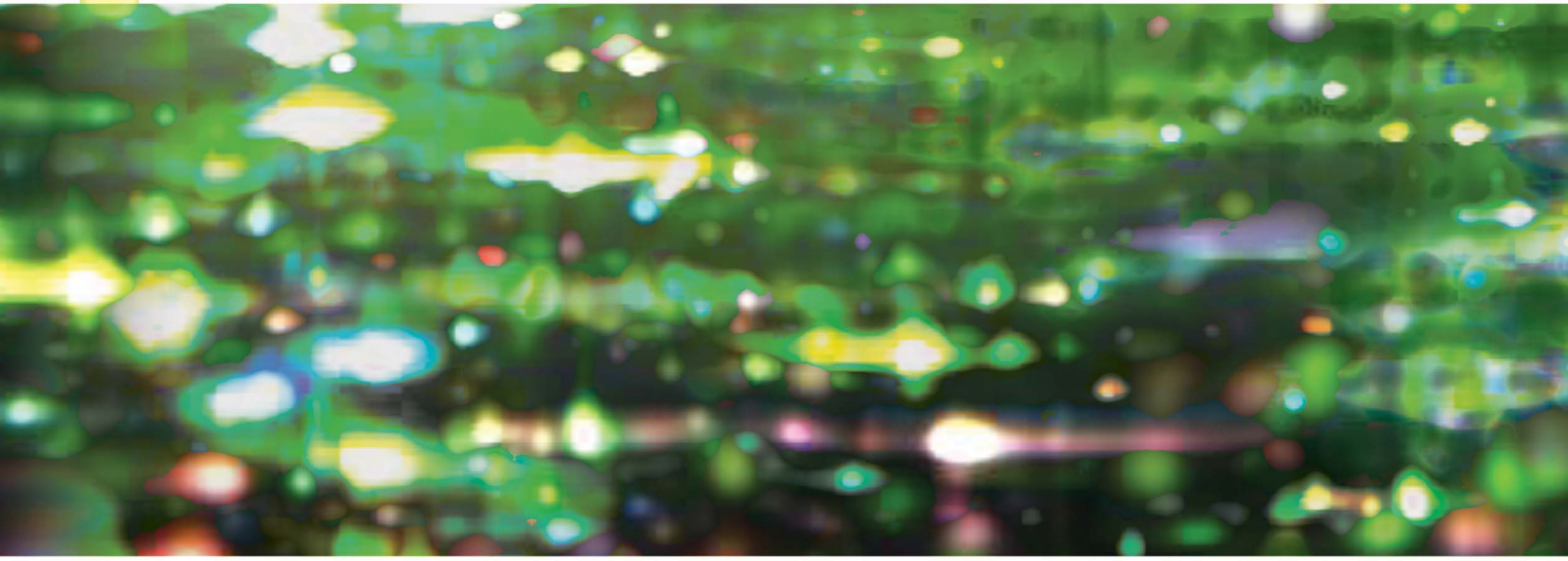
GS
Garland Science

Richard M. Twyman

Principles of PROTEOMICS



SECOND EDITION

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

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

Contents

CHAPTER 1

The origin and scope of proteomics

1.1 INTRODUCTION

Proteomics is the systematic, large-scale analysis of proteins. It is based on the concept of the **proteome** as a complete set of proteins produced by a given cell, tissue, or organism, either as a complete protein catalog or as a list of proteins produced under a defined set of conditions. Proteins are involved in almost every conceivable biological activity, so a comprehensive analysis of the proteins in the cell provides a unique global perspective showing how these molecules interact and cooperate to create and maintain a working biological system. The cell responds to internal and external changes by regulating the level and activity of its proteins, so changes in the proteome, either qualitative or quantitative, provide a snapshot of the cell in action. The proteome is a complex and dynamic entity that can be defined in terms of the sequence, structure, abundance, stability, localization, modification, interaction, and biochemical function of its components, providing a rich and varied source of data. The analysis of these various properties of the proteome requires an equally diverse range of technologies, which are the subject of this book.

This introductory chapter considers the importance of proteomics in the context of large-scale biology, discusses some of the major goals of proteomics, and introduces the major technology platforms. We begin by tracing the origins of proteomics in the genomics revolution of the 1990s and following its evolution from a concept to a mainstream technology with a global market value that is predicted to exceed $6 billion by 2015.

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

The overall goal of molecular biology is to determine the functions of genes and their products. This allows them to be linked into pathways and networks that should ultimately lead to a detailed understanding of how biological systems work. Until the turn of the millennium, molecular biology research focused predominantly on the isolation and characterization of individual genes and proteins because there was neither the information nor the technology available for investigations on a global scale. The only way to study biological systems was to break them down into their components, look at these individually, and then attempt to deduce how the system worked as a whole by proposing hypotheses that could be tested in further experiments. This is known as the **reductionist approach**.

The face of biological research began to change in the 1990s as technological breakthroughs made it possible to carry out **large-scale DNA** (deoxyribonucleic acid) **sequencing**. Until this point, the sequences of individual genes

and proteins had accumulated slowly and steadily as researchers cataloged individual discoveries. This can be seen from the steady growth in the **INSDC nucleotide sequence databases** from 1980 to 1990, when the total amount of stored sequence data reached 10 million base pairs (**Figure 1.1**). During this time, almost all DNA sequencing was performed manually using the **Sanger chain termination method** (**Box 1.1**). The 1990s saw the advent of automated DNA sequencing, which allowed sequence data to be gathered at an increasing rate and ensured that the databases grew exponentially well into the 2000s. In the early 1990s, much of the new sequence data was represented by expressed sequence tags (ESTs), which are short fragments of DNA obtained by the random sequencing of cDNA (complementary DNA) libraries. In 1995, the first complete cellular genome sequence was published, that of the bacterium *Haemophilus influenzae*. This represented a new paradigm in molecular biology because for the first time the data existed to characterize a complete biological system. Over the next few years, more than 100 further genome sequences were completed, including the human genome, which was essentially finished in 2003. A lot of the data added to the databases after this point was in the form of random genomic clones resulting from **whole-genome shotgun** projects, basically massive collections of sequences covering the entire genome, which were then assembled into contigs using powerful computers. The rate of sequence data accumulation continued to increase in the 2000s, mainly because the throughput of automated Sanger sequencing continued to increase despite the inherent limitations of the underlying technology (**Box 1.2**). This involved the development of capillary sequencing machines that could carry out large numbers of automated reactions in parallel, day and night. To cope with this influx of data, two of the INSDC partners collaborated to launch the **Trace Archive** in 2001, to collect raw data produced at sequencing centers around the world. The amount of data in the archive doubled every 10 months between 2001 and 2006.

In 2005, there was another paradigm shift when the first **next-generation sequencing** methods began to displace the Sanger technique. Several next-generation sequencing technologies now exist based on different underlying principles, but they are united by their ability to yield millions of short DNA sequences in parallel (**Box 1.3**). To give some insight into the pace of change,

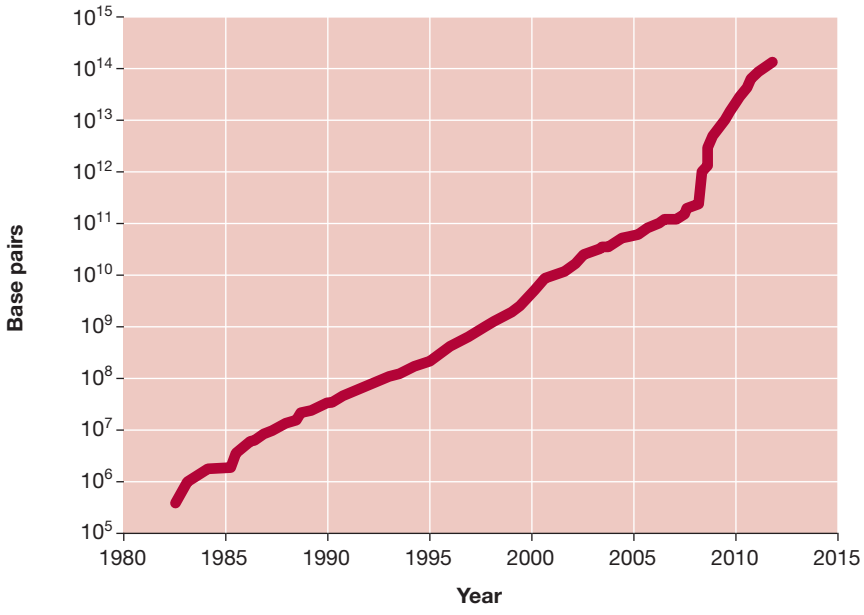


FIGURE 1.1 Cumulative base pairs in the INSDC over time, excluding the Trace Archive. The International Nucleotide Sequence Database Collaboration is a collaborative relationship between the three primary nucleotide sequence databases, that is, GenBank, the European Nucleotide Archive (ENA), and the DNA Data Bank of Japan (DDBJ). This collaboration involves the daily exchange and synchronization of sequence data and the provision of a comprehensive publically accessible nucleotide sequence data resource. (From Karsch-Mizrachi I, Nakamura Y & Cochrane G (2012) International Nucleotide Sequence Database Collaboration, *Nucleic Acids Res.* 40, D33. With permission from Oxford University Press.)

the original Human Genome Project took a decade of work involving a huge consortium of researchers using Sanger sequencing and billions of dollars of funding, but with the advent of next-generation sequencing there are now companies that offer to sequence individual human genomes for less than $100,000 in a few weeks. The first phase of the 1000 Genomes Project was completed in 2010 and this aims to sequence at least 1000 different human genomes every two years, which represents an output of 10 billion bases (three complete human genomes) every 24 hours. The three members of the INSDC began to archive raw next-generation sequence data reads between 2007 and 2008, and in 2009 launched the collaborative **Sequence Read Archive** (**SRA**) to accommodate the additional data, accounting for

BOX 1.1 RELATED TECHNOLOGIES.
Sanger's method for DNA sequencing.

Frederick Sanger's chain termination method for DNA sequencing (also known as the dideoxy method) exploits the ability of DNA polymerases to synthesize complementary strands of varying lengths on a single-stranded DNA template when provided with a short, labeled primer and a mixture of all four standard 2′-deoxynucleoside triphosphates (dNTPs) plus a small amount of a specific, chain-terminating analog (a 2′, 3′-dideoxynucleoside triphosphate, or ddNTP). The normal substrates for DNA synthesis are the four dNTPs representing the nucleosides adenosine (A), cytidine (C), guanosine (G), and thymidine (T). These possess a hydroxyl group at the C3′ position allowing the formation of a phosphodiester bond with the next nucleotide incorporated into the DNA strand. The corresponding ddNTPs lack this hydroxyl group and the strand cannot be elongated once a ddNTP is incorporated, that is, it acts as a chain terminator.

The original Sanger method comprises four parallel reactions, each incorporating the components required for DNA synthesis (the template, a radiolabeled primer, DNA polymerase, and four dNTPs) plus small amount of one of the four corresponding ddNTPs. In each reaction, the ddNTP is incorporated randomly when the template exposes the complementary base, generating a population of DNA molecules with a common 5′ end corresponding to the primer, but a variable 3′ end always representing the same base depending on which analog has been included. The four reaction products are denatured and separated in adjacent lanes by polyacrylamide gel electrophoresis, which has sufficient resolution to separate DNA molecules differing in length by one base. Exposure of the dried gel to X-ray film reveals a ladder of bands, which can be used to read off the sequence.

BOX 1.2 BACKGROUND ELEMENTS.
The limitations of Sanger sequencing.

The Sanger chain termination method for DNA sequencing dominated molecular biology for approximately 25 years (1980–2005). During this time, the throughput of the method increased substantially through cumulative technological improvements, including modifications allowing the use of double-stranded DNA templates, better enzymes, and more sensitive labels. However, the most significant improvement was achieved by switching from the use of radiolabeled primers to the use of four ddNTPs labeled with different fluorophores. This allowed the four reactions to be separated in a single gel lane (because the four sets of products produce different signals) and allowed the sequence to be read automatically by detecting fluorescence in real time during electrophoresis (rather than several days later by autoradiography). This not only increased throughput but also improved sequencing accuracy by reducing the human role in sequence interpretation and providing sufficient capacity to allow both DNA strands to be sequenced a number of times. Even higher throughput was achieved by replacing slab gels (which are labor-intensive and time-consuming) with capillary electrophoresis, which is up to five times faster at separation, reduces artifacts, and involves minimal operator handling. Capillary electrophoresis runs that handle up to 384 reactions simultaneously were the basis of factory-style Sanger sequencing programs that yielded up to 1 million base pairs of sequence each day. The resulting capillary trace data were processed automatically and subjected to rigorous quality control to yield high-quality datasets.

Even so, there are three irreconcilable bottlenecks in Sanger sequencing, namely the requirement to prepare template DNA, then carry out the chain termination reaction, and then separate the products. All these processes take time. These limitations have been addressed by today's "next-generation" sequencing methods, as discussed in Box 1.3.

the sudden surge in deposited sequences from 2009 in Figure 1.1, which is remarkable considering that the *y* axis has a logarithmic scale. Or in other words, the rate of sequence data accumulation is more than exponential at the time of writing. Based on these technological advances, the INSDC

BOX 1.3 RELATED TECHNOLOGIES.
Next-generation sequencing.

The so-called "next-generation" sequencing methods were developed to overcome the inherent limitations of the Sanger chain termination method, namely the need for template preparation and the time taken to complete the chain termination reaction and product separation. The bottleneck caused by template preparation was initially addressed by combining DNA sequencing with the polymerase chain reaction (cycle sequencing), which in its most extreme form can sequence uncloned source DNA directly. More recently, this has been superseded by the use of single-molecule templates immobilized either on a solid substrate or within an oil droplet, which can be sequenced directly or amplified *in situ* (emulsion PCR). The bottleneck caused by chain termination and product separation has been addressed by sequencing DNA in real time and increasing the throughput by extensive miniaturization as discussed below. Many of these methods have now been adopted as the basis of RNA profiling as well as DNA sequencing, as discussed in Box 1.4. They all produce short sequence reads (50–100 bp) but in huge amounts, allowing sequences to be assembled by analyzing overlaps and quality to be tested by sequencing the same DNA segment many times.

454 sequencing
This platform is a high-throughput form of pyrosequencing, in which the incorporation of a nucleotide into DNA is recorded in real time by detecting the release of pyrophosphate. As DNA polymerase moves along the template, each of the four nucleoside triphosphates is fed sequentially into the reaction and then removed. When one of the nucleotides is incorporated, the released pyrophosphate is detected as a flash of light. Multiplexing is achieved by constraining individual sequencing reactions onto microbeads where the template has been amplified by emulsion PCR. The beads are channeled into wells on a picotiter plate, which allows between one and two million reactions to be monitored in parallel.

Illumina/Solexa sequencing
This is based on reversible chain termination, that is, a chain-terminating nucleotide analog is incorporated but can then be cleaved and removed so chain extension can resume after a pause, allowing the fluorescent label to be detected. This method is therefore the closest conceptually to the original Sanger method. The Illumina/Solexa platform involves solid-phase *in situ* template amplification on a glass slide followed by sequencing with four-color blocked reversible terminators that are detected by total internal reflection fluorescence (TIRF) imaging using two lasers. A similar platform known as HeliScope uses non-amplified single-molecule templates.

SOLiD sequencing
This platform is based on the detection of ligation products. Sequencing by ligation involves the "interrogation" of a primed, single-stranded DNA template with a short degenerate oligonucleotide probe containing one or two discriminating bases identified by a specific fluorescent label. If the discriminatory bases match the template immediately adjacent to the primer then the oligonucleotide will anneal and can be ligated to the primer. Otherwise, ligation will not be possible and the probe will be washed away. The sequence adjacent to the primer can therefore be determined by fluorescence detection after washing.

DNA nanoball sequencing
This method uses rolling circle replication to amplify small fragments of genomic DNA into DNA nanoballs, which are then characterized using sequencing by ligation. This platform is offered by CompleteGenomics.

HeliScope sequencing
HeliScope sequencing uses DNA fragments with added poly(A) tail adapters attached to the surface of a flow cell prior to extension-based sequencing achieved by the cyclic addition of individual fluorescently labeled nucleotides prior to washing and signal detection by fluorescence imaging.

SMRT sequencing
Single-molecule real-time sequencing is a sequencing-by-synthesis approach using zero-mode wave guides (small wells containing immobilized DNA polymerase) and fluorescently labeled nucleotides in solution. The wells are constructed so that only fluorescence signals at the base of the well can be detected, allowing the detection of detached fluorescent labels as the corresponding nucleotide is incorporated into the DNA strand.

Emerging methods
Several additional sequencing technologies are considered promising but have yet to reach mainstream development because of technical limitations. These include nanopore sequencing in which a single DNA strand is drawn through a narrow portal and the sequence is determined by measuring the variable but base-specific differences in charge across the pore; ion semiconductor sequencing, which is based on the detection of hydrogen ions that are released during polymerization; and sequencing by hybridization. Although the latter has not been developed into a commercial platform, it is a good example of a next-generation technology because it does not rely on DNA synthesis and therefore does not involve the detection of reaction products. It is the only method that provides instant sequence readout capability, albeit only for short sequences at the current time. The basis of sequencing by hybridization is the annealing of a labeled DNA probe (the sequence to be determined) to an oligonucleotide chip containing arrays of every possible oligonucleotide of a certain length (for example, all possible octanucleotides = 65,536 sequences). The probe will only hybridize to complementary octanucleotides, which should allow the sequence to be reconstructed as a series of overlapping complementary eight-nucleotide fragments.

databases surpassed 100 billion base pairs of DNA in 2009 and reached 100 trillion base pairs in 2011. The ability to produce such massive amounts of sequence data with ever decreasing effort and expense means that it is now considered straightforward to sequence an entire genome as a first step toward characterizing an organism.

The large-scale sequencing projects ushered in the **genomics** era, which led in time to the concept of "**omics**" as a term for genomics and its derivatives, as discussed in the following section. This effectively removed the information bottleneck in accessing the genome and brought about the realization that biological systems, although large and very complex, are ultimately finite. In the 1990s, the idea formed that it might be possible to study biological systems in a global or holistic manner if sufficient amounts of data could be collected and analyzed, simply by cataloging and enumerating the components. However, although the technology for genome sequencing had advanced rapidly, the technology for studying the functions of the newly discovered genes lagged far behind. The databases became clogged with anonymous sequences and gene fragments, and the problem was exacerbated by the unexpectedly large number of new genes found even in well-characterized organisms. As an example, consider the yeast *Saccharomyces cerevisiae*, which was thought to be one of the best-characterized model organisms prior to the completion of its genome-sequencing project in 1996. Over 2000 genes had been characterized in traditional experiments and it was thought that genome sequencing would identify at most a few hundred more. Scientists got a shock when they found the yeast genome contained more than 6000 potential genes, nearly a third of which were unrelated to any previously identified sequence (**Figure 1.2**). Even today, nearly a quarter of the predicted open reading frames in the *S. cerevisiae* genome remain either unconfirmed or without functional annotations.

There are several related terms that describe questionable or unconfirmed sequences. A sequence is described as **unconfirmed** or **questionable** when there is only marginal evidence that it represents a gene. It may be short or may lack certain aspects of a gene while possessing others, suggesting it could be a gene remnant or fragment (that is, a pseudogene) even if it shows homology to known genes. On the other hand, an **orphan gene** has been shown to function as a gene (for example, expression may have been demonstrated) but the sequence is unrelated to any other known gene, that is, it is not a member of a known gene family. This precludes functional annotation by homology but not by independent means, so an orphan gene may not necessarily lack a functional annotation. Several related genes may be grouped into an "orphan family", although this is an oxymoron and a novel family designation is preferred. Finally, a **hypothetical protein** is a protein that is predicted to exist based on the existence of a gene sequence, but direct proof at the protein level does not exist. A hypothetical protein may be the product of an unconfirmed sequence, an orphan gene, or a well-known gene family. Hypothetical proteins can often be promoted to extant proteins by using proteogenomics for the analysis of genomes (Chapter 5).

The availability of masses of anonymous sequence data for hundreds of different organisms has precipitated a number of fundamental changes in the way research is conducted in the molecular life sciences. Traditionally, gene function had been studied by moving from phenotype to gene, an approach sometimes called **forward genetics**. An observed mutant phenotype (or in some cases a purified protein) was used as the starting point to map and identify the corresponding gene, and this led to the functional analysis of that gene and its product. The opposite approach, sometimes termed **reverse genetics**, is to take an uncharacterized gene sequence and modify it to see the effect on phenotype. As more uncharacterized sequences have accumulated in databases, the focus of research has shifted from forward to reverse genetics. Similarly, most research prior to 1995 was **hypothesis-driven**, in

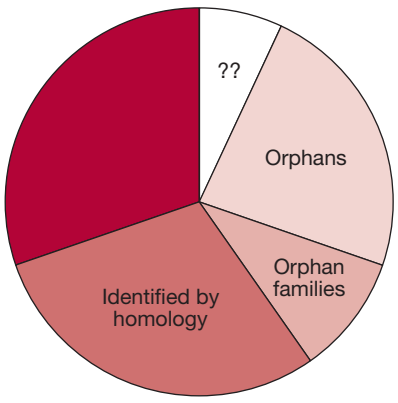


FIGURE 1.2 Distribution of yeast genes by annotation status in the aftermath of the *Saccharomyces cerevisiae* genome project. (?? shows questionable open reading frames)

that the researcher put forward a hypothesis to explain a given observation, and then designed experiments to prove or disprove it. The genomics revolution instigated a progressive change toward what can arguably be called **discovery-driven** research, in which the components of the system under investigation are collected irrespective of any hypothesis about how they might work.

The final paradigm shift concerns the sheer volume of data generated in current experiments. Whereas in the past researchers have focused on individual gene products and generated rather small amounts of data, now the trend is toward the analysis of many genes and their products and the generation of enormous datasets that must be mined for salient information using computers. Advances in genomics have thus forced parallel advances in **bioinformatics**, the computer-aided handling, analysis, extraction, storage, and presentation of biological data.

1.3 THE GENOME, TRANSCRIPTOME, PROTEOME, AND METABOLOME

As large-scale biology has progressively supplanted reductionist experiments, so it has been necessary to re-evaluate the central dogma of molecular biology, which states that a gene is transcribed into RNA (ribonucleic acid) and then translated into protein (**Figure 1.3a**). It has already been necessary to tinker with the dogma to account for new discoveries such as reverse transcription, but large-scale biology has forced a reappraisal of the dogma based on scale. The new paradigm is that the **genome** (all the genes in the organism) gives rise to the **transcriptome** [the complete set of mRNA (messenger RNA) in any given cell], which is then translated to produce the **proteome** (the complete collection of proteins in any given cell) (Figure 1.3b). The proteome is largely responsible for the complete set of chemical compounds found in a cell or organism, which constitutes the **metabolome**. The metabolome is intricately involved in the regulation of the genome, transcriptome, and proteome, thus completing the biological system. By harnessing all this information simultaneously and using it to study and model living organisms, we have now entered the era of systems biology (**Box 1.4**).

The genome differs from the transcriptome and proteome in two important ways. First, the genome has a defined and limited information content because it is a linear sequence of nucleotides. The transcriptome and proteome are much more complex than the genome because a single gene can produce many different mRNAs and proteins. Different transcripts can be generated by alternative splicing, alternative promoter or polyadenylation site usage, and special processing strategies such as RNA editing. Different proteins can be generated by the alternative use of start and stop codons and the proteins synthesized from these mRNAs can be modified in various different ways during or after translation. Some types of modification, such as glycosylation, tend to be permanent. Others, such as phosphorylation, are transient and are often used to regulate protein activity and/or interactions

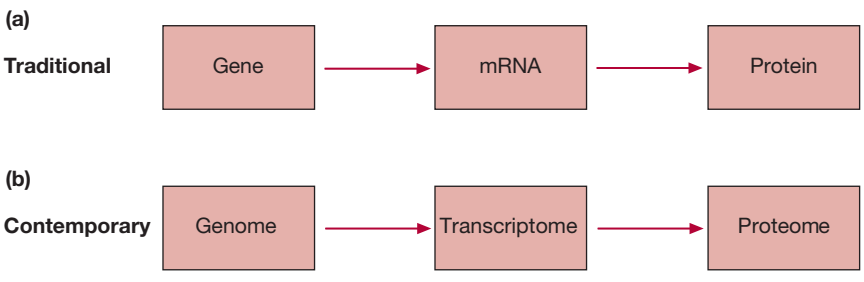


FIGURE 1.3 **The new paradigm in molecular biology—the focus on single genes and their products has been replaced by global analysis.**

BOX 1.4 RELATED TECHNOLOGIES.
Beyond proteomics—metabolomics and systems biology.

Proteomics can be regarded as the global analysis of the final stage of the expression of biological information stored in DNA, resulting in the production of functional molecules—proteins—that carry out a diverse range of activities in the cell (see Box 1.8 for more information on the functions of proteins). However, the end products of cellular processes orchestrated by proteins—for example, in their capacity as enzymes, receptors, transporters, and components of signaling pathways—are the small molecules making up the metabolic profile of the cell, the complete set of which is defined as the metabolome. Metabolomics is thus the global study of metabolites, completing the chain of information from DNA through RNA and protein to the biochemical output of the cell or organism. In many ways, the metabolome can be regarded as an even better snapshot of the functioning cell or organism than the proteome, because it provides an instant readout of physiological status in real time, hence the widespread use of specific metabolites as diagnostic markers.

Like the proteome, the metabolome is dynamic and full of diverse structures that are impossible to analyze with any single method. Depending on the properties of different classes of metabolites, they may be separated by gas chromatography (for volatiles), HPLC (high-performance liquid chromatography; for nonvolatiles), and capillary electrophoresis (for charged molecules) (also see Chapter 2). These techniques can be coupled to various forms of mass spectrometry (discussed in detail in Chapter 3) for detection and identification according to their fragmentation patterns or to nuclear magnetic resonance spectroscopy (Chapter 6).

with other molecules. The same protein can be modified in many different ways, giving rise to innumerable variants. For example, about 70% of human proteins have the potential to be glycosylated and the glycan chains can have many different structures. Often there are several glycosylation sites on the same protein, and different glycan chains can be added to each site. The largest known number of glycosylation sites on a single polypeptide is greater than 20, giving the potential for millions of possible glycoforms. Over 400 different types of post-translational modification have been documented, creating a massive source of proteome diversity. Therefore, although it is estimated that the human genome contains between 20,000 and 25,000 genes, it is likely that the proteome catalog comprises more than a million proteins when post-translational modification is taken into account. The human gene number was initially estimated at 50,000–100,000 based on EST data. This number has been progressively revised downwards following the sequencing and annotation of the human genome, but even with all this information to hand there is still no precise answer. Part of the problem is that different approaches to defining genes give different answers. For example, Ensembl release 67.37 indicates there are 20,115 genes whereas UniProt defines 20,231 genes (see also Box 5.1). Only by increasing diversity at the transcriptome and proteome levels, can the biological complexity of humans be explained compared with nematodes (~18,000 genes), fruit flies (~12,000 genes), and yeast (~6000 genes).

The other major difference between the genome and the transcriptome and proteome is that the genome is a static information resource that, with few exceptions, remains the same regardless of cell type or environmental conditions. In contrast, both the transcriptome and proteome are dynamic entities, whose content can fluctuate dramatically under different conditions due to the regulation of transcription, RNA processing, RNA stability, protein synthesis, protein modification, and protein stability. The transcriptome and proteome vary qualitatively (the type of mRNAs and proteins that are present) and also quantitatively (the levels of different mRNAs and proteins fluctuate over time and in response to internal and external stimuli). Again, much of the increase in biological complexity between simple organisms, such as yeast, and complex organisms, such as mammals and higher plants, is generated at the levels of the transcriptome and proteome.

1.4 FUNCTIONAL GENOMICS

The complete genome sequences that are now available for a large number of important organisms provide potential access to every single gene and therefore pave the way for large-scale functional analysis, an approach known as **functional genomics**. However, even complete gene catalogs provide at best a list of components, and no more explain how a biological system works than a list of mechanical parts explains how to drive a car. Before we can begin to understand how these components build a bacterial cell, a mouse, an apple tree, or a human being, we must understand not only what they do as individual entities, but also how they interact and cooperate with each other. Because the genome is only a blueprint, functional relationships among genes can only be inferred. Direct evidence must be gathered by studying the behavior of gene products at the levels of the transcriptome and proteome. The need for such analysis has encouraged the development of novel technologies that allow large numbers of mRNA and protein molecules to be studied simultaneously.

Transcriptomics is the systematic, global analysis of mRNA

Because the genomics revolution was based on technological advances in large-scale DNA cloning and sequencing, it made good sense to put these technologies to work in the functional analysis of genes. The first functional genomics methods were therefore based on DNA sequencing, and were used to study mRNA expression profiles on a global scale. This gave rise to the field now known as transcriptomics. The expression profile of a gene can reveal much about its role in the cell and can also help to identify functional links to other genes. For example, the expression of many genes is restricted to specific cells or developing structures suggesting that the genes have particular functions in those places (such as insulin, which is expressed solely in pancreatic β-cells). Other genes are expressed in response to external stimuli. For example, they might be switched on or switched off in cells exposed to endogenous signals such as growth factors, or environmental molecules such as DNA-damaging chemicals. Genes with similar expression profiles are likely to be involved in similar processes, and demonstrating that an uncharacterized gene has a similar expression profile to a gene whose function is already known may allow the first gene to be functionally annotated on the basis of "guilt by association." Furthermore, mutating one gene may affect the expression profiles of others, helping to link those genes into functional pathways and networks.

The first transcriptomics technologies were based on a concept now known as **census sequencing**, which refers to the collection and counting of short representative cDNA sequences (**tags**) that are sufficient to identify the corresponding mRNAs. The number of times a given sequence appears is indicative of the relative abundance of that mRNA in the source tissue. In the original method, clones were randomly picked from cDNA libraries and 200–300 bp sequences known as **expressed sequence tags** (**ESTs**) were generated using Sanger's chain termination method. This was an expensive and laborious way to compare mRNA levels within a given sample and it was difficult to compare mRNA levels between samples without carefully prepared comparable cDNA libraries. Alternative methods were therefore devised involving either the rapid quantitative representation of mRNA abundance using techniques such as **differential display PCR** (polymerase chain reaction) or the acquisition of very short **sequence tags** (9–15 bp), many of which could be analyzed at the same time, for example, **serial analysis of gene expression** (**SAGE**) and **massively parallel signature sequencing** (**MPSS**). These tag-based techniques were more reliable than large-scale cDNA sequencing but were complex to realize. The advent of next-generation

sequencing methods (Box 1.3) has made it possible to collect millions of longer sequence tags (~50 bp) rapidly and inexpensively, rendering techniques such as SAGE largely redundant. These new methods (collectively known as **RNA-Seq**) are widely used today. The principles of census sequencing techniques are outlined briefly in **Box 1.5**.

The major alternative transcriptomics technology is based on DNA microarrays, which are miniature devices onto which many different DNA sequences are immobilized in the form of a grid. There are two major types, one made by the mechanical spotting of DNA molecules onto a coated glass slide and one produced by *in situ* oligonucleotide synthesis (the latter are also known as **oligonucleotide chips**). Although manufactured in completely different ways, the principles of mRNA analysis are much the same for each device. Expression analysis is based on **multiplex hybridization** using a complex population of labeled DNA or RNA molecules (**Figure 1.4**

BOX 1.5 RELATED TECHNOLOGIES.
Sequence sampling and display techniques for the global analysis of gene expression.

Sampling of cDNA libraries
Randomly picked clones are sequenced and searched against databases to identify the corresponding genes. The frequency with which each sequence is represented provides a rough guide to the relative abundances of different mRNAs in the original sample. This is an expensive and labor-intensive approach, particularly if several cDNA libraries need to be compared.

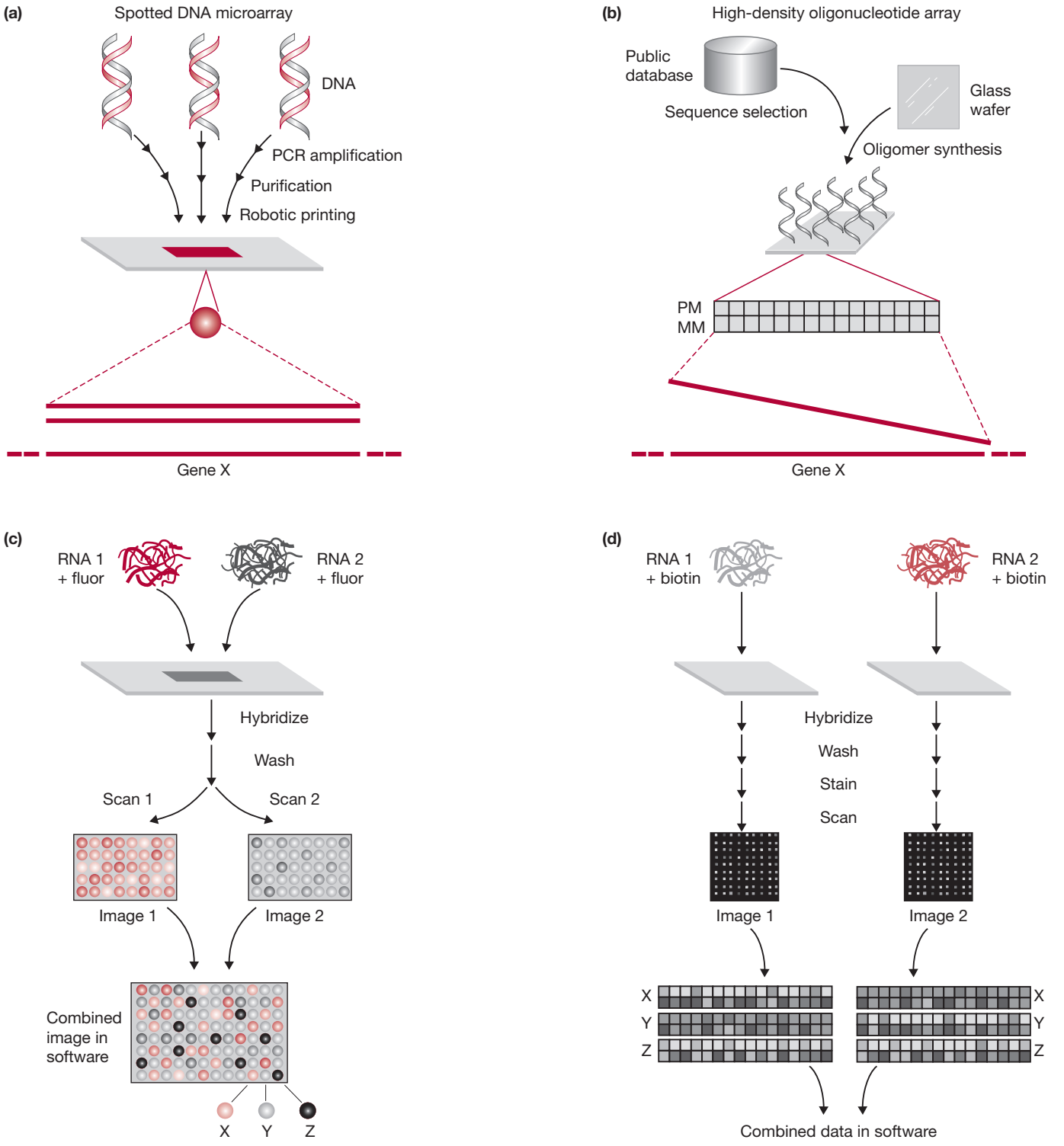
Analysis of EST databases
Expressed sequence tags are signatures generated by the single-pass sequencing of random cDNA clones. If EST data are available for a given library, the abundance of different transcripts can be estimated by determining the representation of each sequence in the database. This is a rapid approach, advantageous because it can be carried out entirely *in silico*, but it relies on the availability of EST data for relevant samples.

Differential display PCR
This is a display method that was devised for the rapid identification of cDNA sequences that are differentially expressed across two or more samples. The method has insufficient resolution to cope with the entire transcriptome in one experiment, so populations of labeled cDNA fragments are generated by RT-PCR (reverse transcriptase polymerase chain reaction) using one oligo-dT primer and one arbitrary primer, producing pools of cDNA fragments representing subfractions of the transcriptome. The equivalent amplification products from two biological samples (that is, products amplified using the same primer combination) are then run side by side on a sequencing gel, and differentially expressed cDNAs are revealed by quantitative differences in band intensities. This technique homes in on differentially expressed genes but false positives are common and other methods must be used to confirm the predicted expression profiles.

Serial analysis of gene expression (SAGE)
In this technique, very short sequence tags representing many cDNAs are joined together in a concatemer, which is sequenced. The tags may be as short as 9–15 bp but this is still adequate to resolve individual cDNA sequences, allowing them to be counted. The method is complex but it essentially involves cleaving a cDNA population with a frequent-cutter restriction enzyme and capturing the poly(A) tail and short exposed fragment. Ligation to a linker containing the recognition site for a type IIS restriction enzyme (which cuts a few base pairs downstream) then generates a sequence tag of defined length. Pairs of linker tags are ligated and the linkers are used as primer annealing sites to amplify the paired tags by PCR. The linkers are then released and the paired tags ligated to form large concatemers for sequencing and counting. SAGE is much more efficient than standard cDNA sampling because 50–100 tags can be counted for each sequencing reaction.

Massively parallel signature sequencing (MPSS)
Like SAGE, the MPSS technique involves the collection of short sequence tags from many cDNAs. However, unlike SAGE (where individual tags are cloned in series for identification by conventional sequencing), MPSS relies on the parallel analysis of thousands of cDNAs attached to microbeads in a flow cell by progressive sequence decoding. As with SAGE, a type IIS restriction enzyme is used to generate the sequence data, but whereas only a single SAGE tag is produced for each cDNA, in MPSS the enzyme is used to expose sequential four-base overhangs on each cDNA, which are "decoded" by a matching adaptor oligonucleotides identified by specific fluorophores. The cDNAs are progressively digested and decoded in four-nucleotide "bites."

RNA-Seq
Sequence sampling comes full circle with RNA-Seq, which is basically the same as the original cDNA sampling method except here the mRNA is reverse-transcribed from source and the resulting cDNA population is randomly sequenced "deeply," that is, millions of short sequences are obtained, using the next-generation sequencing methods described in Box 1.3. This provides statistically highly reliable data about the relative abundance of different mRNA species in a sample and is suitable for direct comparisons between samples.



and color plates). For both devices, a population of mRNA molecules from a particular source is reverse-transcribed en masse to form a representative complex cDNA population. In the case of spotted microarrays, a fluorophore-conjugated nucleotide is included in the reaction mix so that the cDNA population is universally labeled. In the case of oligonucleotide chips, the unlabeled cDNA is converted into a labeled cRNA (complementary

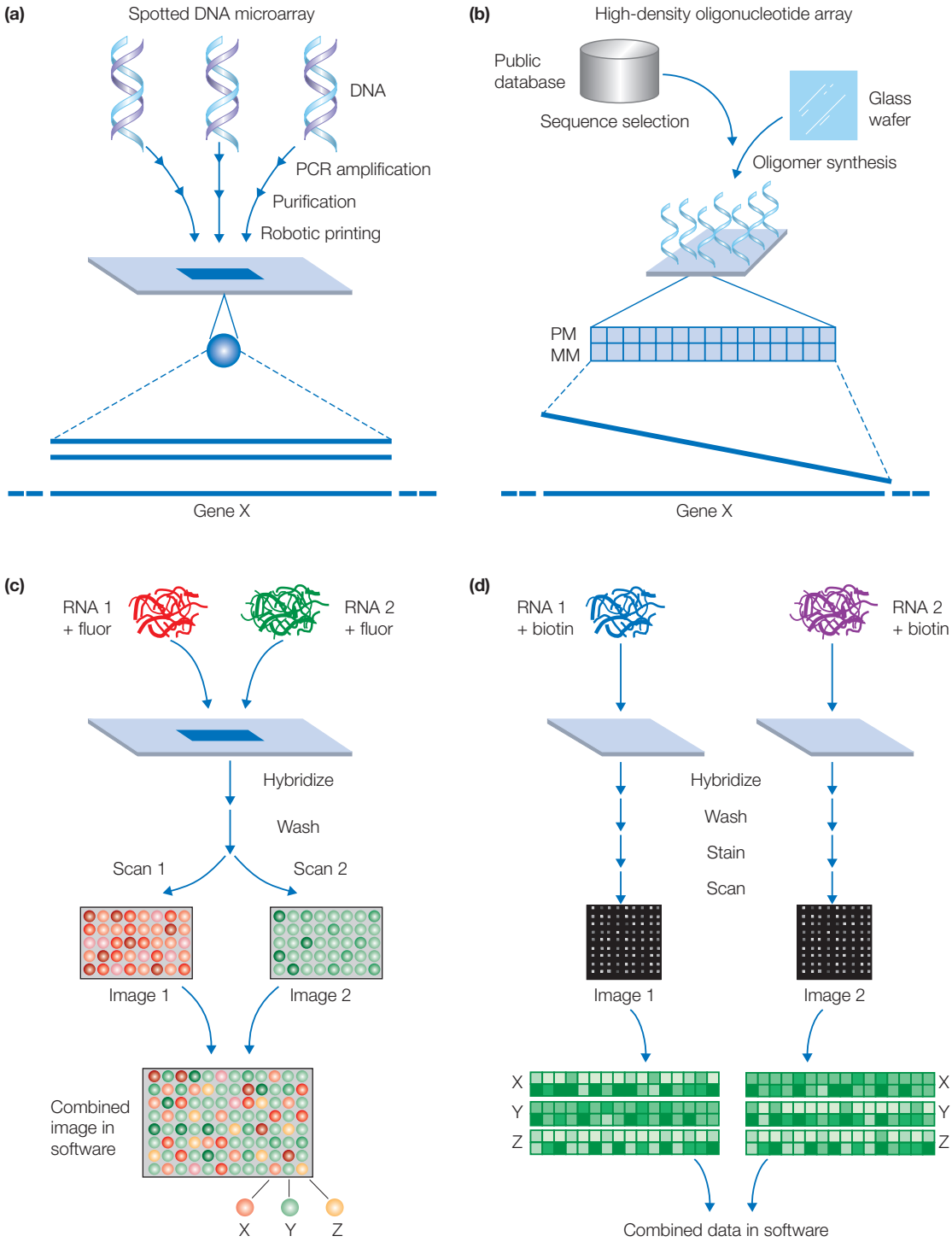


FIGURE 1.4 Expression analysis with DNA microarrays. (a) Spotted microarrays are produced by the robotic printing of amplified cDNA molecules onto glass slides. Each spot or feature corresponds to a contiguous gene fragment of several hundred base pairs or more. (b) High-density oligonucleotide chips 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.)

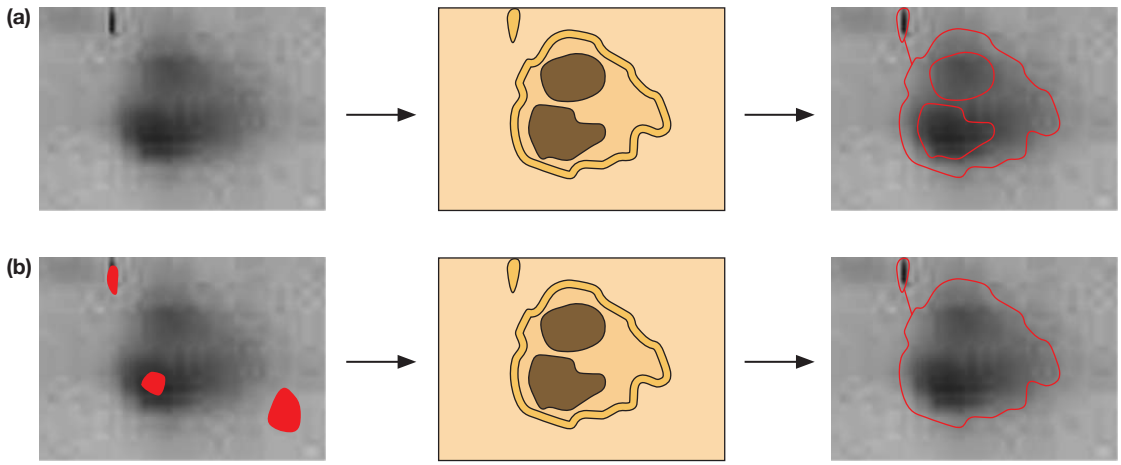


FIGURE 4.1 The watershed method for contour finding on two-dimensional gel images. (a) Any grayscale image can be considered as a topographic surface. If flooded from its minima without allowing water from different sources to merge, the image is partitioned into catchment basins and watershed lines, 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.)

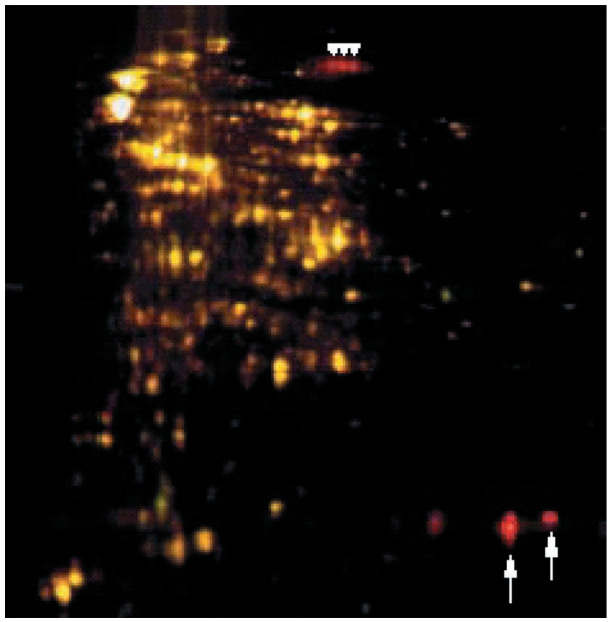
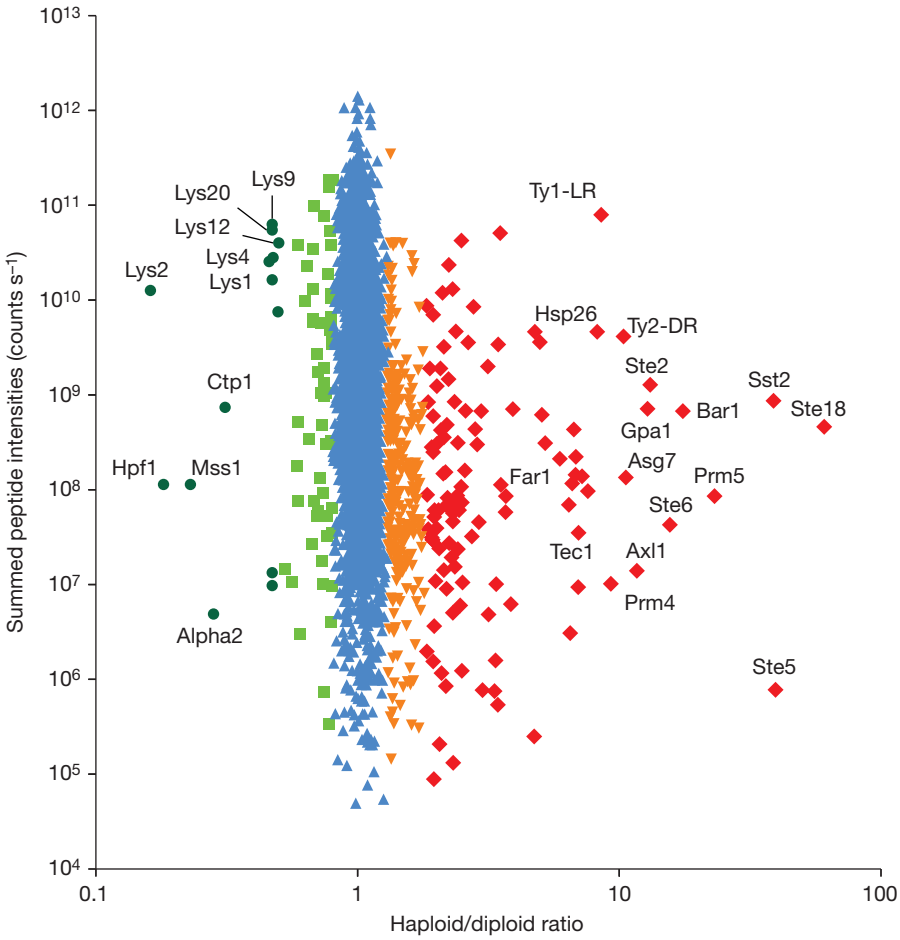
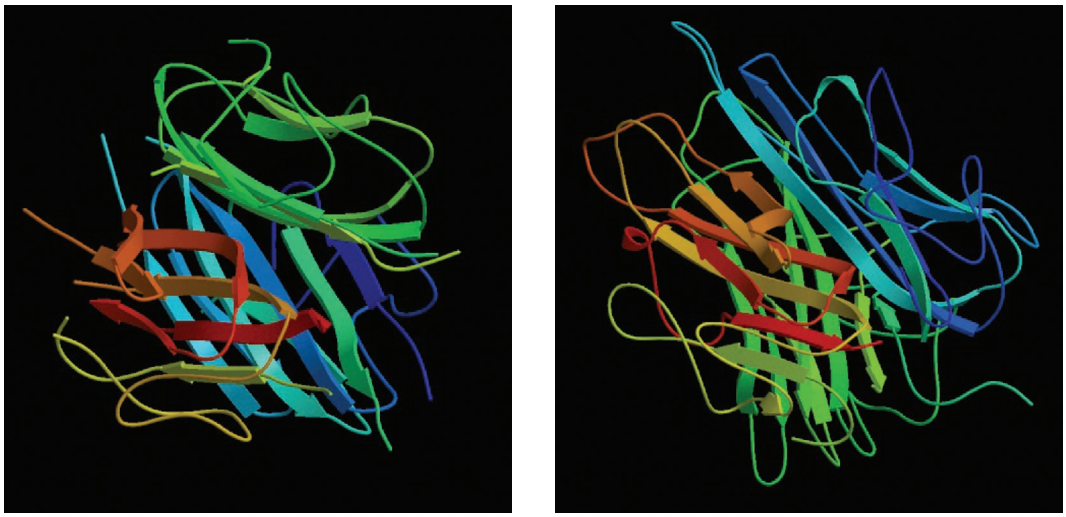


FIGURE 4.4 Two-dimensional DIGE. Overlay image of Cy3- (*green*) and Cy5- (*red*) labeled test-spiked *Erwinia carotovora* proteins. The protein test spikes were three conalbumin isoforms (*arrowheads*) and two myoglobin isoforms (*arrows*). Spots that are of equal intensity between the two channels 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.)

(a)



(b)

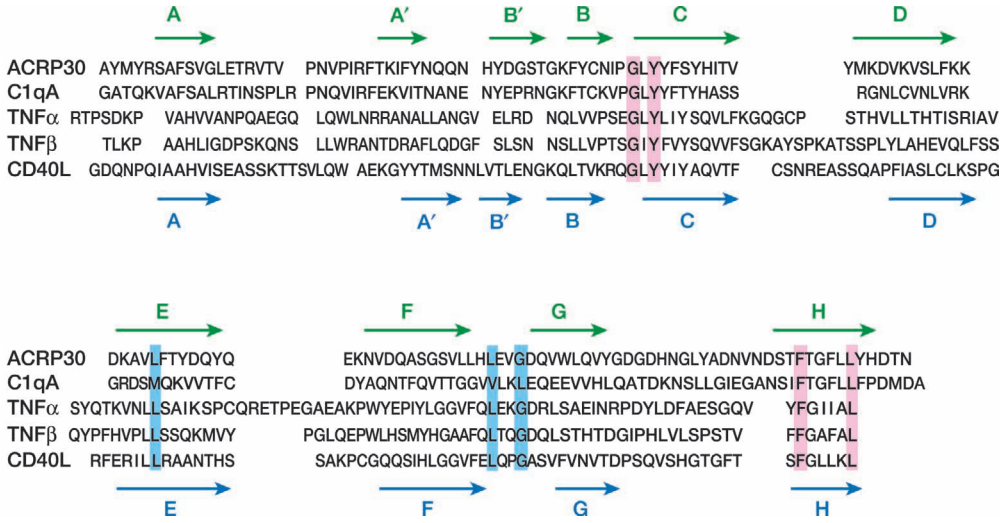


FIGURE 6.1 Identification of related proteins by structural comparison. (a) A ribbon diagram comparison of AdipoQ (left) and TNFα (right). The structural similarity is equivalent to that within the TNF family. (b) Structure-based sequence alignment between several members of the TNF family (CD40L, TNFα, 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.)

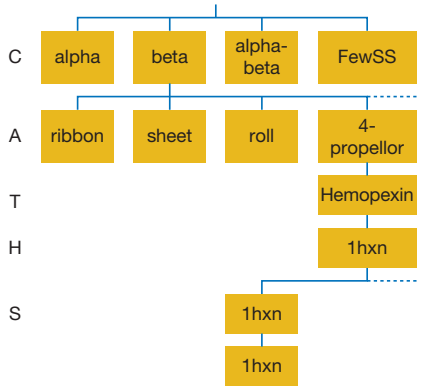


FIGURE 6.8 Structural classification of proteins using the CATH database. The protein shown is hemopexin, a protein rich in β-sheets with few α-helices. (Courtesy of Christine Orengo.)

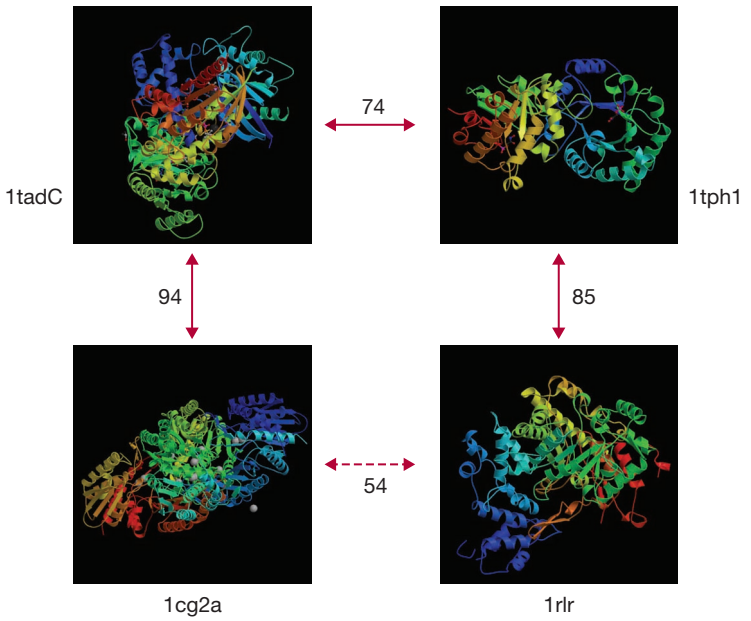


FIGURE 6.9 The Russian doll effect. Four proteins are illustrated that show continuous structural variation over fold space. Each of the proteins shares at least 74 structurally equivalent residues with its nearest neighbor, but the two extreme proteins show only 54 structurally equivalent residues when 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.)

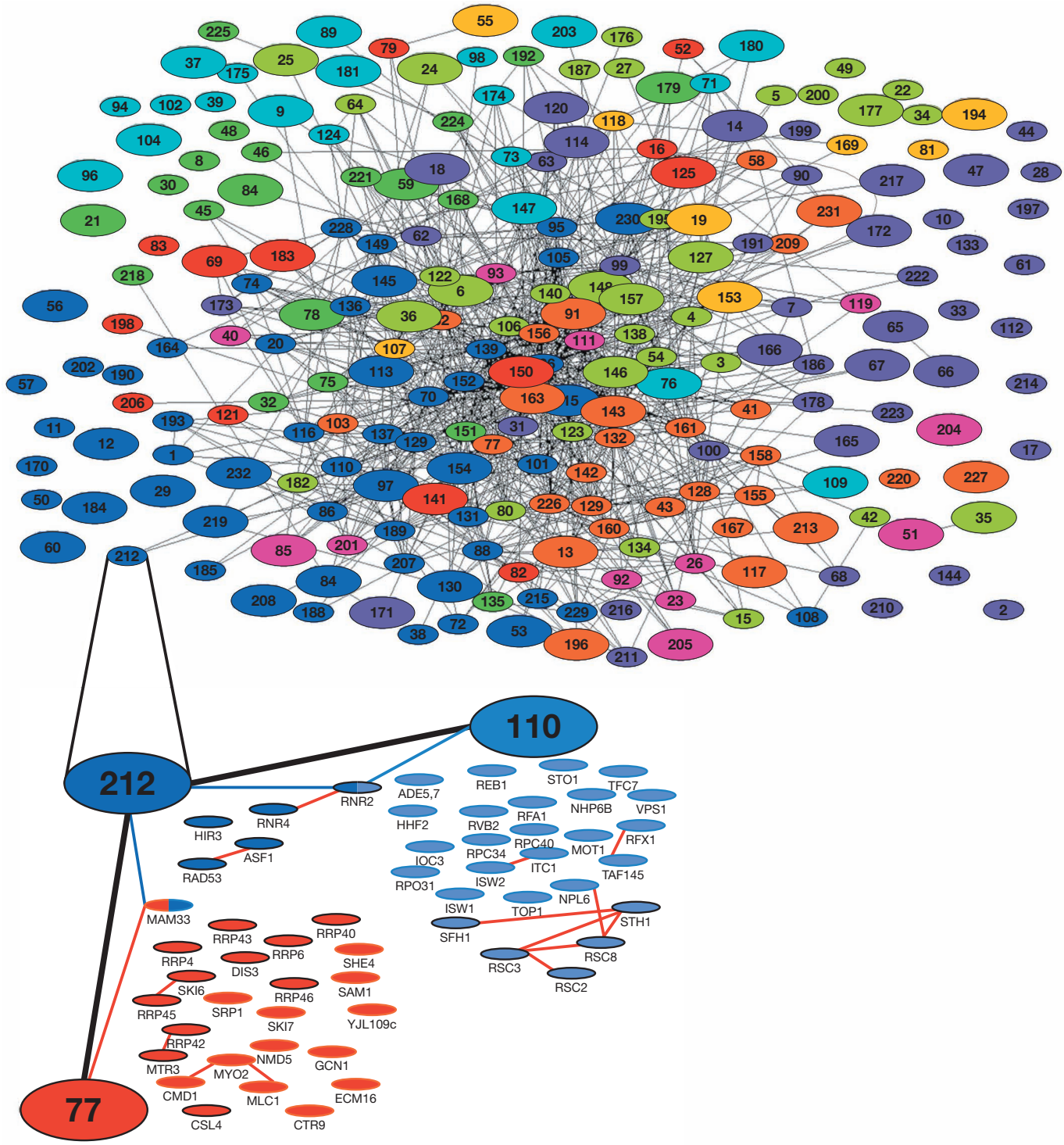


FIGURE 7.19 The protein complex network, and grouping of connected complexes. Links were established between complexes sharing at least one protein. For clarity, proteins found in more than nine complexes were omitted. The graphs were generated automatically by a relaxation algorithm that finds a local 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.)