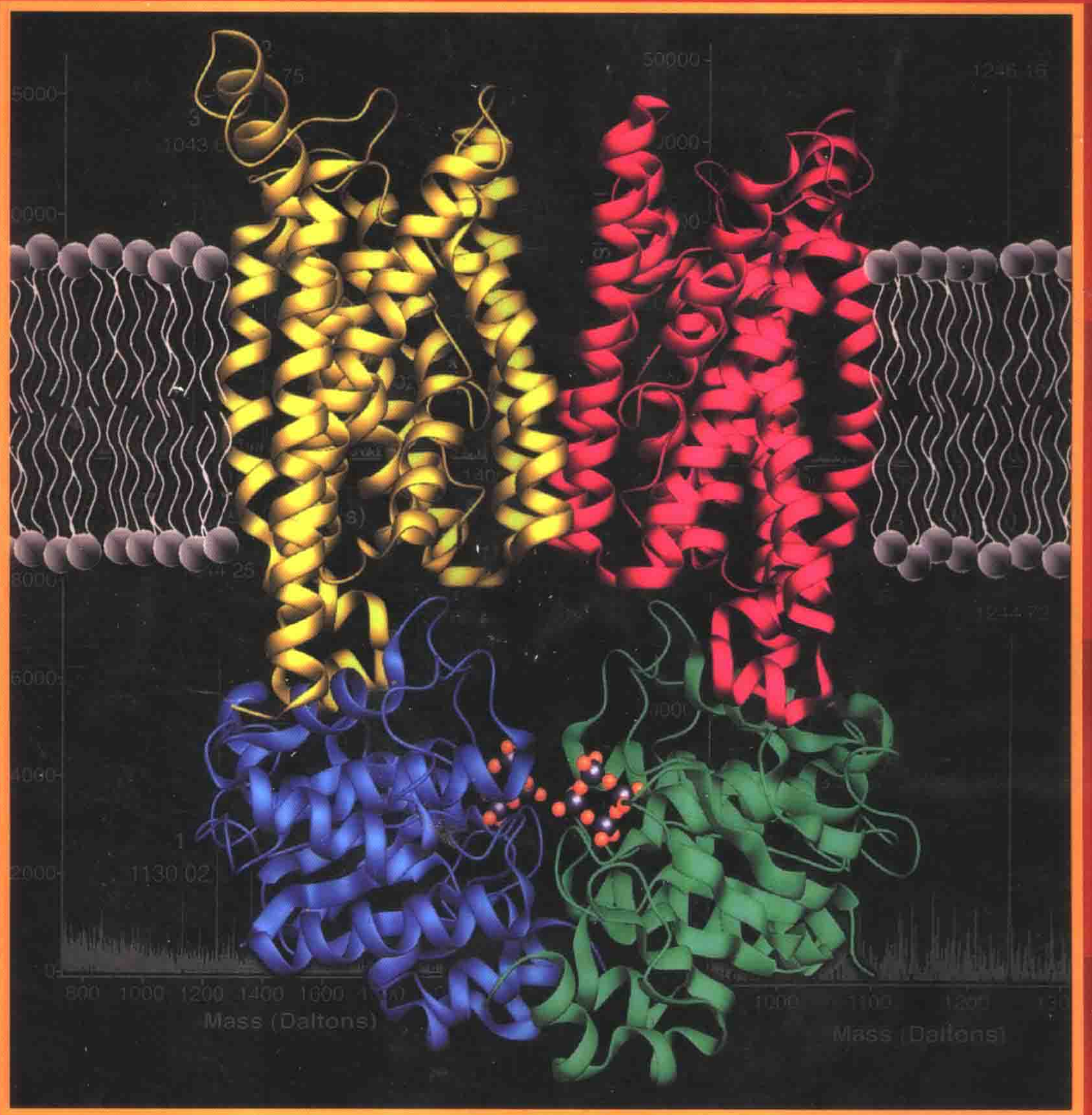
分子克隆
实验指南系列

蛋白质与蛋白质组学实验指南

Proteins and Proteomics: A Laboratory Manual

（影印版）

〔澳〕R. J. 辛普森 著



科学出版社
www.sciencep.com

分子克隆
实验指南系列

蛋白质与蛋白质组学实验指南

〔澳〕 R.J.辛普森 著

科学出版社
北京

图字 01-2003-4092

内容简介

随着人类基因组及其他一系列生物基因组测序的完成，蛋白质组研究的大门已经开启。本书是冷泉港实验室出版社最新推出的《蛋白质与蛋白质组学实验指南》（*Proteins and Proteomics*：*A Laboratory Manual*）的英文影印版，秉承了冷泉港实验手册一贯的先进性、实用性和权威性。本书提供了一系列可操作性强、行之有效的实验方案对蛋白质（细胞及细胞通路中）进行分离、鉴定、定量及功能分析，也有适于分析蛋白质结构包括电泳及色谱制备分析的方法，以及蛋白质芯片和生物信息学的相关内容。

本书内容丰富，信息全面，方案设计完善，技术适用范围较广，并为所涉及的技术的应用和相应原理提供了广泛的背景信息和参考文献，是遗传学、分子生物学研究者从基因组学、基因型研究转向蛋白质组和表型研究的必备工具书，适于生物化学、分子生物学、细胞生物学、遗传学、免疫学、蛋白质组学、功能基因 学等生命科学相关领域研究院所、高校相关院系、实验室的教师、研究生、科研人员，以及生物技术企业的研发者和决策者参考使用。

图书在版编目（CIP）数据

蛋白质与蛋白质组学实验指南/（澳）R.J. 辛普森（Richard J. Simpson）著. —北京：科学出版社，2003.9
（分子克隆实验指南系列）
书名原文：Proteins and Proteomics：A Laboratory Manual
ISBN 7-03-011912-6

Ⅰ. 蛋…　Ⅱ. 辛…　Ⅲ. 蛋白质-实验　Ⅳ. Q51-33

中国版本图书馆 CIP 数据核字（2003）第 071995 号

责任编辑：莫结胜
责任印制：刘士平/封面设计：王　浩

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com

源海印刷有限责任公司 印刷
科学出版社发行　各地新华书店经销

*

2003 年 9 月第 一 版　　开本：890×1240　1/16
2003 年 9 月第一次印刷　　印张：58 1/2
印数：1—3 000　　字数：1 959 000

定价：98.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

Proteins and Proteomics Companion Web Site

A COMPANION WEB SITE (www.proteinsandproteomics.org) to *Proteins and Proteomics: A Laboratory Manual* provides supplemental information about this fast-moving field of research. The site will include:

- References linked to Medline.
- Links to other databases of value to working scientists.
- Selected figures from the book for use in troubleshooting.
- Assistance in troubleshooting HPLC and 2D electrophoresis problems.
- A chapter on the analysis of carbohydrate from proteins (Oxley et al.; see abstract below).

Analysis of Carbohydrate from Glycoproteins

David Oxley,* Graeme Currie,† and Antony Bacic†

**The Babraham Institute, Babraham, Cambridge CB2 4AT, United Kingdom; †Plant Cell Biology Research Centre, School of Botany; University of Melbourne, Victoria 3010, Australia*

ABSTRACT

The identification and analysis of glycans associated with proteins is a formidable challenge due to the complexity and diversity of these carbohydrate structures. This chapter attempts to guide the "nonexpert" through the strategies used for the detection and analysis of carbohydrates on proteins. The "best" approach is very much dictated by the availability of material and the extent of information required. Many of the techniques can be used in most biochemistry/molecular biology laboratories, whereas others are very much in the bailiwick of the specialist laboratories that have access to highly sophisticated instrumentation. This chapter will answer at the very least the question most asked: Is my protein glycosylated? For many investigators, this level of information may suffice. For those inspired to ask the next logical question—What is the structure of this carbohydrate?—this chapter will provide sufficient background to proceed with confidence.

In addition to a brief history of carbohydrate research and a presentation of carbohydrate nomenclature, the chapter explains the fundamental steps involved in the analysis of glycoprotein-derived glycans. Each section within the chapter is designed to provide a concise understanding of the methods available to obtain the goal of the section, some of the limitations of each method, and references to some excellent texts on practical matters. Four widely used protocols are included, which detail the removal of glycans from glycoproteins and the preparation of monosaccharides for analysis by gas chromatography coupled with mass spectrometry.

- A protocol on the use of a multicompartment electrolyzer for the isoelectric fractionation of samples prior to 2D gel electrophoresis (Herbert et al.; see abstract below).

Sample Preparation for High-resolution Two-dimensional Electrophoresis by Isoelectric Fractionation in a Multicompartment Electrolyzer

Ben R. Herbert,[*] Pier Giorgio Righetti,[†] John McCarthy,[*] Jasmine Grinyer,[*] Annalisa Castagna,[†] Matthew Laver,[*] Matthew Durack,[*] Gerard Rummery,[*] Rebecca Harcourt,[*]and Keith L. Williams[*]

[*]*Proteome Systems, North Ryde, Sydney, NSW, 1670, Australia;* [†]*University of Verona, Department of Agricultural and Industrial Biotechnologies, Strada Le Grazie No. 15, 37134 Verona, Italy*

ABSTRACT

Two common problems when using a broad pH gradient with two-dimensional gel electrophoresis for the separation of proteins are low resolution of hydrophobic, highly acidic or basic proteins and poor detection of low-abundance proteins. Increasing resolution and enhancing detection on 2D gels is possible with the use of narrow and ultra-narrow range (1–3 pH units and <1 pH unit, respectively), immobilized pH gradients (IPGs). However, when narrow pH gradients are loaded with an entire cell lysate, a large proportion of the protein sample is not isoelectric within the separation range of the pH gradient. These "extraneous" proteins severely disturb the separation, because they have pIs outside the pH range of the IPG. This phenomenon is aggravated at high protein loads. Although it can be almost eliminated by loading a small amount of protein, this unfortunately is of very limited use for proteomics.

One solution to these problems is to employ a multicompartment electrolyzer (MCE), an instrument that fractionates protein samples isoelectrically prior to the creation of 2D maps. The resulting protein fractions match the pH intervals to be adopted as the first dimensions of the subsequent 2D maps. The fractionated protein mixture, devoid of proteins with isoelectric points outside of the range of the IPGs, can be loaded in a 2D map at much higher levels, thus ensuring greater sensitivity and detection of low-abundance proteins. Isoelectric fractionation using the MCE is fully compatible with subsequent 2D protocols, because it is based on a focusing technique that yields highly concentrated samples devoid of salts and buffers. The current MCE instrument uses commercially available, amphoteric, buffered membranes that are matched to the pH endpoints of commonly used IPGs.

Additional information will be added after the book is published. To access the Web Site:

1. Open the home page of the site.
2. Follow the simple registration procedure that begins on that page (no unique access code is required, since the site is open to anyone who completes the registration process).
3. Your e-mail address and password (selected during the registration process) become your log-in information for subsequent visits to the site.

The FAQ section of the site contains answers about the registration procedure. For additional assistance with registration, to inform us of other Web address changes, and for all other inquiries about the proteinsandproteomics.org Web Site, please e-mail support@proteinsand proteomics.org or call 1-800-843-4388 (in the continental U.S. and Canada) or 516-422-4100 (all other locations) between 8:00 A.M. and 5:00 P.M. Eastern U.S. time.

Preface

NOW THAT THE FIRST DRAFT OF THE HUMAN GENOME SEQUENCE is in the public domain, the primary focus of biologists is rapidly shifting toward gaining an understanding of how genes function, i.e., the functional roles of the full complement of encoded proteins. As well as defining structural characteristics of proteins, this task requires an understanding of the temporal and spatial location of proteins within the cell, including the intricate nature of how proteins interact with one another. Analytical protein chemistry, or proteomics as it is now commonly known, has a vital role in this daunting task. As information began to flow from the various genome projects, it became apparent to Cold Spring Harbor Laboratory Press that there was a growing need to provide researchers with a source of reliable proteomics protocols. Not long after, at the urging of my colleague Joe Sambrook, author of the enormously successful manual *Molecular Cloning*, I was invited to tackle the challenge of writing a laboratory manual of analytical methods and protocols for proteomics studies. Thus, *Proteins and Proteomics: A Laboratory Manual* was conceived.

Proteins and Proteomics is aimed at those who wish to isolate proteins and peptides for subsequent proteomic analysis. It is written for an audience ranging from early graduate students to experienced investigators. *Proteins and Proteomics* is not an encyclopedic book covering all possible proteomics methods. Rather, the book covers only those proteomics methods and technologies that are in current use in my laboratory or those of trusted colleagues. In each chapter, I have endeavored to provide sufficient background knowledge to underpin the accompanying protocols. In areas outside the immediate ken of the protein chemist, such as glycobiology and informatics, I have sought the contributions of experts to cover these specific fields of experimentation. Accordingly, I thank Antony Bacic and Parag Mallick and their colleagues for valued contributions in glycobiology and informatics, respectively.

A work such as this does not see the light of day without much outside help and support. My first thanks go to my friends and colleagues in the Parkville precinct in Melbourne, who have responded generously to my sometimes intemperate requests to provide illustration material or technical review. I am greatly indebted to the editorial and production staff at Cold Spring Harbor Laboratory Press for their dedication and tireless efforts in checking references, facts, faulty constructions, and stylistic abominations and keeping me on schedule (almost). I owe a special debt of gratitude to Judy Cuddihy for her cheerful optimism that raised my spirits, Kaaren Janssen and Maryliz Dickerson for their guidance in the beginning of the project, Tamara Howard for diligent fact checking, Inez Sialiano for coordinating the project, Dorothy Brown for editorial assistance, Susan Schaefer for page layout, Denise Weiss for her elegant design of the book, and most of all, Michael Zierler for his unstinting support as Senior Developmental Editor in steering the book to completion. I also acknowledge the generous support of Jan Argentine, my Managing Editor, and John Inglis, the Director of Cold Spring Harbor Laboratory Press, for overseeing the project.

Finally, I owe a special debt of gratitude to Mary Whitham and Pamela Jones for their secretarial assistance. I must also thank Simone Pakin especially for her superb information gathering skills and for her dedication and shockproof resilience that helped us both survive the process. My thanks also go to the members of the Joint ProteomicS Laboratory (JPSL), especially Robert Moritz, Hong Ji, David Frecklington, Lisa Connolly, James Eddes, Eugene Kapp, and Gavin Reid, for the support that they have given me in more ways that I can list here. I also acknowledge the rich intellectual and collegial environment at the Ludwig Institute for Cancer Research and especially the ungrudging support given to me by its director, Tony Burgess. Without his sustaining enthusiasm, this book would not have been possible.

To this long list of creditors I must finally add my partner, Donna Dorow, who has had to suffer with increasing patience the growing and seemingly never-ending demands that the "book" came to make on our time (the never-ending weekend and late night writing episodes), my attention, and temper.

Richard J. Simpson

Acknowledgments

THE AUTHOR WISHES TO THANK THE FOLLOWING COLLEAGUES for their valuable assistance:

Ruedi Aebersold
Alastair Aitken
Ron D. Appel
Manuel Baca
Antony Bacic
Tomas Bergman
Tom Berkelman
Willy Bienvenut
Reinhard I. Boysen
Edward J. Bures
Annalisa Castagna
Ella Cederlund
Andrea Cinnamon
Lisa Connolly
Patrick W. Cooley
Garry L. Corthals
Graeme Currie
Jenny M. Cutalo
Marc Damelin
Catherine Déon
Leesa J. Deterding
Sam Donohoe
Janice L. Duff
Matthew Durack
Richard H. Ebright
James S. Eddes
David A. Fancy
David Frecklington
Ernesto Freire
Parag S. Ghandi
Robert Goode
David R. Goodlett
Andrew A. Gooley
Robin Gras
Timothy J. Griffin
Jasmine Grinyer
Melanie P. Gygi
Steven P. Gygi
Rebecca Harcourt
Lara G. Hays
Milton T.W. Hearn
Thomas P. Hennessy
Ben R. Herbert
Cameron J. Hill
Denis F. Hochstrasser
Wendy L. Holstein
Femia G. Hopwood
Geoff Howlett
Marion I. Huber
Toshiaki Isobe
Ole N. Jensen
Hong Ji
Hans Jörnvall
Eugene A. Kapp
Hooi Hong Keah
Rosalind Kim
Nancy Laird
Martin R. Larsen
Matthew Laver
Larry J. Licklider
Gavin MacBeath
Gregory S. Makowski
Parag Mallick
Matthias Mann
Edward M. Marcotte
John McCarthy
Scott A. McLuckey
Helmut E. Meyer
Robert L. Moritz
Philippe Mottay
Markus Müeller
Nikolai Naryshkin
Richard A.J. O'Hair
Yoshiya Oda
David Oxley
Sang-Hyun Park
Scott Patterson
Junmin Peng
Ronald T. Raines
Melinda L. Ramsby
Juri Rappsilber
Gavin E. Reid
Andrey Revyakin
Pier G. Righetti
Gerard Rummery
Michael T. Ryan
Jean-Charles Sanchez
David M. Schieltz
Albert Sickmann
Pamela A. Silver
Andrew J. Sloane
Paul E. Smith
Christopher S. Spahr
Hanno Steen
Allan Stensballe
Wayne R. Stochaj
Nobuhiro Takahashi
Masato Taoka
Kenneth B. Tomer
Klaus K. Unger
Adrian Velazquez-Campoy
Anne Verhagen
David B. Wallace
Michael P. Washburn
Valerie C. Wasinger
Keith L. Williams
Yoshio Yamauchi
Eugene C. Yi
Jian-Guo Zhang
Lynn R. Zieske

Foreword

STUDENTS AND EXPERIENCED RESEARCHERS FACE SIMILAR challenges when attempting to delve into a new field of research. They need to learn new terminology, concepts, and theories, define current research topics in the field, and master a new set of methods and techniques. Finding suitable resources may be as difficult as mastering the subject matter itself, especially when the topic is an emerging and rapidly evolving field, such as proteomics. Often, textbooks, glossaries, and reference manuals are sparse—if they exist at all—and the available information must be gleaned from numerous articles in the primary literature.

In this comprehensive book, Richard Simpson and a group of leading proteomics experts attempt the impossible: to condense theory, background information, protocols, and information resources into a single volume. They succeed. This manual contains virtually everything one would hope to find in both a textbook and laboratory manual: overviews, introductory materials, and a theoretical base for proteomics. Detailed protocols for common proteomics experiments, complete source lists for the tested materials, and web links to crucial reagent resources are also thoughtfully provided.

Proteins and Proteomics: A Laboratory Manual is an invaluable information tool both for the experienced protein chemist who bravely ventures into the new world of proteomics and for the novice to proteins and proteomes. By focusing on what is currently considered the bedrock of proteomics technologies, Professor Simpson ensures that—in spite of the rapid advances that characterize contemporary proteomics research—this volume will remain relevant and current for years to come.

Ruedi Aebersold
Professor, Institute for Systems Biology

CHAPTER 1

Introduction to Proteomics

Biological macromolecules are the main actors in the makeup of life.... To understand biology and medicine at a molecular level...we need to visualize the activity and interplay of large macromolecules such as proteins. To study protein molecules, principles for their separation and determination of their individual characteristics had to be developed. One of the most important chemical techniques used today for the analysis of biomolecules is mass spectrometry (MS), one of the subjects of the 2002 Nobel prize award.

The 2002 Nobel prize in Chemistry was awarded "for the development of methods for identification and structure analyses of biological macromolecules" with one half going jointly to John B. Fenn (Virginia Commonwealth University, Richmond, USA) and Koichi Tanaka (Shimadzu Corporation, Kyoto, Japan) "for their development of soft desorption ionization methods for mass spectrometric analyses of biological macromolecules."

This is a revolutionary breakthrough. Chemists and biologists can now rapidly and reliably identify what proteins a sample contains. Hence, scientists can both "see" the proteins and understand how they function within cells.

NOW THAT MORE THAN 40 GENOMES, INCLUDING THE HUMAN GENOME, have been fully sequenced and are in the public domain, the next challenge for biologists will be to connect gene to function, genotype to phenotype, to find out what the genes really do! The rapid pace of genome sequencing efforts during the past several years has resulted in many newly discovered genes that have been ascribed no function or a function that at best has been poorly described. For an up-to-date monitor of complete and ongoing sequencing projects, see the GOLD Web Site (Genomes OnLine Databases at http://wit.integratedgenomics.com/GOLD). This impetus to understand the function of newly discovered genes is leading biologists toward the systematic analysis of the expression levels of the components that constitute a biological system, chiefly, mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) (see Figure 1.1). Because proteins are central to biological function and obvious candidates for drug targeting, proteomics is enjoying a rapidly increasing level of attention.

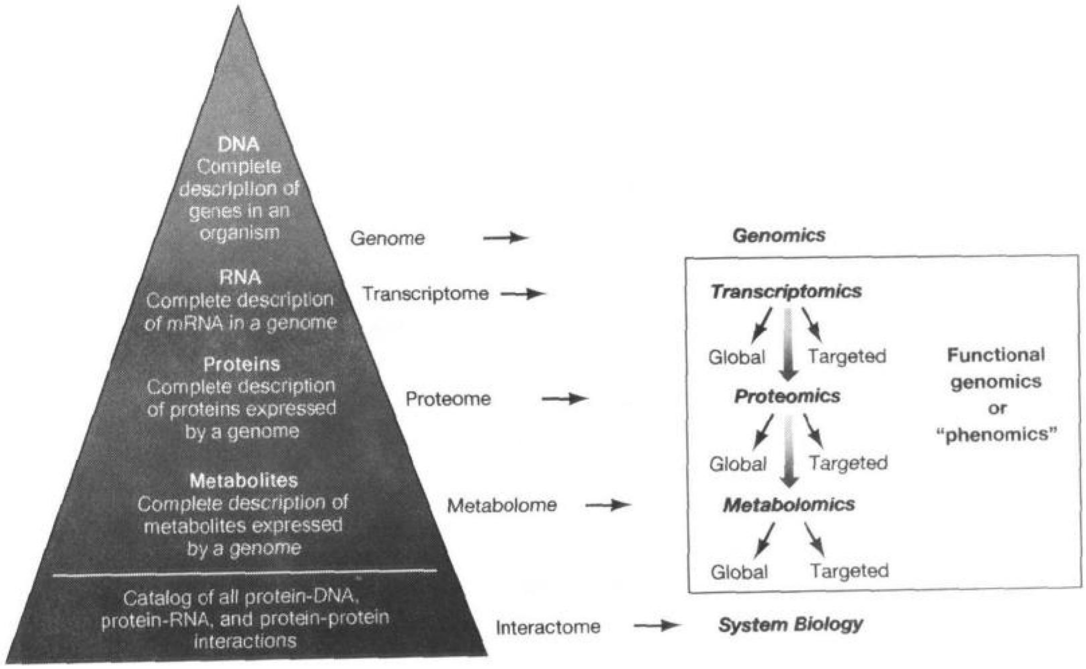

DNA
Complete description of genes in an organism
RNA
Complete description of mRNA in a genome
Proteins
Complete description of proteins expressed by a genome
Metabolites
Complete description of metabolites expressed by a genome
Catalog of all protein-DNA, protein-RNA, and protein-protein interactions
Genome
Transcriptome
Proteome
Metabolome
Interactome
Genomics
Transcriptomics
Global
Targeted
Proteomics
Global
Targeted
Metabolomics
Global
Targeted
System Biology
Functional genomics or "phenomics"

FIGURE 1.1. Functional genomics or phenomics. Genomics provides an overall description of the complete set of genetic instructions (the genes) contained within the genome that are available to a cell, i.e., the "blueprint" of a cell. Functional genomics, on the other hand, represents a systematic approach to elucidating the function of the novel genes revealed by complete gene sequences. (Phenomics has been suggested as an all-embracing term to describe functional genomics.) Functional genomics adopts a hierarchical strategy aimed at gaining a comprehensive and integrative view of the workings of living cells. There are a number of different approaches for studying the functional analysis of novel genes. These can be grouped into four domains: genome (the complete set of genes for an organism and its organelles), transcriptome (the complete set of mRNA molecules), proteome (the complete set of proteins), and metabolome (the complete set of metabolites, the low-molecular-weight intermediates). Researchers have now added the suffix "*ics*" to describe the utility for analyzing these domains. For example, the task of comparing the mRNA profiles using DNA arrays is now referred to as cellular (or tissue) transcriptomics, and the task of separating the cell's proteins and comparing their expression profiles is referred to as expression proteomics. Studying all of the proteins encoded by a genome (the proteome) without focusing on a particular cell type, growth conditions, and subcellular localization is the domain of global proteomics. Focus on protein expression within a particular cell type and/or subcellular organelle is the domain of targeted proteomics. It is now clear that these domains are not an end in themselves, but a vehicle to understanding an organism's entire metabolism, now referred to as metabolomics (Raamsdonk et al. 2001; Oliver 2002). Thus far, the fully sequenced genome studies have yielded many insights into the functional properties of proteins, especially the emergence of networks of interacting proteins (the term "interactome" has been coined to describe protein-protein networks). Understanding interactions between encoded proteins of a given genome is a critical first step in functional genomic analysis (Xenarios and Eisenberg 2001; Gerstein et al. 2002). To understand biology at the system level, and to develop models that explain the dynamics of cellular and organismal function (rather than the characteristics of isolated parts), all of these domains must be integrated (the "legome," total systems biology can be likened to assembling all of the component parts of a "Lego" set) in a quantitative and temporal manner (for a review on systems biology, see Brenner 1999b; Kitano 2002).

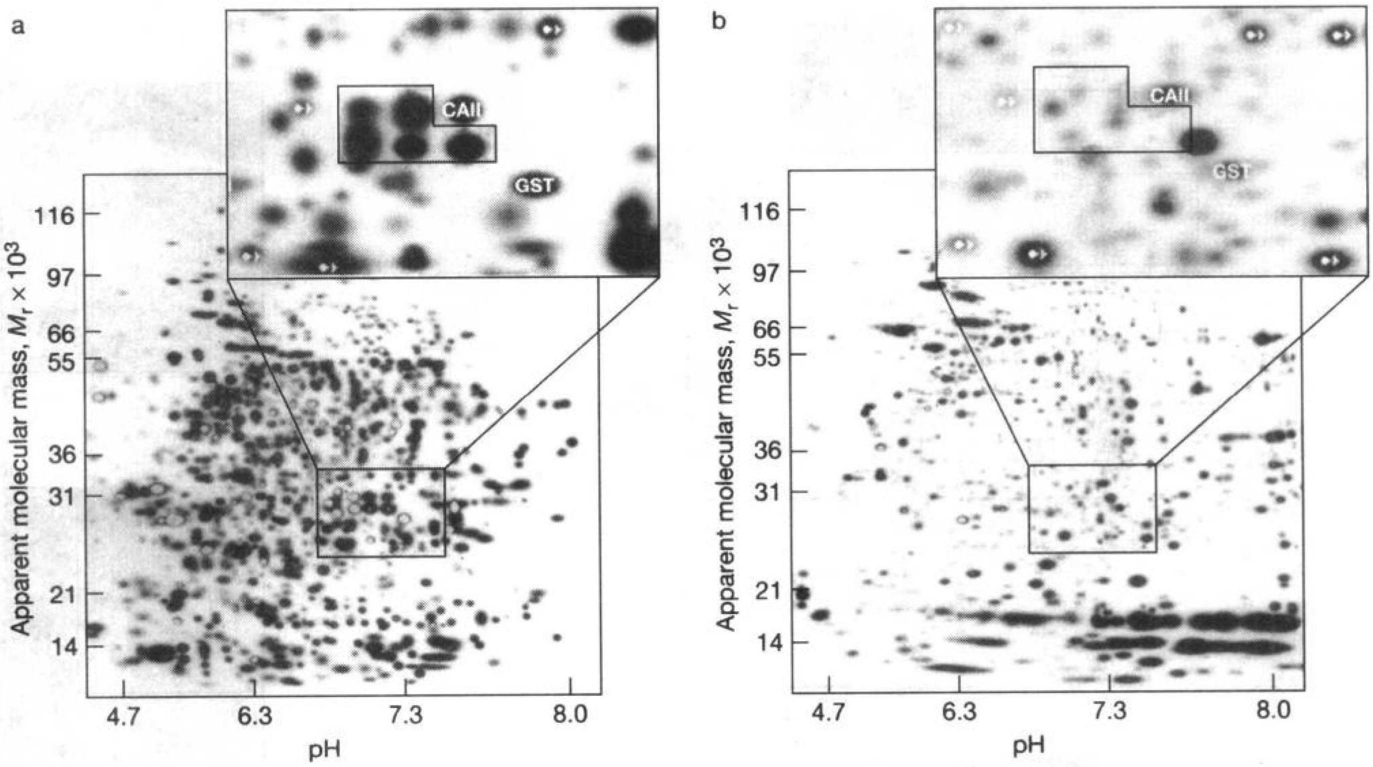


FIGURE 1.2. 2D gel electrophoresis of proteins from a whole-cell extract stained with Coomassie Brilliant Blue. (*a*) Wild-type C57/Black/6J murine colonic crypts; (*b*) polyps from multiple intestinal neoplasia (MIN) mice (Cole et al. 2000). The synthetic gel images were generated using PDQuest software. Differentially expressed proteins are marked in light gray. The insets show that carbonic anhydrase (CAII) and GST are both highly expressed in wild-type crypts and MIN polypts, whereas expression of several CAII isoforms (*outlined box*) is dramatically reduced in MIN polypts. (Reproduced, with permission, from Simpson and Dorow 2001 [© Elsevier Science].)

DEFINING PROTEOMICS

Proteomics or, more appropriately, functional proteomics refers to the branch of discovery science focusing on proteins. Initially, the term was used to describe the study of the expressed proteins of a genome using two-dimensional (2D) gel electrophoresis, and mass spectrometry (MS) to separate and identify proteins and sophisticated informatics approaches for deconvoluting and interrogating data. This approach is now referred to as "expression" or "global profiling" proteomics (Figures 1.2 and 1.3). The scope of proteomics has now broadened to embrace the study of "protein-protein" interactions (protein complexes), referred to as cell-mapping proteomics (Blackstock and Weir 1999) (see panel below on THE MANY FACES OF PROTEOMICS).

The term proteome, coined in 1994 as a linguistic equivalent to the concept of genome, is used to describe the complete set of proteins that is expressed, and modified following expression, by the entire genome in the lifetime of a cell. It is also used in a less universal sense to describe the complement of proteins expressed by a cell at any one time (from *Nature* 1999). Today, proteomics is a scientific discipline that promises to bridge the gap between our understanding of genome sequence and cellular behavior; it can be viewed as more of a biological assay or tool for determining gene function.

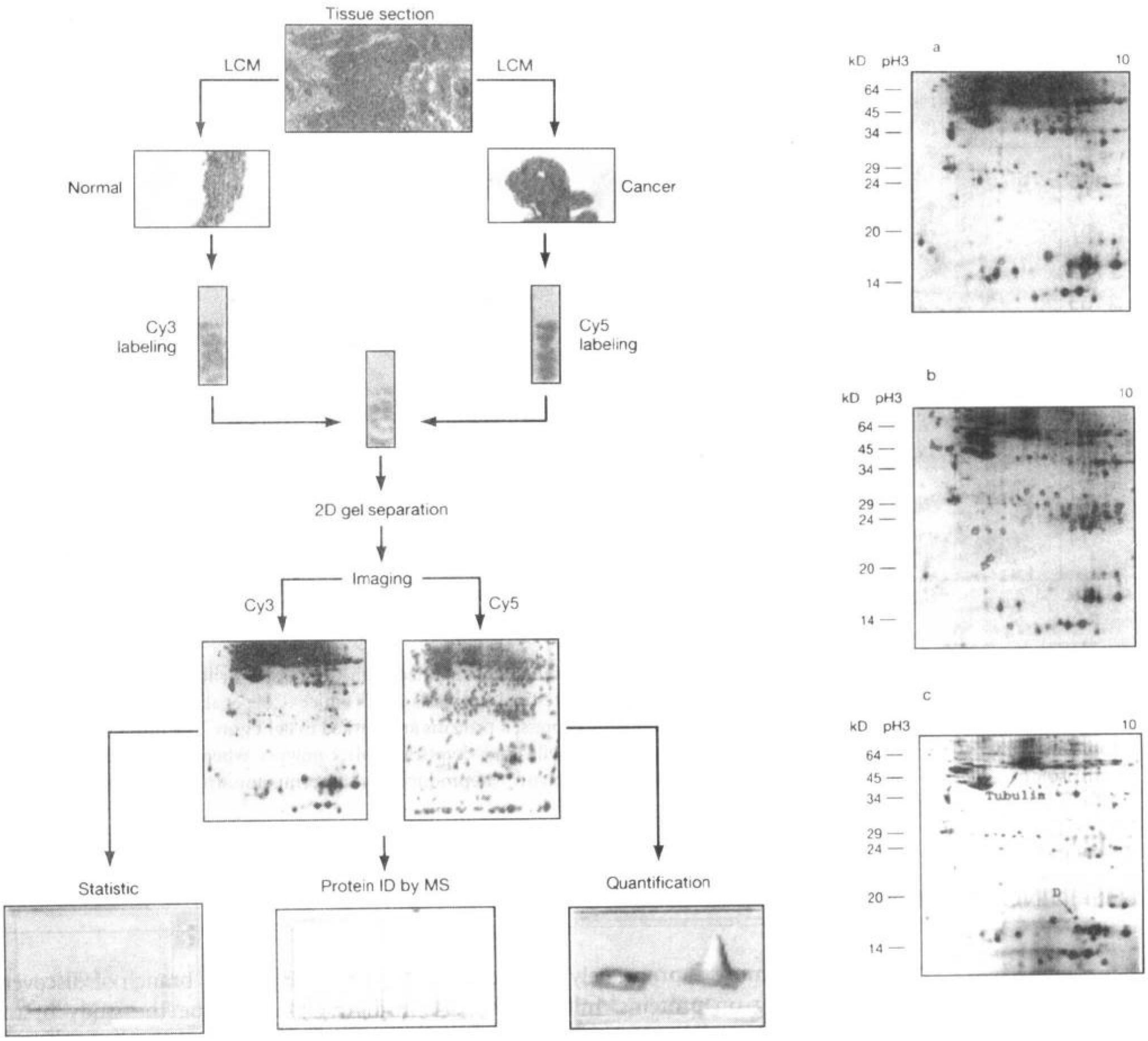


FIGURE 1.3. Differential in-gel electrophoresis (DIGE) for the identification of cancer markers. (*Top, left*) In DIGE, an emerging technology for proteome analysis (Unlu et al. 1997), two pools of proteins (e.g., normal cells and cancer cells procured from the same tumor sample using laser-capture microdissection) (Emmert-Bock et al. 1996) are labeled with 1-(5-carboxypentyl)-1′-propylindocarbocyanine halide (Cy3), N-hydroxy-succinimidyl ester and 1-(5-carboxypentyl)-1′-methylindodi-carbocyanine halide (Cy5) N-hydroxysuccinimidyl ester fluorescent dyes, respectively. The Cy3- and Cy5-labeled proteins are mixed and then separated in the same 2D gel. The 2D gel protein profiles can be rapidly imaged by the fluorescent excitation of either the Cy3 or Cy5 dye. (Cy3-labeled gel images are collected at an excitation wavelength of 540 nm and at an emission wavelength of 590 nm, whereas the Cy5-labeled gel images are collected at an excitation wavelength of 620 nm and an emission wavelength of 680 nm.) A comparison of the resulting images allows quantitation of each protein spot. (*Right, a*) Cy3 image of proteins from normal cells; (*b*) Cy5 image of proteins from tumor cells. Because both protein pools are electrophoretically separated in the same gel, those proteins existing in both pools will migrate to the same location in the 2D gel, thereby minimizing the inherent reproducibility problem associated with 2D gels. Quantitation of the protein profile can be rapidly and accurately achieved over a wide dynamic range (e.g., four orders of magnitude) based on fluorescent intensity (Patton 2000). The separated proteins in the 2D gel are next visualized by SYPRO Ruby staining (*right, panel c*). (The SYPRO Ruby-stained image is scanned at an excitation wavelength of 400 nm and an emission wavelength of 630 nm.) Protein spots of interest are excised and in-gel-digested with trypsin, and the peptides extracted for the purpose of identification by mass spectrometry (MS) methods described in Chapters 7 and 8. (ID) Identification. (Adapted, with permission from Zhou et al. 2002.)

The Many Faces of Proteomics

- ***Proteomic analysis (or analytical protein chemistry).*** The large-scale identification and characterization of proteins, including their posttranslational modifications, such as phosphorylation and glycosylation. Analysis is done with the aid of mass spectrometry or Edman degradation. For analysis of protein phosphorylation, see Chapter 9; and for amino-terminal sequence analysis using the Edman degradation procedure, see Chapter 6.
- ***Expression proteomics (or differential display proteomics).*** Two-dimensional gels are used for global profiling of expressed proteins in cell lysates and tissues. This conventional approach is being challenged by non-2D gel methods, such as liquid-based isoelectric focusing (IEF) or ion-exchange chromatography/reversed-phase high-performance liquid chromatography (RP-HPLC). Proteins are typically identified by mass spectrometry (MS). In many situations, these methods are complemented by DNA-based array methods. Includes *quantitative proteomics* (for a review of proteomics strategies for the quantitative analysis of paired protein samples [e.g., normal vs. diseased] utilizing stable isotope labeling combined with chromatographic separations, see Chapter 8 and Patterson 2000a,b).
- ***Cell-mapping proteomics (or cataloging of protein-protein interactions).*** Protein-protein interactions and intracellular signaling circuitry are determined by the identification of protein complexes (obtained by affinity purification and protein identifications by MS) or by direct DNA readout (e.g., yeast two-hybrid, phage display, ribosome display, and RNA-peptide fusions). For reviews on protein networks, see Legrain (2002) and Mayer and Hieter (2000); on cell-mapping proteomics, see Blackstock and Weir (1999), Lakey and Raggett (1998), and Duan et al. (2002); and on mapping protein-protein interactions with combinatorial biology methods that rely on direct DNA readout, see Pelletier and Sidhu (2001).

WHY PROTEOMICS IN ADDITION TO GENOMICS?

Large-scale Genome Sequencing: What Have We Learned?

One of the most exciting biological achievements to emerge during the past 40 years has been the completion of draft DNA sequences of the human genome, published by the International Human Genome Sequencing Consortium (a publicly funded project) (Lander et al. 2001) and by Celera Genomics (a commercial effort) (Venter et al. 2001). These Herculean efforts provide a blueprint of the information needed to create a human being and reveal, for the first time, the organization of a vertebrate's DNA (for an overview of this project, see Baltimore 2001). One of the interesting findings about the human genome is the number of genes found. The public project estimates that there are 31,000 protein-encoding genes, whereas Celera finds ~26,000, with many more still to be found. (A current estimate is that the number of protein-encoding genes may be on the order of 60,000.)

Interestingly, the number of coding genes in the human sequence is not dramatically different from the numbers reported for phylogenetically remote organisms: 6,000 for a yeast cell, 13,000 for a fly, 18,000 for a worm, and 26,000 for a plant (Genomes OnLine Databases at http://wit.integratedgenomics.com/GOLD). The number of genes reported for multicellular organisms is not highly accurate because of the limitations of existing ab initio gene prediction methods used to identify genes (Dunham et al. 1999). The existence of an open reading frame (ORF) in genomic data does not necessarily imply the existence of a functional gene. In human DNA, gene prediction by ab initio methods is notoriously difficult because of the extensive alternative splicing (Black 2000), lower density of exons, and high proportion of interspersed repetitive sequences. Given the unreliability of ab initio gene prediction software, all genes will need to be experimentally identified and annotated. For example, the error rate in the annotations for 340 genes from the *Mycoplasma genitalium* genome was ~8% (Brenner 1999a). Hence, verification of a gene product by proteomic analysis is an important first step in annotating the genome.

Disparity between mRNA Profiling and Protein Profiling

No simple correlation exists between changes in mRNA expression levels (transcriptomics) and those in protein levels (proteomics). Indeed, the link between transcript levels and protein levels in a given cell or tissue is tenuous, to say the least, and it is clear that array-based gene expression monitoring or other gene expression methods for measuring mRNA abundances, alone, are insufficient for analyzing the cell's protein complement (for a review of global gene expression methodologies, see Lockhart and Winzeler 2000). Recent studies show a marked disparity between the relative expression levels of mRNAs and those of their corresponding proteins (Anderson and Seilhamer 1997; Gygi et al. 1999a). A further complication arises when considering the complementarity of genomics and proteomics. Despite the adage that one gene gives rise to one protein, the situation in eukaryotic cells is more likely six to eight proteins per gene (Strohman 1994). Thus, there may be several hundred thousand human proteins after splice variants and essential posttranslational modifications are included. For example, 22 different forms of human α-1-antitrypsin have been observed in human plasma (Hoogland et al. 1999). Fortunately, such biological complexity can be unraveled using proteomic studies to understand how cells modulate and integrate signals.

Origins of Cellular Complexity

From the genome sequencing efforts to date, it is clear that the physiological complexity of organisms is not merely a consequence of gene numbers. For instance, humans (although composed of ~10,000,000,000,000 cells) have fewer than twice as many genes as the 959-cell nematode, *Caenorhabditis elegans*. Rather, evolution of the increased complexity of higher-order organisms is due to a number of other mechanisms, such as alternative splicing (Mironov et al. 1999; Black 2000), diversification of gene regulatory networks, and the ability of intracellular signaling pathways to interact with one another (Weng et al. 1999; Davidson et al. 2002). Biological signaling pathways can interact to form complex networks comprising a large number of components. Such complexity arises from the overlapping functions of components, from the connections among components, and from the spatial relationship between components in the cell. Additionally, many cellular processes are performed and regulated not by individual proteins but by proteins acting in large protein assemblies or macromolecular complexes. For instance, the eukaryotic ribosome, which translates RNA into protein, consists of ~80 unique proteins (Wool et al. 1995), and the RNA polymerase II transcription complexes in eukaryotic cells, which is involved in DNA replication, comprises at least 50 different proteins (Pugh 1996). For a further discussion of protein-protein interactions, see Chapter 10.

The traditional view of protein function tends to focus on the biochemical activity of a single protein molecule such as the catalysis of a given reaction or the binding of a ligand to its cognate receptor. This local function is often referred to as the "molecular function" of a protein. However, an expanded view of protein function is beginning to emerge in the post-genome era, with a protein being defined as an element in its network of interactions. This notion of expanded function has been variously referred to as "contextual function' or "cellular function" (see Kim 2000). The contemporary view of function is that each protein in living matter operates as an integral component of an intricate web of interacting molecules. For excellent reviews on this subject, see Weng et al. (1999) and Eisenberg et al. (2000).

Integrated Biology (Total Systems Biology)

To achieve a full understanding of how a complex organism works, biologists must develop an integrated (or global) view of a cell's mRNA and protein complements and a detailed knowledge of how these complements change with development and the environment (especially in disease). Mathematically, expression profiles from both mRNA and protein are required to fully understand how a gene network operates (see Hatzimanikatis and Lee 1999). For example, an integrated genomic and proteomic analysis of a systematically perturbed glucose/galactose-utilizable pathway in yeast concluded that an analysis of both mRNAs and proteins is crucial for understanding biological systems (see Ideker et al. 2001).

Protein-protein interactions are a crucial component of this integrated biology. Already, a large proportion of known protein-protein interactions in yeast have been identified by genome-scale yeast two-hybrid assays (Legrain and Selig 2000; Schwikowski et al. 2000; Uetz et al. 2000; Hazbun and Fields 2001; Ito et al. 2001a,b; Legrain et al. 2001) and direct affinity capture methods (Gavin et al. 2002; Ho et al. 2002). However, the interactions detected by these physical methods may include nonspecific interactions of no biological significance. Biologically important protein-protein interactions require that the interacting partners be in specific *protein states* (interactions may result in the transition of one protein state to another), but physical methods, like two-hybrid assays and affinity capture, do not distinguish between protein states of a given protein molecule. The following is a list of attributes that define protein states (also see Figure 1.4).

- ***Covalent modification*** (e.g., phosphorylation, glycosylation, lipidation, nitrosylation, acetylation, and ubiquitination). A protein may occur in its "active" or "inactive" form depending on the state of covalent modifications, such as phosphorylation (Hunter 2000a,b).
- ***Cellular localization.*** Depending on the biological status of the cell, a protein molecule may reside in one or several cellular locations, such as the nucleus, cytosol, plasma membrane, mitochondria, and endoplasmic reticulum.
- ***Presence of ligands.*** The binding of small molecules and ions (e.g., heme, metal ion, glucose, ATP, ADP, GTP, and GDP) to proteins alters protein states, affecting properties such as rates of enzyme catalysis and allostery.
- ***Alternate splicing.*** Different forms of a protein molecule may result from alternate splicing of the gene product.
- ***Proteolytic cleavage.*** Truncated forms of a protein molecule may result from specific amino- or carboxy-terminal cleavage or internal cleavage. These truncations alter the

state and activity of the protein. For exampl^e, certain proteolytic enzymes are produced as inactive precursors (zymogens), which must be cleaved to generate an active enzyme. By restricting the synthesis, location, or activity of the necessary proteases, cells have a means of regulating proteolytic activity with respect to time and cellular localization (Khan and James 1998; Kobe and Kemp 1999). Examples include proteases involved in blood clotting, catabolic digestion (e.g., pepsin, rennin, trypsin, chymotrypsin, and carboxypeptidase are secreted as inactive precursors), apoptosis (e.g., the key effector molecules, the caspases, are present as inactive zymogens [Shi 2002]), cleavage of viral precursor proteins to functional units, and pattern formation in multicellular organisms (Khan and James 1998). Proteolytic activity is further controlled by specialized proteins that specifically inhibit the active proteases (Bode and Huber 1992; Khan and James 1998).

- ***Oligomeric state.*** A protein molecule may exist in a multiprotein complex or as a homodimer or homo-oligomer. From yeast interactome studies, it is estimated that at least 78% of yeast proteins occur in complexes (Gavin et al. 2002).
- ***Protein conformation.*** Three-dimensional structure information on different protein states is important for understanding their biological behavior. For example, protein function is often regulated by allosteric mechanisms (Monod et al. 1963), in which effector molecules bind to regulatory sites distinct from the active site, usually inducing conformational changes that alter the activity. Allosteric effectors usually bear no structural resemblance to the substrates of their target protein, the classic example being end products of metabolic pathways acting at early steps of the pathway to exert feedback control. Intrasteric regulation (Kemp and Pearson 1991), on the other hand, includes autoregulation of protein kinases and phophatases by internal amino acid sequences that resemble the substrate (such internal amino acids are often referred to as pseudosubstrates). This type of regulation is considered the counterpart of allosteric control (see Figure 1.5). Examples of protein kinases whose regulation is mediated by intrasteric autoregulatory sequences include Twitchen kinase, Titin kinase, CaMK-1, insulin receptor kinase, and MAP kinase ERK2 (for reviews, see Kobe and Kemp 1999; Huse and Kuriyan 2002).

For a descriptive database of biological protein interactions organized in terms of protein states and state transitions, see LiveDIP (http:www.dip.doe-mbi.ucla.edu/) (Duan et al. 2002). Additional information on DIP (the Database of Interacting Proteins) can be found in Chapter 11 and in Table 1.1, which provides a list of Web-accessible databases containing information on protein-protein interactions.

One of the difficulties (and challenges) in studying protein-protein interactions is that at any given time, the pool of molecules of a protein inside a cell most likely represents one or several of the protein states of that particular protein, depending on the cellular context. This is a major impetus of proteomics, especially cell-mapping proteomics, which aims to describe all protein-protein interactions (both spatially and temporally) within a given cell. The challenge of proteomics is to utilize existing technologies (and to develop new technologies) to define all of the protein states for a given protein molecule. Such information is of crucial importance in post-genome biology, especially total systems biology, since it will shed light on the molecular mechanisms underlying biological processes. For reviews, see Xenarios and Eisenberg (2001) and Gerstein et al. (2002) for protein-protein interactions, and for systems biology, see Brenner (1999b) and Kitano (2002).