GS
Garland Science

Richard M. Twyman

Principles of PROTEOMICS

SECOND EDITION

Principles of Proteomics

Second Edition

Richard M. Twyman

Garland Science
Taylor & Francis Group
NEW YORK AND LONDON

Garland Science
Vice President: Denise Schanck
Assistant Editor: David Borrowdale
Production Editor and Layout: Ioana Moldovan
Copy Editor: Sally Huish
Proofreader: Mac Clarke
Illustrations: Oxford Designers & Illustrators
Cover Design: Armen Kojoyian
Indexer: Bill Johncocks

Front cover image:
Courtesy of Gabriel Mazzucchelli, Mass Spectrometry Laboratory, GIGA Proteomics, University of Liège, Belgium.

About the author:
Richard M. Twyman studied genetics at Newcastle University, where he gained a first-class honors degree, and then obtained his doctorate in molecular biology at Warwick University. After working as a postdoctoral research fellow at the MRC Laboratory of Molecular Biology in Cambridge, he became a full-time scientific writer, initially at the John Innes Centre in Norwich and then as the director of Twyman Research Management Ltd, a company that develops and manages scientific projects and provides assistance with the preparation of scientific manuscripts. He is the author of many science textbooks and is actively involved in many current research projects and lecture courses. He is a visiting professor of biotechnology at the University of Lleida in Spain.

ISBN 978-0-8153-4472-8

Library of Congress Cataloging-in-Publication Data
Twyman, Richard M.
Principles of proteomics / Richard M. Twyman. -- Second edition.
pages cm
Includes bibliographical references.
ISBN 978-0-8153-4472-8 (alk. paper)
1. Proteomics. 2. Proteins. I. Title.
QP551.T94 2014
572'.6--dc23
2013021694

Published by Garland Science, Taylor & Francis Group, LLC,
an informa business,
711 Third Avenue, New York, NY, 10017, USA, and 3 Park Square,
Milton Park, Abingdon, OX14 4RN, UK.

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Garland Science
Taylor & Francis Group

Visit our website at http://www.garlandscience.com

Principles of Proteomics

Second Edition

Preface to the second edition

When I wrote the first edition of *Principles of Proteomics* in 2003, it was the first book that had attempted to cover the entire field of proteomics in broad strokes rather than focusing on specialized individual technologies. The first edition was published when proteomics was an emerging discipline, still unsure of its footing although confident in its abilities, with many technology platforms jostling for attention and consideration. Nearly a decade later, writing the second edition has proven a significant challenge. Although proteomics has stabilized, with certain technologies becoming unshakably established and others becoming obsolete, the cutting edge still boasts a rich and diverse source of novel technology platforms seeking to capture the proteome in ever more detail and on a scale barely conceived at the beginning of the millennium. But proteomics has also become increasingly commercialized. It is a billion-dollar industry, with many companies vying for attention, providing technologies, solutions, and contract research to other companies, who are in turn interested in using proteomics to find disease biomarkers, drug targets, vaccine candidates, novel chemical inhibitors, improved enzymes for industrial processes, and products to protect plants, the food chain, and the environment. Keeping up with the pace of change while still being aware of the fundamental aspects of proteomics, the core principles that make it possible in the first place, is a difficult task made more difficult by the dominant position of proprietary technologies, and the explosion in patents relating to proteomic technologies and strategies for processing proteomics data.

Despite the above, we must remember that proteomics is still about the global analysis of proteins. It seeks to achieve what genomics cannot—that is, a complete description of living cells in terms of all their functional components, brought about by the direct analysis of those components rather than the genes that encode them. Proteins offer a rich source of data, including sequences, structures, and biochemical and biological functions, which are influenced by modifications, subcellular localization, and, perhaps most important of all, the interactions among proteins and with other molecules. If genes are the instruction carriers, proteins are the molecules that execute those instructions. Genes are the instruments of change over evolutionary timescales, but proteins are the molecules that define which changes are accepted and which are discarded. It is from proteins that we shall learn how living cells and organisms are built and maintained and what leads to their dysfunction.

Although now firmly established, proteomics is still a difficult subject to penetrate for those not familiar with the terminology and technology, including experts in one area of proteomics venturing into another. There is still a great deal of jargon and many hyphenated acronyms that make sense once explained but otherwise remain mystifying; and there is still a high turnover of methods at the cutting edge, making it difficult to keep up. This situation is exacerbated by the increasing integration of proteomics with other areas of large-scale biology as researchers attempt to model cellular processes by looking not only at the functional components, but also at the information (genes, transcripts) and the outputs (metabolites, phenotypes) and how these are linked into networks and systems.

As I stated in the preface to the first edition, it is my hope that this book will be useful to those who need a broad overview of proteomics and what it has to offer. It is not meant to provide expertise in any particular area: there are plenty of other books that deal with specific technologies and their applications, the processing and archiving of proteomic data, and the integration of proteomics with other disciplines. The aim of this book is to pull together the different proteomics technologies and their applications, and present them in what I hope is a simple, logical, and user-friendly manner. After a brief introductory chapter providing an updated perspective on the history of proteomics since the turn of the millennium, the major proteomics technologies are discussed in more detail: two-dimensional gel electrophoresis, multidimensional liquid chromatography, mass spectrometry, sequence analysis, structural analysis, methods for studying protein interactions and modifications, and the development and applications of protein microarrays. These chapters have been broadened to account for new developments since the first edition, but I have made every effort to keep the material as concise as possible, since the brevity of the first edition was one of its strengths. I have assumed necessarily that the reader has a working knowledge of molecular biology and biochemistry. Each chapter has a short bibliography listing classic papers and useful reviews that will help the interested reader delve deeper into the literature.

The second edition would not have been possible without the help and support of the editorial team at Garland Science, so I extend special thanks to Gina Almond, David Borrowdale, and Ioana Moldovan for their dedication and assistance during the writing and revision process. I would also like to thank friends and colleagues who provided feedback on the first edition or suggestions for the second edition or who pointed out errors and omissions.

As ever, this book is dedicated with love to my parents, Peter and Irene, to my children, Emily and Lucy, and to Hannah, Joshua, and Dylan.

Richard M. Twyman

August 2013

Instructor Resources Website

Accessible from www.garlandscience.com, the Instructor Resource Site requires registration and access is available only to qualified instructors. To access the Instructor Resource Site, please contact your local sales representative or email science@garland.com.

The images in *Principles of Proteomics* are available on the Instructor Resource Site in two convenient formats: PowerPoint® and JPEG, which have been optimized for display. The resources may be browsed by individual chapter or a search engine.

Resources available for other Garland Science titles can be accessed via the Garland Science Website.

Acknowledgments

The author and publisher of *Principles of Proteomics* gratefully acknowledge the contributions of the following reviewers in the development of this book:

Vasco A. de Carvalho Azevedo	Universidade Federal de Minas Gerais, Brazil
Venkatesha Basrur	University of Michigan, USA
Richard Edwards	University of Southampton, UK
Rob Ewing	Case Western Reserve University, USA
Yao-Te Huang	College of Life Sciences, China Medical University, Taiwan
André Klein	Hogeschool Leiden, Netherlands
Sunny Liu	North Carolina State University, USA
Metodi Metodiev	University of Essex, UK
Peter Nilsson	AlbaNova University Center, Sweden
Joanna Rees	University of Cambridge, UK
Dacheng Ren	Syracuse University, USA
Anikó Váradi	The University of the West of England, UK

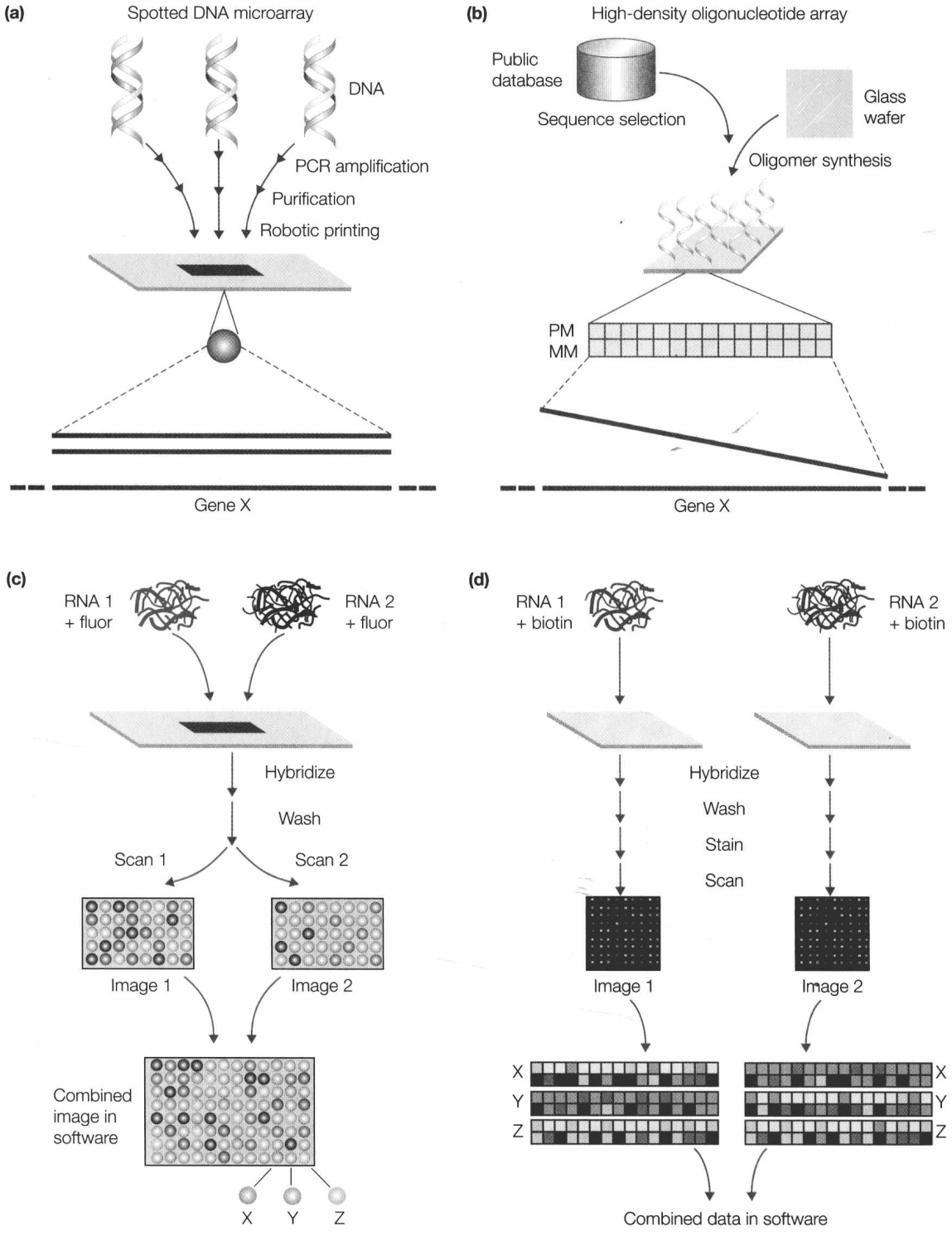


FIGURE 1.4 Expression analysis with DNA microarrays. (a) Spotted microarrays are produced by the robotic printing of amplified cDNA molecules onto glass slides. Each spot or feature corresponds to a contiguous gene fragment of several hundred base pairs or more. (b) High-density oligonucleotide chips are manufactured using a process of light-directed combinatorial chemical synthesis to produce thousands of different sequences in a highly ordered array on a small glass chip. Genes are represented by 15–20 different oligonucleotide pairs (PM, perfectly matched; MM, mismatched) on the array. (c) On spotted arrays, comparative expression assays are usually carried out by differentially labeling two mRNA or cDNA samples with different fluorophores. These are hybridized to features on the glass slide and then scanned to detect both fluorophores independently. Colored dots labeled X, Y, and Z at the bottom of the image correspond to transcripts present at increased levels in sample 1 (X), increased levels in sample 2 (Y), and similar levels in samples 1 and 2 (Z). (d) On Affymetrix GeneChips, biotinylated cRNA is hybridized to the array and stained with a fluorophore conjugated to avidin. The signal is detected by laser scanning. Sets of paired oligonucleotides for hypothetical genes present at increased levels in sample 1 (X), increased levels in sample 2 (Y), and similar levels in samples 1 and 2 (Z) are shown. (From Harrington CA, Rosenow C & Retief J (2000) *Curr. Opin. Microbiol.* 3, 285. With permission from Elsevier.)

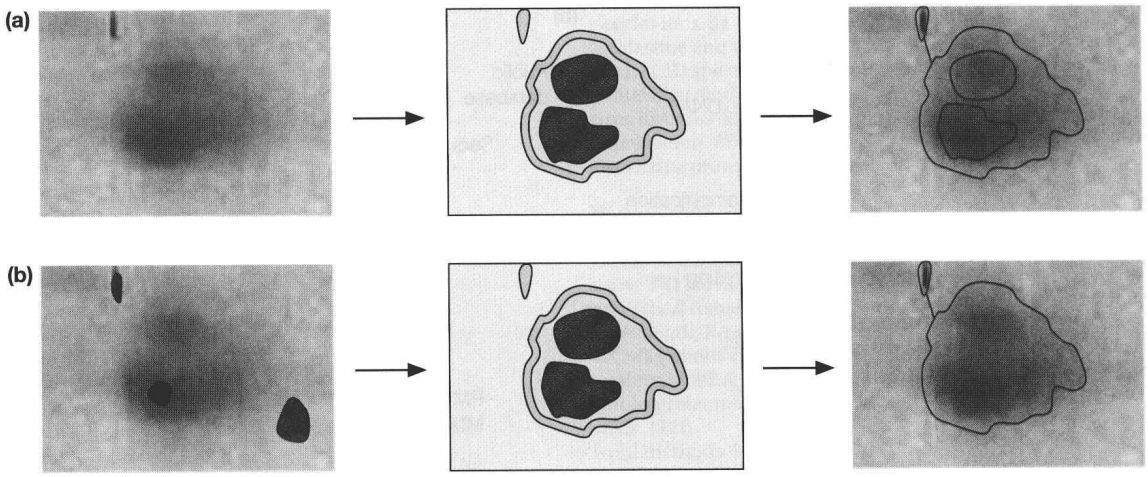


FIGURE 4.1 The watershed method for contour finding on two-dimensional gel images. (a) Any grayscale image can be considered as a topographic surface. If flooded from its minima without allowing water from different sources to merge, the image is partitioned into catchment basins and watershed lines, but in practice this leads to over-segmentation. (b) Therefore, markers (*red shapes*) are used to initiate flooding, and this reduces over-segmentation considerably. (Adapted from images by Serge Beucher, CMM/École Nationale Supérieure des Mines de Paris.)

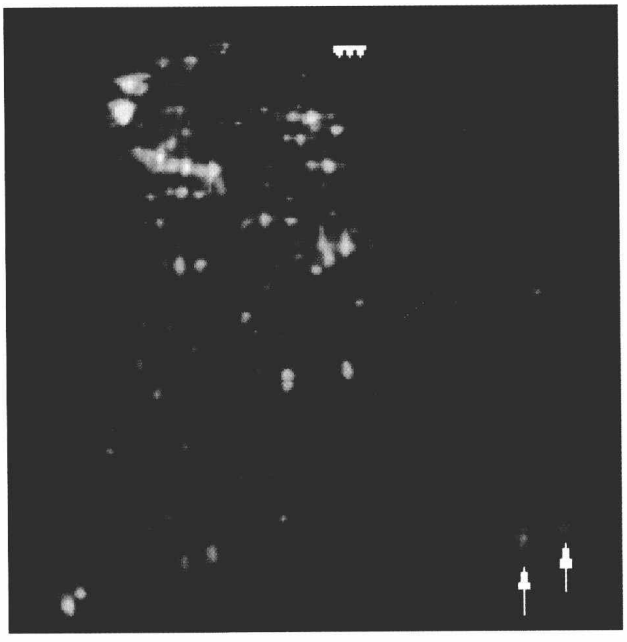
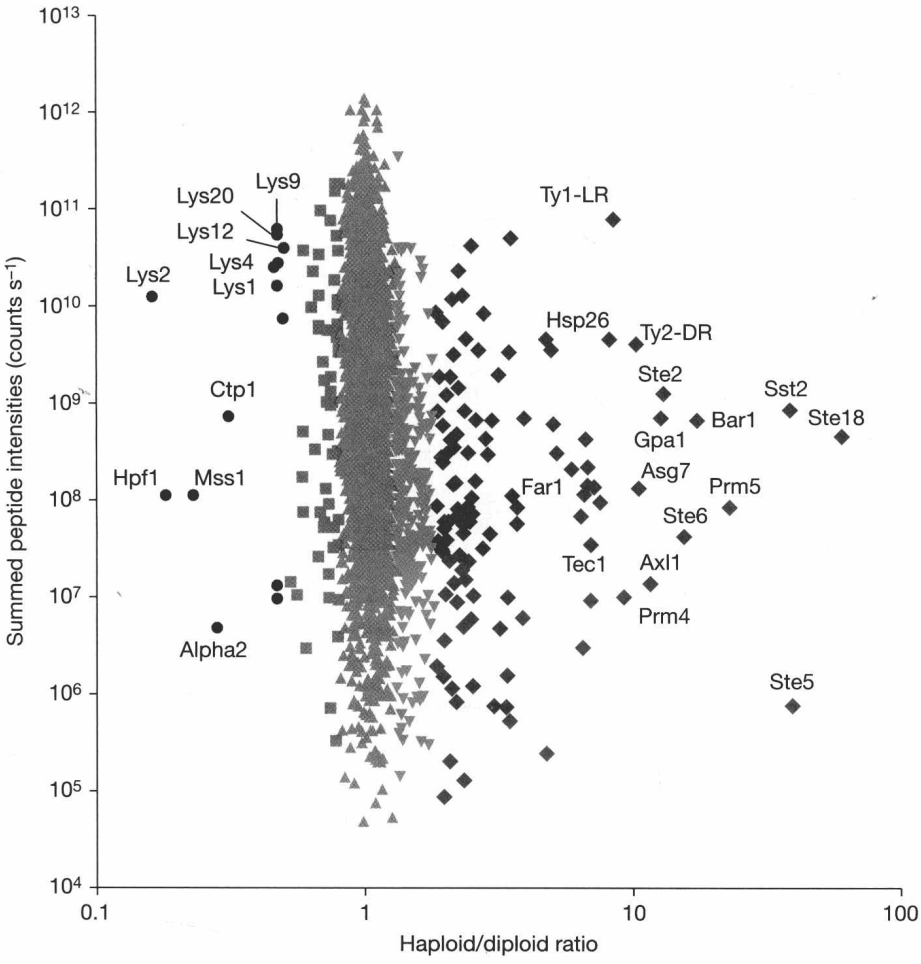


FIGURE 4.4 Two-dimensional DIGE. Overlay image of Cy3- (*green*) and Cy5- (*red*) labeled test-spiked *Erwinia carotovora* proteins. The protein test spikes were three conalbumin isoforms (*arrowheads*) and two myoglobin isoforms (*arrows*). Spots that are of equal intensity between the two channels appear *yellow* in the overlay image. As spike proteins were eight times more abundant in the Cy5 channel, they appear as *red* spots in the overlay. The gel is oriented with the acidic end to the left. (From Lilley KS, Razzaq A & Dupree P (2002) *Curr. Opin. Chem. Biol.* 6, 46. With permission from Elsevier.)



BOX 4.5 FIGURE 2 Quantitative difference between the haploid and diploid yeast proteome (overall fold change). Proteins to the left (becoming deeper *green*) are more strongly represented in haploid cells. Proteins to the right (becoming deeper *red*) are more strongly represented in diploid cells. (From de Godoy LMF, Olsen JV, Cox J et al. (2008) *Nature* 455, 1251–1254. With permission from Macmillan Publishers Ltd.)

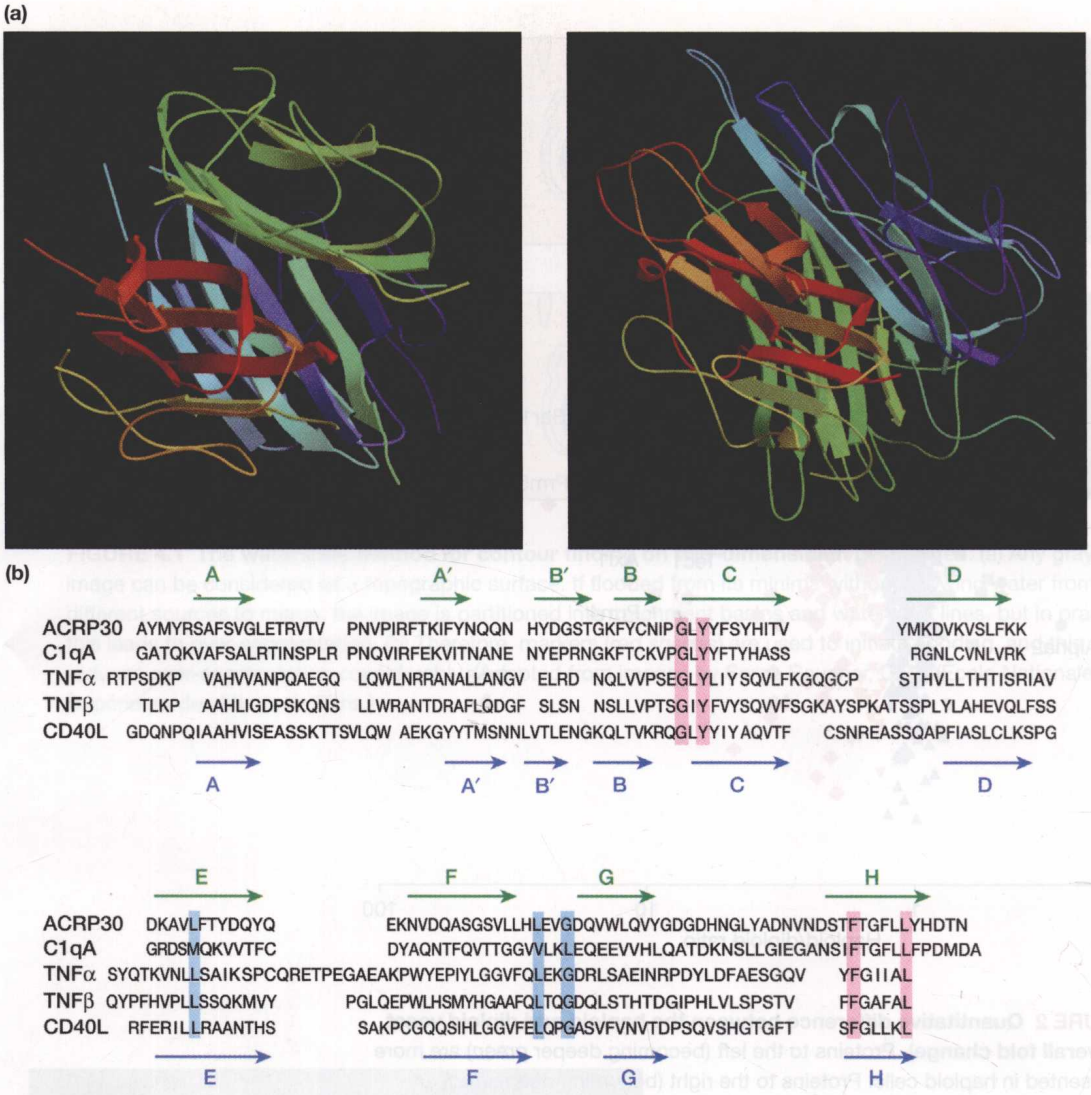


FIGURE 6.1 Identification of related proteins by structural comparison. (a) A ribbon diagram comparison of AdipoQ (left) and TNFα (right). The structural similarity is equivalent to that within the TNF family. (b) Structure-based sequence alignment between several members of the TNF family (CD40L, TNFα, and TNFβ) and two members of the C1q family (C1qA and AdipoQ, the latter labeled ACRP30). Highly conserved residues (present in at least four of the proteins) are shaded, and arrows indicate β-strand regions in the proteins. There is little sequence similarity between AdipoQ and the TNF proteins (for example, 9% identity between AdipoQ and TNFα), so BLAST searches would not identify a relationship. (Adapted from Shapiro L & Harris T (2000) *Curr. Opin. Biotechnol.* 11, 31. With permission from Elsevier. Images courtesy of Protein Data Bank.)

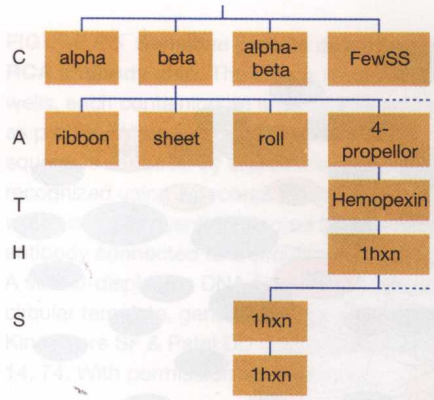
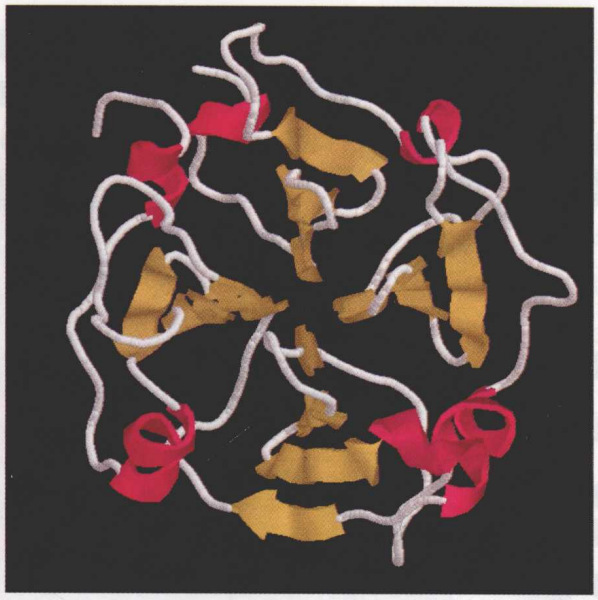


FIGURE 6.8 Structural classification of proteins using the CATH database. The protein shown is hemopexin, a protein rich in β-sheets with few α-helices. (Courtesy of Christine Orengo.)



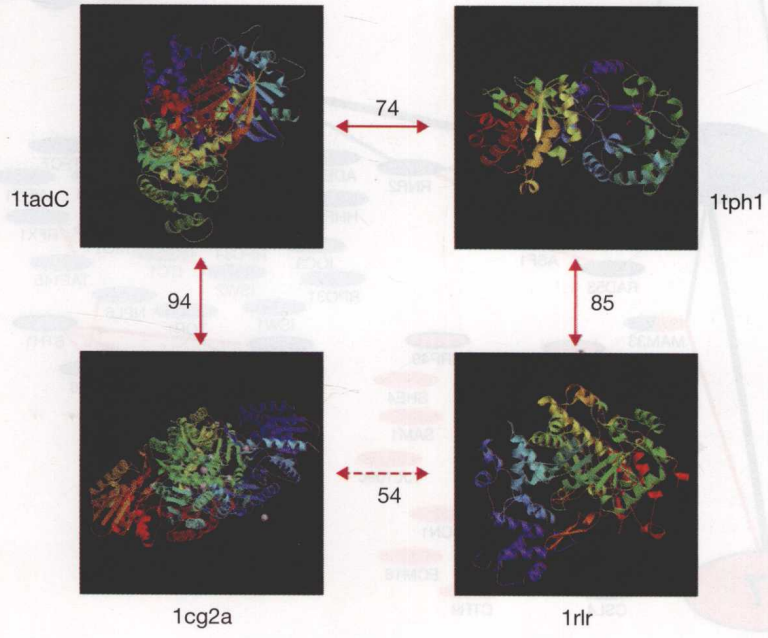


FIGURE 6.9 The Russian doll effect. Four proteins are illustrated that show continuous structural variation over fold space. Each of the proteins shares at least 74 structurally equivalent residues with its nearest neighbor, but the two extreme proteins show only 54 structurally equivalent residues when compared directly. Key: 1cg2a, carboxypeptidase G2; 1tadC, transducin-K; 1tph1, triose phosphate isomerase; 1rlr, ribonucleotide reductase protein R1. (From Domingues FS, Koppensteiner WA & Sippl MJ (2000) *FEBS Lett.* 476, 98. With permission from Elsevier. Images courtesy of Protein Data Bank.)

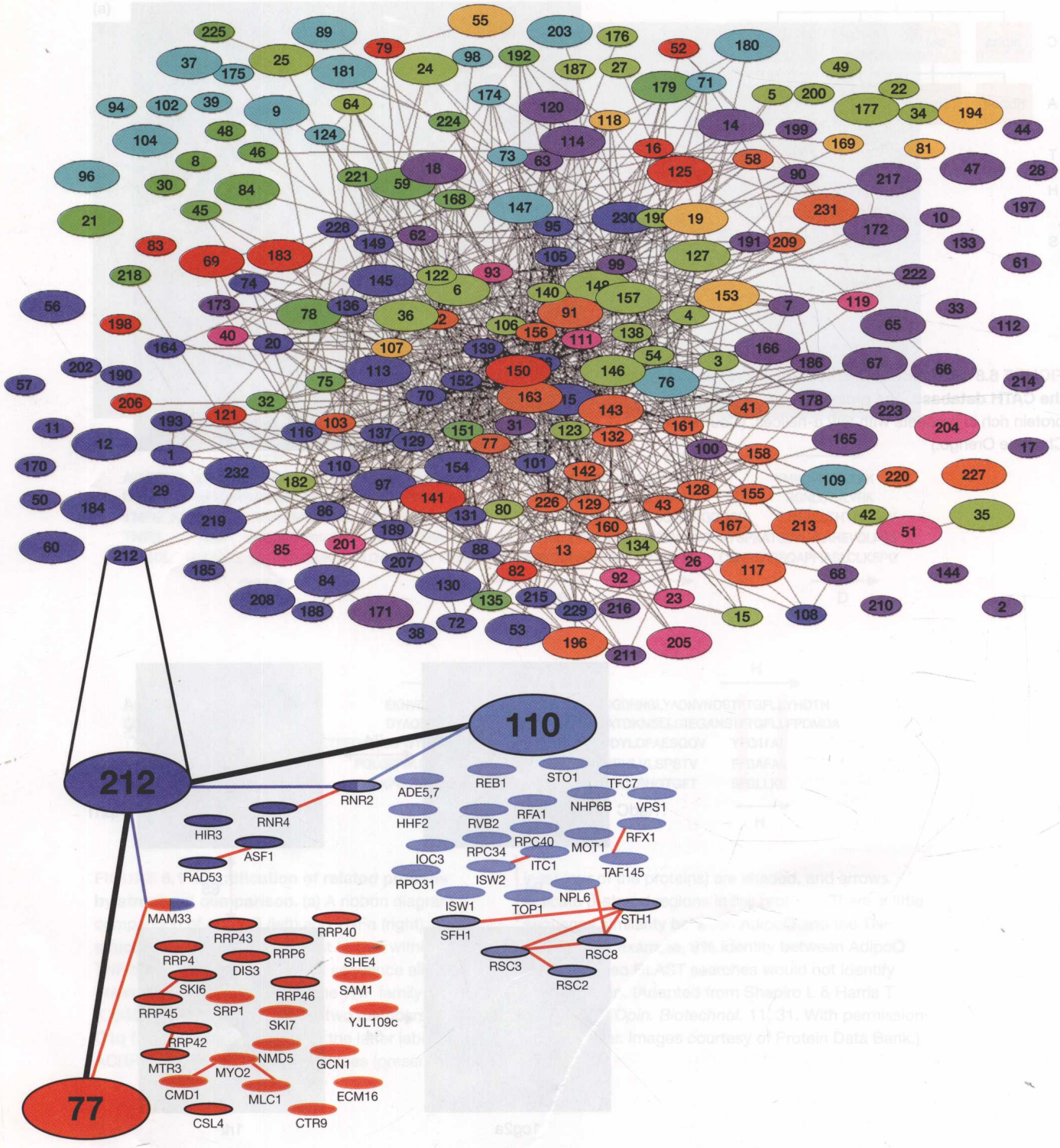


FIGURE 7.19 The protein complex network, and grouping of connected complexes. Links were established between complexes sharing at least one protein. For clarity, proteins found in more than nine complexes were omitted. The graphs were generated automatically by a relaxation algorithm that finds a local minimum in the distribution of nodes by minimizing the distance of connected nodes and maximizing the distance of unconnected nodes. In the upper panel, cellular roles of the individual complexes are color-coded: *red*, cell cycle; *dark green*, signaling; *dark blue*, transcription, DNA maintenance, chromatin structure; *pink*, protein and RNA transport; *orange*, RNA metabolism; *light green*, protein synthesis and turnover; *brown*, cell polarity and structure; *violet*, intermediate and energy metabolism; *light blue*, membrane biogenesis and traffic. The lower panel is an example of a complex (TAP-C212) linked to two other complexes (TAP-C77 and TAP-C110) by shared components. It illustrates the connection between the protein and complex levels of organization. *Red* lines indicate physical interactions as listed in the Yeast Proteome Database. (From Gavin AC, Bösche M, Krause et al. (2002) *Nature* 415, 141. With permission from Macmillan Publishers Ltd.)

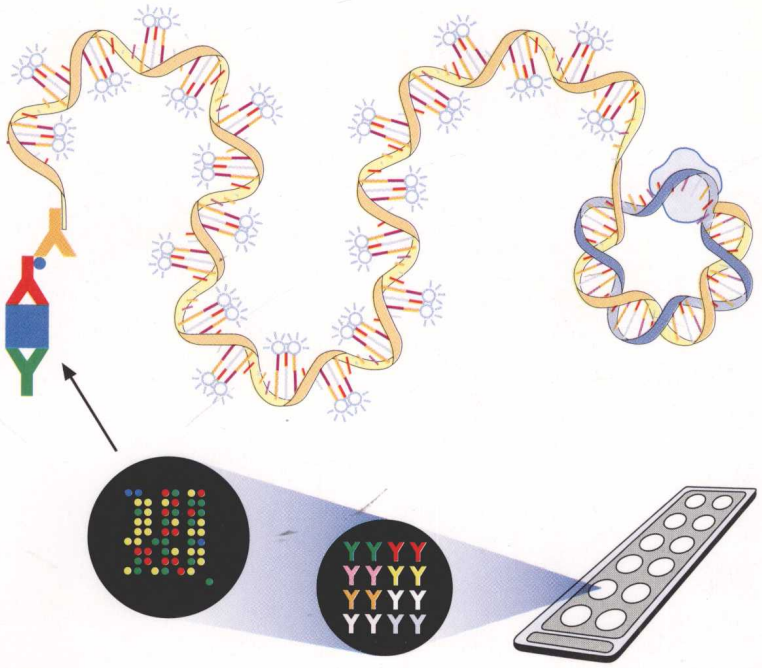


FIGURE 9.5 Sensitive protein detection using the RCA antibody chip. The chip is divided into 16 Teflon wells, each containing an array of 256 antibodies as probes. When a protein, represented by the *blue* square, is captured by one of the probes, it can be recognized using a second, biotinylated antibody (*red*), which is subsequently detected by a tertiary universal antibody connected to a circular oligonucleotide. A strand-displacing DNA polymerase can use this circular template, generating a long concatemer. (From Kingsmore SF & Patel DD (2003) *Curr. Opin. Biotechnol.* 14, 74. With permission from Elsevier.)

Contents

Chapter 1 The origin and scope of proteomics 1

1.1 INTRODUCTION 1

1.2 THE BIRTH OF LARGE-SCALE BIOLOGY AND THE "OMICS" ERA 1

1.3 THE GENOME, TRANSCRIPTOME, PROTEOME, AND METABOLOME 6

1.4 FUNCTIONAL GENOMICS 8

Transcriptomics is the systematic, global analysis of mRNA 8

Large-scale mutagenesis and interference can also determine the functions of genes on a global scale 11

1.5 THE NEED FOR PROTEOMICS 15

1.6 THE SCOPE OF PROTEOMICS 17

Protein identification and quantitation are the most fundamental aspects of proteomic analysis 17

Important functional data can be gained from sequence and structural analysis 18

Interaction proteomics and activity-based proteomics can help to link proteins into functional networks 19

1.7 CURRENT CHALLENGES IN PROTEOMICS 20

Chapter 2 Strategies for protein separation 23

2.1 INTRODUCTION 23

2.2 GENERAL PRINCIPLES OF PROTEIN SEPARATION IN PROTEOMICS 23

2.3 PRINCIPLES OF TWO-DIMENSIONAL GEL ELECTROPHORESIS 25

Electrophoresis separates proteins by mass and charge 25

Isoelectric focusing separates proteins by charge irrespective of mass 26

SDS-PAGE separates proteins by mass irrespective of charge 28

2.4 THE APPLICATION OF 2DGE IN PROTEOMICS 29

The four major advantages of 2DGE are robustness, reproducibility, visualization, and compatibility with downstream microanalysis 29

The four major limitations of 2DGE are resolution, sensitivity, representation, and compatibility with automated protein analysis 30

The resolution of 2DGE can be improved with giant gels, zoom gels, and modified gradients, or by pre-fractionating the sample 30

The sensitivity of 2DGE depends on the visualization of minor protein spots, which can be masked by abundant proteins 31

The representation of hydrophobic proteins is an intractable problem reflecting the buffers required for isoelectric focusing 32

Downstream mass spectrometry requires spot analysis and picking 34

2.5 PRINCIPLES OF MULTIDIMENSIONAL LIQUID CHROMATOGRAPHY 34

Protein and peptide separation by chromatography relies on differing affinity for stationary and mobile phases 34

Affinity chromatography exploits the specific binding characteristics of proteins and/or peptides 36

Size exclusion chromatography sieves molecules on the basis of their size 36

Ion exchange chromatography exploits differences in net charge 37

Reversed-phase chromatography and hydrophobic interaction chromatography exploit the affinity between peptides and hydrophobic resins 38

2.6 MULTIDIMENSIONAL LIQUID CHROMATOGRAPHY STRATEGIES IN PROTEOMICS 39

Multidimensional liquid chromatography is more versatile and more easily automated than 2DGE but lacks a visual dimension 39

The most useful MDLC systems achieve optimal peak capacity by exploiting orthogonal separations that have internally compatible buffers 40

MudPIT shows how MDLC has evolved from a laborious technique to virtually hands-free operation 41

RP-RPLC and HILIC-RP systems offer advantages for the separation of certain types of peptide mixtures 44

Affinity chromatography is combined with MDLC to achieve the simplification of peptide mixtures 44

Chapter 3 Strategies for protein identification 47

3.1 INTRODUCTION 47

3.2 PROTEIN IDENTIFICATION WITH ANTIBODIES 47

3.3 DETERMINATION OF PROTEIN SEQUENCES BY CHEMICAL DEGRADATION 48

Complete hydrolysis allows protein sequences to be inferred from the content of the resulting amino acid pool 48

Edman degradation was the first general method for the *de novo* sequencing of proteins 49

Edman degradation was the first protein identification method to be applied in proteomics, but it is difficult to apply on a large scale 50

3.4 MASS SPECTROMETRY—BASIC PRINCIPLES AND INSTRUMENTATION 52

Mass spectrometry is based on the separation of molecules according to their mass/charge ratio 52

The integration of mass spectrometry into proteomics required the development of soft ionization methods to prevent random fragmentation 52

Controlled fragmentation is used to break peptide bonds and generate fragment ions 53

Five principal types of mass analyzer are commonly used in proteomics 54

3.5 PROTEIN IDENTIFICATION USING DATA FROM MASS SPECTRA 58

Peptide mass fingerprinting correlates e xperimental and theoretical intact peptide masses 58

Shotgun proteomics can be combined with database searches based on uninterpreted spectra 61

MS/MS spectra can be used to derive protein sequences *de novo* 61

Chapter 4 Strategies for protein quantitation 69

4.1 INTRODUCTION 69

4.2 QUANTITATIVE PROTEOMICS BASED ON 2DGE 70

The quantitation of proteins in two-dimensional gels involves the creation of digital data from analog images 70

Spot detection, quantitation, and comparison can be challenging without human intervention 71

4.3 MULTIPLEXED IN-GEL PROTEOMICS 75

Difference in-gel electrophoresis involves the simultaneous separation of comparative protein samples labeled with different fluorophores 75

Parallel analysis with multiple dyes can also be used to identify particular structural or functional groups of proteins 76

4.4 QUANTITATIVE MASS SPECTROMETRY 77

Label-free quantitation may be based on spectral counting or the comparison of signal intensities across samples in a narrow *m/z* range 77

Label-based quantitation involves the incorporation of labels that allow corresponding peptides in different samples to be identified by a specific change in mass 77

ICAT reagents are used for the selective labeling of proteins or peptides 79

Proteins and peptides can also be labeled nonselectively 80

Isobaric tagging allows protein quantitation by the detection of reporter ions 80

Metabolic labeling introduces the label before sample preparation but is limited to simple organisms and cultured cells 83

Chapter 5 The analysis of protein sequences 87

5.1 INTRODUCTION 87

5.2 PROTEIN FAMILIES AND EVOLUTIONARY RELATIONSHIPS 89

Evolutionary relationships between proteins are based on homology 89

The function of a protein can often be predicted from its sequence 92

5.3 PRINCIPLES OF PROTEIN SEQUENCE COMPARISON 93

Protein sequences can be compared in terms of identity and similarity 93

Homologous sequences are found by pairwise similarity searching 93

Substitution score matrices rank the importance of different substitutions 96

Sequence alignment scores depend on sequence length 98

Multiple alignments provide more information about key sequence elements 98

5.4 STRATEGIES TO FIND MORE DISTANT RELATIONSHIPS 100

PSI-BLAST uses sequence profiles to carry out iterative searches 100

Pattern recognition methods incorporate conserved sequence signatures 101
5.5 THE RISK OF FALSE-POSITIVE ANNOTATIONS 104

Chapter 6 The analysis of protein structures 107
6.1 INTRODUCTION 107
6.2 STRUCTURAL GENOMICS AND STRUCTURE SPACE 110
Coverage of structure space is currently uneven 110
Structure and function are not always related 113
6.3 TECHNIQUES FOR SOLVING PROTEIN STRUCTURES 114
X-ray diffraction requires well-ordered protein crystals 114
NMR spectroscopy exploits the magnetic properties of certain atomic nuclei 116
Additional methods for structural analysis mainly provide supporting data 118
6.4 PROTEIN STRUCTURE PREDICTION 119
Structural predictions can bridge the gap between sequence and structure 119
Protein secondary structures can be predicted from sequence data 120
Tertiary structures can be predicted by comparative modeling if a template structure is available 122
Ab initio prediction methods attempt to construct structures from first principles 123
Fold recognition (threading) is based on similarities between nonhomologous folds 123
6.5 COMPARISON OF PROTEIN STRUCTURES 124
6.6 STRUCTURAL CLASSIFICATION OF PROTEINS 125
6.7 GLOBAL STRUCTURAL GENOMICS INITIATIVES 126

Chapte 7 Interaction proteomics 131
7.1 INTRODUCTION 131
7.2 METHODS TO STUDY PROTEIN–PROTEIN INTERACTIONS 134
Genetic methods suggest interactions from the combined effects of two mutations in the same cell or organism 134
Protein interactions can be suggested by comparative genomics and homology transfer 135
Affinity-based biochemical methods provide direct evidence that proteins can interact 138
Interactions between proteins *in vitro* and *in vivo* can be established by resonance energy transfer 142
Surface plasmon resonance can indicate the mass of interacting proteins 142
7.3 LIBRARY-BASED METHODS FOR THE GLOBAL ANALYSIS OF BINARY INTERACTIONS 143
7.4 TWO-HYBRID/PROTEIN COMPLEMENTATION ASSAYS 145
The yeast two-hybrid system works by assembling a transcription factor from two inactive fusion proteins 145
Several large-scale interaction screens have been carried out using different yeast two-hybrid screening strategies 146
Conventional yeast two-hybrid screens have a significant error rate 148
7.5 MODIFIED TWO-HYBRID SYSTEMS FOR MEMBRANE, CYTOSOLIC, AND EXTRACELLULAR PROTEINS 149
7.6 BACTERIAL AND MAMMALIAN TWO-HYBRID SYSTEMS 150
7.7 LUMIER AND MAPPIT HIGH-THROUGHPUT TWO-HYBRID PLATFORMS 151
7.8 ADAPTED HYBRID ASSAYS FOR DIFFERENT TYPES OF INTERACTIONS 152
7.9 SYSTEMATIC COMPLEX ANALYSIS BY TANDEM AFFINITY PURIFICATION–MASS SPECTROMETRY 153
7.10 ANALYSIS OF PROTEIN INTERACTION DATA 155
7.11 PROTEIN INTERACTION MAPS 156
7.12 PROTEIN INTERACTIONS WITH SMALL MOLECULES 158

Chapter 8 Protein modification in proteomics 165
8.1 INTRODUCTION 165
8.2 METHODS FOR THE DETECTION OF POST-TRANSLATIONAL MODIFICATIONS 167
8.3 ENRICHMENT STRATEGIES FOR MODIFIED PROTEINS AND PEPTIDES 168
8.4 PHOSPHOPROTEOMICS 170
Protein phosphorylation is a key regulatory mechanism 170
Separated phosphoproteins can be detected with specific staining reagents 172