Advances in PROTEIN CHEMISTRY

VOLUME 65

Proteome Characterization and Proteomics

ACADEMIC PRESS

ADVANCES IN PROTEIN CHEMISTRY

EDITED BY

FREDERIC M. RICHARDS
Department of Molecular Biophysics
and Biochemistry, Yale University
New Haven, Connecticut

DAVID S. EISENBERG
Department of Chemistry and Biochemistry
University of California, Los Angeles
Los Angeles, California

JOHN KURIYAN
Department of Molecular Biophysics
Howard Hughes Medical Institute
Rockefeller University
New York, New York

VOLUME 65

Proteome Characterization and Proteomics

EDITED BY

RICHARD D. SMITH
Pacific Northwest National Laboratory
Richland, Washington

TIMOTHY D. VEENSTRA
National Cancer Institute
Frederick, Maryland

ACADEMIC PRESS

An imprint of Elsevier Science

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

This book is printed on acid-free paper. ♾

Academic Press
An Elsevier Science Imprint.
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.academicpress.com

Academic Press
84 Theobald's Road, London WC1X 8RR, UK
http://www.academicpress.com

International Standard Book Number: 0-12-034265-2

PRINTED IN THE UNITED STATES OF AMERICA
03 04 05 06 07 08 9 8 7 6 5 4 3 2 1

ADVANCES IN PROTEIN CHEMISTRY

Volume 65

Proteome Characterization and Proteomics

CONTENTS

Application of Separation Technologies to Proteomics Research

HALEEM J. ISSAQ

Proteomics of Membrane Proteins

JULIAN P. WHITELEGGE, STEPHEN M. GÓMEZ, AND KYM F. FAULL

Proteomics in Drug Discovery

RODNEY M. HEWICK, ZHIJIAN LU, AND JACK H. WANG

From Clone to Crystal: Maximizing the Amount of Protein Samples for Structure Determination

CHRIS M. KOTH AND ALED M. EDWARDS

Proteomics and Bioinformatics

CAROL S. GIOMETTI

PROTEOMICS IN THE POSTGENOMIC AGE

By RICHARD S. MORRISON,* YOSHITO KINOSHITA,* MARK D. JOHNSON,* AND THOMAS P. CONRADS†

*Department of Neurological Surgery, University of Washington School of Medicine, Seattle, Washington 98195, and †SAIC-Frederick, National Cancer Institute at Frederick, Frederick, Maryland 21702

I. Introduction

Advances in molecular biology and bioinformatics are making it possible to simultaneously analyze the entire complement of genes expressed in a particular cell or tissue. These advances have created unique opportunities in the field of medicine, where the results of gene expression studies are expected to help identify cellular alterations associated with disease etiology, progression, outcome, and response to therapy. These rapidly emerging technologies are also expected to result in the identification of novel therapeutic targets for a host of maladies, including infectious diseases, behavioral disorders, developmental defects, neurodegenerative diseases, aging, and cancer.

Technical advances have facilitated characterization of the three major genetic units: the genome, the transcriptome, and the proteome (Fig. 1). The *genome* describes the entire set of genes encoded by the DNA of an organism. The *transcriptome* encompasses the entire complement of messenger RNA (mRNA) transcripts transcribed from the genome of a cell. The transcriptome varies from cell to cell and fluctuates in response to numerous physiological signals, including developmental status, stress, changes in the extracellular milieu, and disease. The *proteome* describes the entire complement of proteins expressed by a cell at a point in time. Proteomic investigations also aim to determine protein localization, modifications, interactions, and, ultimately, protein function. Because the

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65

0065-3233/03 $35.00

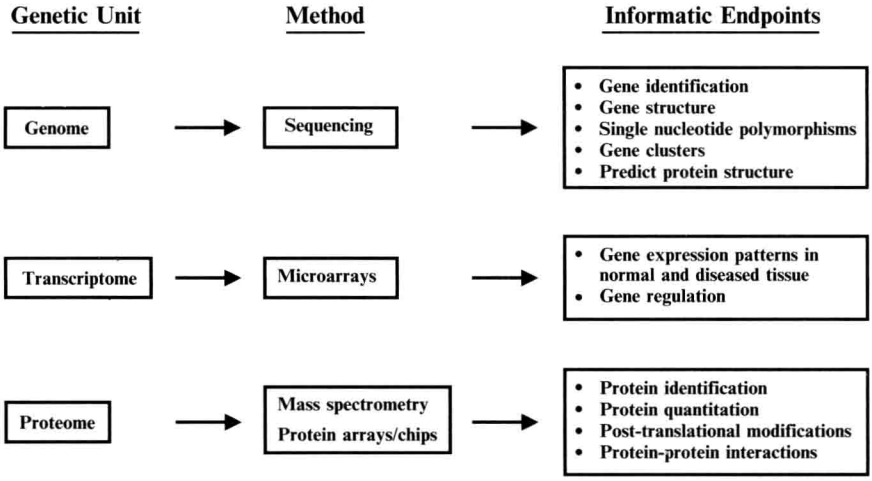

FIG. 1. Information obtained from genomic and proteomic analysis. Advances in genomic and proteomic analysis have facilitated characterization of the three major genetic units: the genome, the transcriptome, and the proteome. The data gleaned from each of these fields is in most cases unique to each genetic unit and, therefore, provide complementary information about the organization and regulation of living systems. The genome describes the entire set of genes that is encoded by the DNA of an organism and this has been obtained through a massive DNA sequencing effort. The transcriptome encompasses the entire complement of messenger RNA (mRNA) transcripts transcribed from the genome of a cell. These data have been obtained largely through the application of cDNA microarrays. The proteome describes the entire complement of proteins expressed by a cell at a point in time. Whereas two-dimensional (sodium dodecyl sulfate and isoelectric focusing) polyacrylamide gel electrophoresis has been the mainstay of proteomics analysis, the field is now embracing new techniques such as multidimensional capillary liquid chromatography coupled with tandem mass spectrometry (Link *et al.*, 1999; Washburn *et al.*, 2001), and single-dimension ultrahigh-resolution capillary liquid chromatography in combination with FTICR mass spectrometry (Jensen *et al.*, 1999). The development of high-throughput mass spectrometry methods and protein arrays will greatly accelerate the pace of proteomics research.

function of a gene is dependent on the activity of its translated protein, there has been significant impetus to develop methods that will enable high-throughput analysis of cellular proteomes. To understand the significance and impact of the rapid advances in analysis of the proteome, it is necessary to consider the utility and limitations of data obtained from analyses of the genome and transcriptome. Taken together, the study of the genome, transcriptome, and proteome provides complementary insights into a host of biological processes, and provides a greater understanding of the regulation of these processes.

II. Deciphering the Genome

Advances in nucleic acid sequencing and the software necessary to store and annotate sequence data have been instrumental in characterizing the genome of humans (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001) and other species (Fleischmann *et al.*, 1995; Blattner *et al.*, 1997; Goffeau *et al.*, 1996). The emphasis in the field of genomics has been to both generate and evaluate whole genome sequence data. The size and complexity of these genomic databases have necessitated the development of new informatic tools to organize and analyze data. The results obtained from this enormous effort will be substantial. The first major benefit of this work will be the complete identification and sequencing of the estimated 40,000 human genes that comprise the human genome (Venter *et al.*, 2001). Genes will be mapped to specific chromosomes, which will contribute to an understanding of normal development, the origin of phenotypic variability, and disease etiology and disease susceptibility in humans. Genomic data will also provide important information about intron and regulatory DNA sequences that influence such critical processes as mRNA transcription (Thieffry *et al.*, 1998) and mRNA splicing. In addition, homology alignment of new genome sequences with previously characterized genes will facilitate structural and functional predictions of expressed proteins (Gough *et al.*, 2001).

Although there is clearly much to be gleaned from genomic sequence data, interpreting the genome is complicated. It is well recognized that a single gene may encode multiple different proteins. Moreover, because coding regions are interspersed with noncoding regions of DNA and because there can be differential mRNA splicing, the genomic sequence cannot be used to reliably predict the entire spectrum of mRNA transcripts (transcriptome) or corresponding proteins expressed by a cell or tissue at any point in time.

III. Gene Expression Profiles

Most biological studies have been limited in scope to the analysis of individual mRNA transcripts or proteins. These studies employed such techniques as Northern blot analysis, RNase protection assays, reverse transcription–polymerase chain reaction, or Western blot analysis. The data derived from these techniques are specific and generally quantitative, but are limited to a small number of genes/proteins. Unfortunately, it is often difficult to appreciate how an individual gene or protein relates to an injury response, a signal transduction cascade, or a complex biological

state such as cancer. However, developments in DNA microarray technology have made it possible to simultaneously evaluate mRNA transcripts on a comprehensive level.

DNA microarrays are glass slides or nylon membranes to which cDNA sequences or oligonucleotides corresponding to select genes are affixed. Total or poly(A) RNAs are isolated from the cells or tissues being compared and reverse-transcribed to cDNAs. These are differentially labeled with fluorescent dyes or other markers. The labeled cDNAs are concomitantly hybridized to the array. The glass slides or nylon membranes are subsequently washed and scanned for intensity. A comparison of the two label intensities allows the relative expression levels of thousands of genes to be analyzed in a single experiment.

Alternatively, serial analysis of gene expression (SAGE) also provides a comprehensive and quantitative measure of gene expression (Velculescu *et al.*, 1995). SAGE is based on the generation of unique nucleotide sequence tags (10 base pairs) from a fixed position in each species of mRNA. The tags are initially prepared from mRNA that is transcribed into double-stranded cDNA and the frequency with which a tag appears in the cDNA pool reflects its relative abundance. Analyses of mRNA transcript levels, using either microarray or SAGE technology, are generally well correlated (Ishii *et al.*, 2000; Nacht *et al.*, 1999).

The most common research application of cDNA microarrays is gene expression profiling. Utilizing this approach, investigators have begun to identify subsets of genes associated with particular biological states (e.g., cancer) or that vary in response to different environmental conditions. With the completion of the Human Genome Project and through the ongoing annotation efforts, it will be possible to assess the entire subset of mRNAs (transcriptome) expressed in a tissue or cell of interest. It will also be possible to determine how the transcriptome of a cell or tissue changes with age, changing environmental conditions, or in response to injury and disease. Gene expression profiling has already proved effective in distinguishing between normal cells and tumor cells. Two subtypes of non-Hodgkin's lymphoma that could not be distinguished by traditional histological methods were distinguished by profiling 17,856 genes in patient samples (Alizadeh *et al.*, 2000). Distinct subtypes of malignant melanoma (Bittner *et al.*, 2000) and breast cancer (Perou *et al.*, 1999, 2000) have also been classified on the basis of gene expression profiling. Monitoring patterns of gene expression in malignant tissues is having a significant impact on the diagnosis and classification of many human cancers (DeRisi *et al.*, 1996; Golub *et al.*, 1999).

In a study aimed at understanding the molecular mechanisms that underlie the tumorigenesis and progression of clear cell renal cell

carcinoma (ccRCC), gene expression profiles of 29 ccRCC tumors obtained from patients with diverse clinical outcomes were analyzed with 21,632 cDNA-containing microarrays (Takahashi *et al.*, 2001). Gene expression profiles of each tumor sample were compared with cognate patient-matched normal tissue to identify gene expression alterations that occur in most ccRCCs. In addition, because all the experiments shared a "common" normal tissue reference, results from each experiment could be compared to identify gene expression patterns that correlated with differences in observed clinical features of the tumors. Changes in gene expression that were common to most of the ccRCCs studied and unique to clinical subsets were identified. There was a significant distinction in gene expression profiles between patients with a relatively nonaggressive form of the disease (100% survival after 5 years with 88% of the patients having no clinical evidence of metastasis) versus patients with a relatively aggressive form of the disease (average survival time of 25.4 months with a 0% 5-year survival rate). Approximately 40 genes, some of which have previously been implicated in tumorigenesis and metastasis, were identified. Moreover, the identified genes provide insight into the molecular mechanisms of aggressive ccRCC and suggest intervention strategies.

Many of the 40 genes that most effectively discriminated between patients with good outcome and those with poor outcome gave insight into the biology of the two groups of ccRCC (Takahashi *et al.*, 2001). *Sprouty*, the mammalian homolog of the *Drosophila melanogaster* angiogenesis inhibitor, was found to be exclusively upregulated in the good outcome group, which suggests that failure to properly inhibit angiogenesis may contribute to aggressive forms of ccRCC. Transforming growth factor (TGF-β), TGF-β receptor II (TGF-βRII), and its downstream effector, tissue inhibitor of metalloproteinase 3(TIMP3), were exclusively downregulated in the poor outcome group. Loss of the TGF-βII signaling pathway has previously been shown to contribute in aggressive cancer development (Engel *et al.*, 1999), and loss of TIMP3 expression by promoter methylation was shown to increase tumorigenicity (Bachman *et al.*, 1999). The identification of this pathway as downregulated in aggressive ccRCC suggests numerous targets for intervention to supplement the still low response rate of current adjuvant therapies.

There are a multitude of biological questions in addition to cancer that can be addressed by gene expression profiling. For example, this technology is being used to determine the molecular basis of apoptosis (Voehringer *et al.*, 2000) and to unlock the secrets of the aging brain (Lee *et al.*, 2000). Despite the utility that gene expression profiling provides, however, there are significant questions that cannot be answered by this

powerful technology. Genomics and gene expression profiling convey only limited information about the translated proteins that are ultimately encoded by the genome. The varied and complex properties of proteins cannot be reliably predicted by a simple linear readout of the genomic blueprint or the transcriptome.

IV. A Niche for Proteomics

Fortunately, the rapidly evolving field of proteomics (study of the proteome) is directed toward providing a comprehensive view of the characteristics and activity of every cellular protein. The proteome is clearly more complicated than the genome. The concept that one gene corresponds to one protein no longer holds true. A single gene can encode multiple different proteins. This can be attributed to (1) alternative splicing of the mRNA transcript, (2) the use of alternative translation start or stop sites, and (3) the occurrence of frame-shifting, during which a different set of triplet codons in the mRNA is translated. The net result of these activities is the generation of a proteome that contains many proteins derived from shared or overlapping genomic sequences (Fig. 2).

Another powerful impetus for moving beyond the transcriptome is the demonstration by several researchers that protein levels do not faithfully correlate with mRNA levels (e.g., Anderson and Seilhamer, 1997; Gygi *et al.*, 1999; O'Shaughnessy *et al.*, 2000). An analysis of 106 genes in the yeast *Saccharomyces cerevisiae* demonstrated that the levels of protein expression attributed to mRNA species of equal abundance could vary by as much as 30-fold. Conversely, the mRNA levels for proteins that were expressed at comparable levels varied as much as 20-fold. Experience from our own laboratory with cDNA microarray analysis yielded similar results. We identified a novel transcript in malignant mouse astrocytes, pescadillo, which was upregulated approximately 3-fold relative to nontransformed mouse astrocytes (Kinoshita *et al.*, 2001). Despite a 3-fold difference in the abundance of pescadillo transcripts, pescadillo protein levels were elevated more than 50-fold in the malignant mouse astrocytes (Y. Kinoshita, G. Foltz, J. Schuster, P. S. Nelson and R. S. Morrison, unpublished results). These results demonstrate that it is not always possible to predict changes in protein levels on the basis of changes in mRNA abundance.

One additional characteristic of proteins that is difficult to predict from genomic sequence data is the nature of their posttranslational modifications. In contrast to DNA and RNA, proteins can be modified

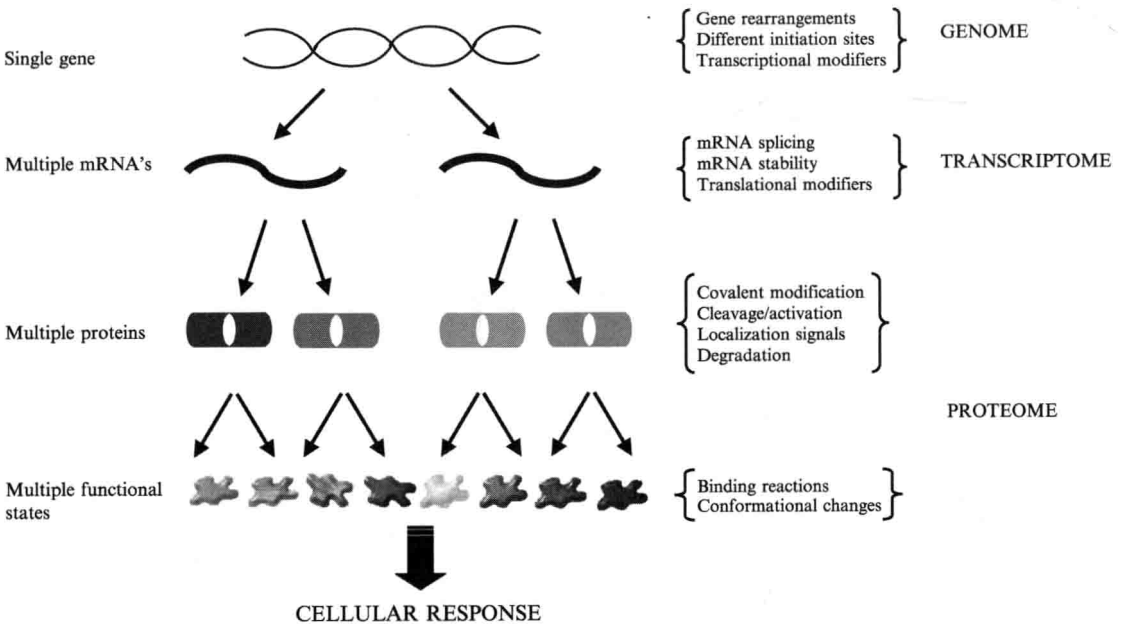

FIG. 2. The case for analyzing the proteome. At the level of the genome, a single gene may undergo rearrangement, or its transcription can be altered by several mechanisms to produce several different mRNAs. These mRNAs may then be alternately spliced, translated at varying rates, or differentially degraded to yield a potentially large and dynamically variable mRNA pool (transcriptome). Further complexity arises from a variety of posttranslational modifications and intermolecular interactions that yield additional functional protein states. Thus, dozens of unique functional protein states may be derived from a single gene. Because the process of protein production from DNA passes through so many levels of increasing complexity, it is difficult to predict the expressed complement of functional protein states (and the associated cellular response) from the genetic blueprint alone. Only a direct analysis of the proteome itself can answer these questions.

by phosphorylation, glycosylation, acetylation, nitrosation, poly(ADP-ribosyl)ation, ubiquitination, farnesylation, sulfation, linkage to glycophosphatidylinositol anchors, and SUMOylation (SUMO, small ubiquitin related modifier). In total, there are about 300 different posttranslational modifications that have been reported (Aebersold and Goodlett, 2001). These modifications can profoundly affect protein conformation, stability, localization, binding interactions, and function. Proteins are often modified at multiple sites, and it is not possible to predict from a sequence with complete certainty which sites will be modified in response to a specific set of conditions. The p53 tumor suppressor protein is a striking example of a protein that is modified at multiple sites in response to different stimuli. It is a nuclear phosphoprotein that is modified, in response to DNA damage, by the addition of phosphate to multiple seryl residues. The phosphorylation of certain seryl residues is required for p53-mediated transcription of several downstream targets associated with cell cycle arrest, including p21WAF1/Cip1 (p21) and mdm-2 (Jabbur *et al.*, 2000; Oda *et al.*, 2000). In contrast to these specific seryl residues, phosphorylation at other seryl residues regulates the transcriptional activation of apoptosis (Oda *et al.*, 2000). Moreover, the phosphorylation pattern of the p53 molecule can vary in an injury-dependent manner. Exposing cultured fibroblasts to nitric oxide induces a different pattern of p53 phosphorylation than exposure to γ irradiation, UV light, and doxorubicin (Adriamycin) (Nakaya *et al.*, 2000). p53 also requires multiple forms of posttranslational modification to manifest its activity. In addition to phosphorylation, conjugation of the ubiquitin-like molecule, SUMO-1, enhances the transcriptional activity of p53 (Gostissa *et al.*, 1999; Muller *et al.*, 2000; Rodriguez *et al.*, 1999). Although it is possible to analyze a genetic sequence for the presence of putative consensus sites for various posttranslational modifications, the mere presence of such sites does not indicate whether they are utilized, under what circumstances they are utilized, or if they are utilized simultaneously.

V. The Proteome: Greater than the Sum of its Parts

The disparity between mRNA levels and protein expression suggests that characterizing the proteome, rather than the transcriptome, under different conditions may provide a more accurate representation of the biological state of a cell. At the present time, evaluating the proteome is more difficult and labor intensive than characterizing the transcriptome. Two-dimensional (sodium dodecyl sulfate and isoelectric focusing)