Proteomics for Biological Discovery

Timothy D. Veenstra
Laboratory of Proteomics and Analytical Technologies
SAIC-Frederick, Inc.
Frederick, Maryland

John R. Yates
Department of Cell Biology
Scripps Research Institute
La Jolla, California
Proteomics for Biological Discovery
Proteomics for Biological Discovery

Timothy D. Veenstra
Laboratory of Proteomics and Analytical Technologies
SAIC-Frederick, Inc.
Frederick, Maryland

John R. Yates
Department of Cell Biology
Scripps Research Institute
La Jolla, California
Contents

CONTRIBUTORS vii
FOREWORD xi
PREFACE xv

PART I FOUNDATIONS OF PROTEOMICS 1

1. Mass Spectrometry: The Foundation of Proteomics 1
   Timothy D. Veenstra

2. Proteomic Analysis by Two-Dimensional Polyacrylamide Gel Electrophoresis 19
   Pavel Gromov, Irina Gromova, and Julio E. Celis

3. Isotope Labeling in Quantitative Proteomics 47
   Kristy J. Brown and Catherine Fenselau

4. Mass Spectrometric Characterization of Post-translational Modifications 63
   Thomas P. Conrads, Brian L. Hood, and Timothy D. Veenstra

5. Technologies for Large-Scale Proteomic Tandem Mass Spectrometry 91
   David L. Tabb and John R. Yates

6. Protein Fractionation Methods for Proteomics 111
   Thierry Rabilloud
 Contributors

CASSIO DA SILVA BAPTISTA
SAIC–Frederick, Inc.
National Cancer Institute at Frederick
Research Technology Program
Laboratory of Molecular Technology
Frederick, Maryland

KIRSTY J. BROWN
CTL Bio Services
Rockville, Maryland

JULIO E. CELIS
Department of Proteomics in Cancer
Institute of Cancer Biology
Danish Cancer Society and Danish Centre for Translational Research in Breast Cancer
Copenhagen, Denmark

G. MARIUS CLORE
Laboratory of Chemical Physics
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health
Bethesda, Maryland

THOMAS P. CONRADS
SAIC–Frederick, Inc.
National Cancer Institute at Frederick
Frederick, Maryland
AMY DAMBROWITZ
Department of Chemistry
University of Washington
Seattle, Washington

TRISHA N. DAVIS
Department of Biochemistry
University of Washington
Seattle, Washington

DON L. DEVOE
Department of Mechanical Engineering and Institute for System Research
University of Maryland
College Park, Maryland
and Calibrant Biosystems
Rockville, Maryland

NORMAN J. DOVICH
Department of Chemistry
University of Washington
Seattle, Washington

CATHERINE FENSELAU
Department of Chemistry and Biochemistry
University of Maryland
College Park, Maryland

PAVEL GROMOV
Department of Proteomics in Cancer
Institute of Cancer Biology
Danish Cancer Society and Danish Centre for Translational Research in Breast Cancer
Copenhagen, Denmark

IRINA GROMOVA
Department of Proteomics in Cancer
Institute of Cancer Biology
Danish Cancer Society and Danish Centre for Translational Research in Breast Cancer
Copenhagen, Denmark

BRIAN L. HOOD
SAIC–Frederick, Inc.
National Cancer Institute at Frederick
Frederick, Maryland
SHEN HU
Department of Chemistry
University of Washington
Seattle, Washington
Present address: UCLA School of Dentistry
Los Angeles, California

LEOPOLD L. ILAG
Department of Chemistry
University of Cambridge
Cambridge, United Kingdom

HALEEM J. ISSAQ
Laboratory of Proteomics and Analytical Technologies
SAIC–Frederick, Inc.
National Cancer Institute at Frederick
Frederick, Maryland

CHENG S. LEE
Department of Chemistry and Biochemistry
University of Maryland
College Park, Maryland 20742
and Calibrant Biosystems
Rockville, Maryland

YAN LI
Department of Chemistry and Biochemistry
University of Maryland
College Park, Maryland

DANIEL C. LIEBLER
Proteomics Laboratory
Mass Spectrometry Research Center
Vanderbilt University School of Medicine
Nashville, Tennessee

DANQIAN MAO
Department of Chemistry
University of Washington
Seattle, Washington

DAVID MICHELS
Department of Chemistry
University of Washington
Seattle, Washington

ERIC G. D. MULLER
Department of Biochemistry
University of Washington
Seattle, Washington
CONTRIBUTORS

David J. Munroe
SAIC–Frederick, Inc.
National Cancer Institute at Frederick
Research Technology Program
Laboratory of Molecular Technology
Frederick, Maryland

DaRue A. Prieto
Laboratory of Proteomics and Analytical Technologies
SAIC–Frederick, Inc.
National Cancer Institute at Frederick
Frederick, Maryland

Thierry Rabilloud
DRDC/BECP
CEA-Grenoble
Grenoble, France

Carol V. Robinson
Department of Chemistry
University of Cambridge
Cambridge, United Kingdom

David L. Tabb
Department of Genome Sciences
University of Washington
Seattle, Washington
Present address: Life Sciences Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Timothy D. Veenstra
Laboratory of Proteomics and Analytical Technologies
SAIC–Frederick, Inc.
National Cancer Institute at Frederick
Frederick, Maryland

John R. Yates
Department of Cell Biology
The Scripps Research Institute
La Jolla, California
Foreword

The cell can arguably be viewed as the basic unit of life, and a key focus of biological research is therefore to understand how cells are put together. What are the design principles through which the molecular constituents of the cell are organized? How do they respond dynamically to a changing environment, and how do they associate to form tissues and organs within a multicellular animal? Equally important and puzzling, how do a mere 5000 genes or so provide sufficient information to build a viable, free-living cell with remarkably complex properties, and how do a paltry 30,000 genes specify a human being, containing cells as diverse in their functions as a lymphocyte, a neuron, or a myocyte?

In considering these challenges, it is worth recalling how far we have come over the last thirty years, both technically and conceptually. Thirty years ago we couldn’t sequence DNA, molecular cloning was in its infancy, live imaging of cells was nonexistent, we had little understanding of the extent or functions of post-translational modifications of proteins, RNA splicing was not thought of, oligonucleotide-directed mutagenesis had not been conceived, the genetic manipulation of mammalian genomes was the stuff of science fiction, the term bioinformatics had not been coined, biological mass spectrometry and the yeast 2-hybrid system lay in the future, and solving a single protein structure was a Herculean effort. It was clear that an individual gene product, however, could have profound effects on many different aspects of cellular behavior.

This multieffect was especially evident for the proteins encoded by viral oncogenes that induce malignant transformation of cultured cells and tumors in vivo. The expression of a single oncoprotein such as v-Src, for example, causes changes in cell shape, adhesion, metabolism, growth, survival, and proliferation. This observation suggested that these distinct facets of cellular function must all be interconnected, be it directly or indirectly, and that it should be possible to define a logic that explains the inner working of the cell. For this enterprise, we need to know the complete coding potential of cellular genomes, and, more daunting, we need ways of globally investigating the expression, modifications, interactions, and
subcellular locations of their products. Furthermore, we need databases, computational tools, and modeling approaches to collate and interpret this information, to investigate how cellular networks function to generate complex properties, and to provide new hypotheses regarding biological function that can be tested experimentally. This new area of science is not only explanatory in nature. By understanding the basic principles of cellular design, we can learn to reengineer cell signaling networks, and thus to endow cells with new properties. This synthetic approach to biology may be especially valuable in the treatment of diseases such as advanced cancer, in which the normal organization of cells and tissues becomes severely deranged, and in infection, in which there are complex interactions between the pathogen and the host. First, however, to quote an old recipe for rabbit stew, “catch your rabbit” (or in this case catch your proteins).

The extensive sequencing of cDNAs and genomic DNA has now given us a fairly comprehensive account of the protein coding potential of prokaryotic and eukaryotic genomes, although for most of the predicted proteins there remains a significant degree of uncertainty about their true identity, their splice variants, their functions, and their regulation. Nonetheless, we can argue that we have in hand an increasingly complete set of the protein building blocks through which cells are assembled. The primary amino acid sequence of a protein can potentially give us a large amount of information, in part because proteins are commonly constructed in a cassette-like fashion from multiple smaller domains with characteristic conserved sequences. These domains can have either an enzymatic activity (such as a protein kinase) or a binding function (such as a phosphotyrosine-binding SH2 domain) and have been used repeatedly in a wide range of proteins. It seems as though cells and organisms may have evolved primarily through the increasingly sophisticated use of a limited set of protein domains, joined in increasingly elaborate combinations, and have only occasionally resorted to the invention of entirely new biochemical functions. Thus, the presence of particular domains in a protein of unknown biological function can give us strong clues as to its physiological properties. In addition, as we better understand the abilities of signaling enzymes to modify their intracellular targets at specific sites, and learn the rules determining the binding of interaction domains to defined peptide motifs, we can search the proteome in silico for potential substrates and binding partners of a protein of interest. In silico analysis, however, cannot replace experimental analysis, and fortunately there has been a veritable revolution in our capacity to analyze protein expression, post-translational modifications, and interactions, that in aggregate has led to the burgeoning field of proteomics, a discipline that is proving essential for our understanding of cell biology.

Genome sequencing and associated techniques, such as microarray analysis of RNA expression, have as their underlying theme that cells and organisms cannot be fully understood by studying one gene or transcript at a time. This statement is especially true at the level of the proteome, which presents challenges with a heightened degree of difficulty related to its dynamic nature. Proteins are not equally stable; some have a half-life of a few minutes, while others persist for days. Indeed, a large family of proteins is dedicated to the selection of specific polypeptides for ubiquitination and degradation, often in response to changing cellular conditions. In addition, since a single gene can potentially encode many different products, each of which may have a distinct function, one must identify not simply an individual protein but the complement of related splice variants expressed in a particu-
lar cell or tissue. To complicate matters, proteins can undergo a number of modifications, such as phosphorylation, acetylation, methylation, hydroxylation, ubiquitination, and nitrosylation, among others. These modifications can alter a protein's enzymatic activity but also serve as switches to induce or antagonize modular protein–protein interactions and thus the assembly of regulatory complexes. This complexity is the tip of the proteomic iceberg, since a single protein can be modified simultaneously by several different groups, with each combination of modifications potentially generating a distinct biological function. This multimodification phenomenon has been studied intensively in the context of proteins such as histones and p53, but there is every reason to suppose that it is the norm rather than the exception. Adding to the complexity, a single modification, such as ubiquitination, can come in several different flavors. Addition of a single ubiquitin to a lysine residue in a target protein creates a binding site for interaction domains, such as the ubiquitin interaction motif found in proteins involved in endocytosis and intracellular trafficking. The linking of further ubiquitins to the initial site of modification to form a polyubiquitin chain, however, can lead to recognition by the proteasome and degradation.

Fortunately, powerful new proteomic techniques have been introduced just at the moment when they are most needed to address these issues. The forerunner of this approach, still extraordinarily useful, is the yeast 2-hybrid technique, which allows the investigator to measure binary protein–protein interactions within the confines of the yeast cell. While initially used to search a library for binding partners of a single protein, it has more recently been employed for comprehensive screens involving entire proteomes or large subsets thereof. An interesting lesson from these efforts is that the use of orthogonal techniques can greatly increase the reliability of proteomic data. For example, combining a 2-hybrid screen involving all 28 yeast SH3 domains with data concerning their binding preferences for peptide motifs, identified by phage display analysis, has yielded a more reliable view of the interaction network controlled by SH3 domains in a yeast cell.

A parallel technique of exceptional power involves the use of mass spectrometry (MS) to analyze proteins, either in their intact state or, more commonly, following peptide digestion. Peptide fragmentation can give sufficient sequence information to unambiguously identify a protein by MS, by comparison with a database of potential products inferred from DNA sequence information. Through the use of isotopic labeling and selective modification with reagents such as isotope-coded affinity tags (ICATs), it is possible to use MS to compare protein expression and modifications in two related cell samples, and the use of an isotopically labeled reference peptide allows for quantitation of protein levels. In addition, through the affinity isolation of one protein, it is possible to identify its associated polypeptides, as demonstrated through analysis of the yeast interaction map (Ho, Gruhler, Heilbut et al. 2002. Nature 415: 180–183). While this latter approach has typically involved gel purification of the complex protein mixture prior to analysis, advances in peptide separation have enabled the use of gel-free techniques to analyze protein complexes, which will potentially enhance the speed and completeness with which sets of interacting proteins can be identified. This advancement will be a necessity as we approach the more complex proteomes of mammalian cells.

These advances have given us an unprecedented ability with which to explore the expression and modifications of cellular proteins and to establish a wiring diagram of the cell. To be truly useful, such proteomic data must be linked to...