A, Albar JP, Lalioti V, Lei C, Hong W, Merida I (2007) Proteomics identification of sorting nexin 27 as a diacylglycerol kinase zeta-associated protein: new diacylglycerol kinase roles in endocytic recycling. *Mol Cell Proteomics* 6(6): 1073–1087
- Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci USA* 103(22): 8390–8395
- Rosenblatt KP, Sun ZP, Heller S, Hudspeth AJ (1997) Distribution of Ca^{2+} -activated K^{+} channel isoforms along the tonotopic gradient of the chicken's cochlea. *Neuron* 19(5): 1061–1075
- Roy M, Xu Q, Lee C (2005) Evidence that public database records for many cancer-associated genes reflect a splice form found in tumors and lack normal splice forms. *Nucleic Acids Res* 33(16): 5026–5033
- Scherer SE, Muzny DM, Buhay CJ, Chen R, Cree A, Ding Y, Dugan-Rocha S, Gill R, Gunaratne P, Harris RA, et al. (2006) The finished DNA sequence of human chromosome 12. *Nature* 440(7082): 346–351
- Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL (2000) *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101(6): 671–684
- Smith CW, Valcarcel J (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci* 25(8): 381–388
- Soeding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33(Web Server issue): W244–W248
- Stojic J, Stoehr H, Weber BHF (2007) Three novel ABCC5 splice variants in human retina and their role as regulators of abcc5 gene expression. *BMC Mol Biol* 8: 42
- Talavera D, Vogel C, Orozco M, Teichmann SA, de la Cruz X (2007) The (in)dependence of alternative splicing and gene duplication. *PLoS Comput Biol* 3(3): e33
- Taylor TD, Noguchi H, Totoki Y, Toyoda A, Kuroki Y, Dewar K, Lloyd C, Itoh T, Takeda T, Kim D-W, She X, Barlow KF, Bloom T, Bruford E, Chang JL, Cuomo CA, Eichler E, FitzGerald MG, Jaffe DB, LaButti K, Nicol R, Park H.-S, Seaman C, Sougnez C, Yang X, Zimmer AR, Zody MC, Birren BW, Nusbaum C, Fujiyama A, Hattori M, Rogers J, Lander ES, Sakaki Y (2006) Human chromosome 11 DNA sequence and analysis including novel gene identification. *Nature* 440(7083): 497–500

- The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* 306(5696): 636–640
- The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146): 799–816
- The UniProt Consortium (2008). The universal protein resource (UniProt) *Nucleic Acids Res* 36 (Database issue): D190–D195
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22(22): 4673–4680
- Tramontano A, Morea V (2003) Assessment of homology-based predictions in CASP5. *Proteins* 53 (Suppl 6): 352–368
- Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PL, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski RA, Lopez G, Sadowski MI, Watson JD, Fariselli P, Rossi I, Nagy A, Kai W, Stirling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramirez F, Schlicker A, Denoeud F, Jones P, Kerrien S, Orchard S, Antonarakis SE, Reymond A, Birney E, Brunak S, Casadio R, Guigo R, Harrow J, Hermjakob H, Jones DT, Lengauer T, Orengo CA, Patthy L, Thornton JM, Tramontano A, Valencia A (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc Natl Acad Sci USA* 104(13): 5495–5500
- Vaz FM, Houtkooper RH, Valianpour F, Barth PG, Wanders RJA (2003) Only one splice variant of the human TAZ gene encodes a functional protein with a role in cardiolipin metabolism. *J Biol Chem* 278(44): 43089–43094
- Vielhauer GA, Fujino H, Regan JW (2004) Cloning and localization of hFP(S): a six-transmembrane mRNA splice variant of the human FP prostanoid receptor. *Arch Biochem Biophys* 421(2): 175–185
- Viklund H, Elofsson A (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden markov models and evolutionary information. *Protein Sci* 13(7): 1908–1917
- Wells CA, Chalk AM, Forrest A, Taylor D, Waddell N, Schroder K, Himes SR, Faulkner G, Lo S, Kasukawa T, Kawaji H, Kai C, Kawai J, Katayama S, Carninci P, Hayashizaki Y, Hume DA, Grimmond SM (2006) Alternate transcription of the Toll-like receptor signaling cascade. *Genome Biol* 7(2): R10
- Wernersson R, Rapacki K, Staerfeldt H-H, Sackett PW, Mlgaard A (2006) FeatureMap3D – a tool to map protein features and sequence conservation onto homologous structures in the PDB. *Nucleic Acids Res* 34(Web Server issue): W84–W88
- Wojtowicz WM, Flanagan JJ, Millard SS, Zipursky SL, Clemens JC (2004) Alternative splicing of *Drosophila* Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell* 118(5): 619– 633
- Xing Y, Lee C (2006) Alternative splicing and RNA selection pressure– evolutionary consequences for eukaryotic genomes. *Nat Rev Genet* 7(7): 499–509
- Xu Y, Malhotra A, Ren M, Schlame M (2006) The enzymatic function of tafazzin. *J Biol Chem* 281(51): 39217–39224

CONTRIBUTORS

Ben Adams

Max-Planck-Institut für Informatik
66123 Saarbrücken, Germany
E-mail: b.adams@bath.ac.uk

Mario Albrecht

Max-Planck-Institute for Informatics
66123 Saarbrücken, Germany
E-mail: mario.albrecht@mpi-sb.mpg.de

Tyler Alioto

Center for Genomic Regulation
08003 Barcelona, Spain
E-mail: tyler.alioto@crg.es

Irena Artamonova

German Research Center for
Environmental Health
85764 Neuherberg, Germany

Vavilov Institute of General Genetics RAS
119991 Moscow, Russia
E-mail: irena.artamonova@gsgf.de

L. Banyai

Hungarian Academy of Sciences
Institute of Enzymology
1518 Budapest, Hungary
E-mail: banyai@enzim.hu

Hagen Blankenburg

Max-Planck-Institute for Informatics
66123 Saarbrücken, Germany
E-mail: hagen@mpi-inf.mpg.de

Brigitte Boeckmann

Swiss Institute of Bioinformatics
1211 Geneva 4, Switzerland
E-mail: brigitte.boeckmann@isb-sib.ch

Peer Bork

European Molecular Biology Laboratory
69117 Heidelberg, Germany
Max-Delbrück-Centre for Molecular
Medicine
13092 Berlin, Germany
E-mail: bork@embl.de

P.-Y. Bourguignon

CEA – Institut de Genomique –
Genoscope
91057 Evry, France
E-mail: pbourgui@genoscope.cns.fr

Alvis Brazma

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD, UK
E-mail: brazma@ebi.ac.uk

Alan J. Bridge

Swiss Institute of Bioinformatics
1211 Geneva 4, Switzerland
E-mail: alan.bridge@isb-sib.ch

Rita Casadio

University of Bologna
40126 Bologna, Italy
E-mail: casadio@alma.unibo.it

Richard Coulson

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD, UK
E-mail: coulson@ebi.ac.uk

Antoine Danchin

Institut Pasteur
75724 Paris Cedex 15, France
E-mail: antoine.danchin@normalesup.org

Francisco S. Domingues

Max-Planck-Institute for Informatics
66123 Saarbrücken, Germany
E-mail: doming@mpi-sb.mpg.de

Piero Fariselli

University of Bologna
40126 Bologna, Italy
E-mail: piero@biocomp.unibo.it

K. Farkas

Hungarian Academy of Sciences
Institute of Enzymology
1518 Budapest, Hungary
E-mail: farkas@enzim.hu

Dmitrij Frishman

German Research Center for
Environmental Health
85764 Neuherberg, Germany

Technische Universität München
Wissenschaftszentrum Weihenstephan
85350 Freising, Germany
E-mail: d.frishman@wzw.tum.de

Alejandro Giorgetti

University of Verona
37134 Verona, Italy

University of Rome “La Sapienza”
00185 Rome, Italy
E-mail: alejandro.giorgetti@uniroma1.it

Roderic Guigó

Center for Genomic Regulation
08003 Barcelona, Spain
E-mail: roderic.guigo@crg.es

Peter F. Hallin

Technical University of Denmark
2800 Lyngby, Denmark
E-mail: pfh@cbs.dtu.dk

Eoghan D. Harrington

European Molecular Biology
Laboratory
69117 Heidelberg, Germany
E-mail: eogham.harrington@embl.de

H. Hegyi

Hungarian Academy of Sciences
Institute of Enzymology,
1518 Budapest, Hungary
E-mail: hegyi@emzim.hu

Henning Hermjakob

EMBL-EBI,
Wellcome Trust Genome Campus,
Hinxton, Cambridge CB10 1SD, UK
E-mail: hhe@ebi.ac.uk

Rafael C. Jimenez

EMBL-EBI,
Wellcome Trust Genome Campus,
Hinxton, Cambridge CB10 1SD, UK

National Bioinformatics Network
Cape Town, South Africa
E-mail: rafael@nbn.ac.za

David Jones

University College London
London WC1E 6BT, UK
E-mail: D.Jones@cs.ucl.ac.uk

David A. Juan

Spanish National Cancer Research Centre
(CNIO)
28029 Madrid, Spain
E-mail: dadejuan@cnio.es

Lars Juhl Jensen

European Molecular Biology Laboratory
69117 Heidelberg, Germany

University of Copenhagen
2200 Copenhagen, Denmark
E-mail: lars.juhl.jensen@gmail.com

Agnieszka S. Juncker

Technical University of Denmark
2800 Lyngby, Denmark
E-mail: ag@cbs.dtu.dk

E. Kozma

Hungarian Academy of Sciences
Institute of Enzymology
1518 Budapest, Hungary
E-mail: kozma@enzim.hu

Stefan Kramer

Technische Universität München
85748 Garching b. München, Germany
E-mail: kramer@in.tum.de

Eleonora Kulberkyte

Technical University of Denmark
2800 Lyngby, Denmark
E-mail: eleonora@cbs.dtu.dk

Eugene Kulesha

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD, UK
E-mail: ek@ebi.ac.uk

Thomas Lengauer

Max-Planck-Institute for Informatics
66123 Saarbrücken, Germany
E-mail: lengauer@mpi-sb.mpg.de

Gonzalo Lopez

Spanish National Cancer Research Centre
(CNIO)
28029 Madrid, Spain
E-mail: glopez@cnio.es

Claus Lundegaard

Technical University of Denmark
2800 Lyngby, Denmark
E-mail: lunde@cbs.dtu.dk

Thomas Manke

Max Planck Institute for Molecular
Biology
14195 Berlin, Germany
E-mail: manke@molgen.mpg.de

Pierluigi Martelli

University of Bologna
40126 Bologna, Italy
E-mail: gigi@lipid.biocomp.unibo.it

Claudine Médigue

CNRS UMR8030
91057 Evry Cedex, France

Commissariat à l'Énergie Atomique
(CEA)

Direction des Sciences du Vivant
91057 Evry Cedex, France
E-mail: cmedigue@infobiogen.fr

Alice Carolyn McHardy

Max-Planck-Institute for Informatics
66123 Saarbrücken, Germany
E-mail: mchardy@mpi-inf.mpg.de

Richard Mott

Wellcome Trust Centre for Human
Genetics, University of Oxford
Oxford OX3 9JN, UK
E-mail: richard.mott@well.ox.ac.uk

Nicola J. Mulder

EMBL-EBI,
Wellcome Trust Genome Campus,
Hinxton, Cambridge CB10 1SD, UK

National Bioinformatics Network Node
Institute for Infectious Disease
and Molecular Medicine
University of Cape Town, South Africa
E-mail: Nicola.Mulder@uct.ac.za

A. Nagy

Hungarian Academy of Sciences
Institute of Enzymology,
1518 Budapest, Hungary
E-mail: nagy@enzim.hu

Christine A. Orengo

University College London
London WC1E 6BT, UK
E-mail: orengo@biochemistry.ucl.ac.uk

Christos Ouzounis

King's College London
London WC2R 2LS
E-mail: ouzounis@kcl.ac.uk

Philipp Pagel

Technische Universität München
Wissenschaftszentrum Weihenstephan
85350 Freising, Germany
E-mail: p.pagel@wzw.tum.de

Kimmo Palin

University of Helsinki
FIN-00014 Helsinki, Finland
E-mail: kimmo.palin@cs.helsinki.fi

Laszlo Patthy

Hungarian Academy of Sciences
Institute of Enzymology
1518 Budapest, Hungary
E-mail: patthy@enzim.hu

Andrea Pierleoni

University of Bologna
40126 Bologna, Italy
E-mail: andrea@biocomp.unibo.it

Andreas Prlic

The Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA, UK
E-mail: andreas@sdsc.edu

Domenico Raimondo

University of Rome "La Sapienza"
00185 Rome, Italy
E-mail: domenico.raimondo@uniroma1.it

Antonio Rausell

Spanish National Cancer Research Centre
(CNIO)
28029 Madrid, Spain
E-mail: arausell@cnio.es

Oliver Redfern

University College London
London WC1E 6BT, UK
E-mail: o.redfern@ucl.ac.uk

Gabrielle A. Reeves

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD, UK
E-mail: gabby@ebi.ac.uk

Helge Roider

Max Planck Institute for Molecular
Biology
14195 Berlin, Germany
E-mail: roider@molgen.mpg.de

Ivan Rossi

University of Bologna
40126 Bologna, Italy
E-mail: ivan@biocomp.unibo.it

Leszek Rychlewski

BioInfoBank Institute
60-744 Poznan, Poland
E-mail: leszek@bioinfo.pl

Olivier Sand

Université Libre de Bruxelles
1050 Bruxelles, Belgium
E-mail: oly@scmbb.ulb.ac.be

Vincent Schächter

Genoscope – CEA
91057 Evry Cedex, France
E-mail: vs@genoscope.cns.fr

Ingolf Sommer

Max-Planck-Institute for Informatics
66123 Saarbrücken, Germany
E-mail: sommer@mpi-sb.mpg.de

Janet M. Thornton

EMBL-EBI,
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD, UK
E-mail: doffice@ebi.ac.uk

H. Tordai

Hungarian Academy of Sciences
Institute of Enzymology
1518 Budapest, Hungary
E-mail: tordaih@enzim.hu

Anna Tramontano

University of Rome “La Sapienza”
00185 Rome, Italy
E-mail: anna.tramontano@uniroma1.it

Michael L. Tress

Spanish National Cancer Research Centre
(CNIO)
28029 Madrid, Spain
E-mail: mtress@cnio.es

Jean-Valéry Turatsinze

Université Libre de Bruxelles
1050 Bruxelles, Belgium
E-mail: jturatsi@ulb.ac.be

Esko Ukkonen

University of Helsinki
FIN-00014 Helsinki, Finland
E-mail: ukkonen@cs.helsinki.fi

Alfonso Valencia

Spanish National Cancer Research Centre
(CNIO)
28029 Madrid, Spain
E-mail: valencia@cnio.es

Jacques van Helden

Université Libre de Bruxelles
1050 Bruxelles, Belgium
E-mail: Jacques.van.Helden@ulb.ac.be

Anne-Lise Veuthey

Swiss Institute of Bioinformatics
1211 Geneva 4, Switzerland
E-mail: anne-lise.veuthey@isb-sib.ch

Martin Vingron

Max Planck Institute for Molecular
Biology
14195 Berlin, Germany
E-mail: vingron@molgen.mpg.de

Gunnar von Heijne

Stockholm University
106 91 Stockholm, Sweden
E-mail: unnar@dbb.su.se

G. Vriend

NCMLS, Radboud University Nijmegen
Medical Centre
6525 GA Nijmegen, The Netherlands
E-mail: vriend@cmbi.ru.nl

B. Vroling

NCMLS, Radboud University Nijmegen
Medical Centre
6525 GA Nijmegen, The Netherlands
E-mail: bvroling@cmbi.ru.nl

Kai Wang

Technical University of Denmark
2800 Lyngby, Denmark
E-mail: wangkai@cbs.dtu.dk

James D. Watson

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD, UK
E-mail: Watson@ebi.ac.uk

Jan-Jaap Wesselink

Spanish National Cancer Research Centre
(CNIO)
28029 Madrid, Spain
E-mail: jjwesselink@cnio.es

Corin Yeats

University College London
London WC1E 6BT, UK
E-mail: yeats@biochemistry.ucl.ac.uk