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, Galifornia

WILEY-LISS
A John Wiley & Sons, Inc., Publication

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, Galifornia

WILEY-LISS
A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data is available.

ISBN-13: 978-0-471-16005-2
ISBN-10: 0-471-16005-9

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Contributors

Cassio Da Silva Baptista
SAIC–Frederick, Inc.
National Cancer Institute at Frederick
Research Technology Program
Laboratory of Molecular Technology
Frederick, Maryland

Kristy 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. Dovichi
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 Cantre 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

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 proteosome 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

functional analysis. In yeast this can readily be accomplished through the use of genetic deletion sets, in which each open reading frame has systematically been disrupted. Among other things, this tool has allowed a high-throughput screen of genetic interactions to complement the data regarding physical protein–protein interactions. In mammalian cells, gene targeting is much more laborious than in yeast, but short interfering (si) RNAs provide a powerful tool with which to down-regulate gene expression in cultured cells and in intact animals. Indeed, siRNAs can be used to analyze families of proteins and are being employed in genome-wide screens for proteins that control specific aspects of cellular function. Another source of functional information involves the use of genetically encoded fluorescent protein derivatives to track protein localization in live cells. This technique is suitable for automation and provides a further level of annotation regarding protein activity.

Finally, we need computational and mathematical tools to synthesize these data and build models that illuminate cellular organization and complexity. An important advantage should be our ability to compare data from different species, in an effort to identify common threads and significant differences between the proteomes and interaction maps of distinct organisms. We are entering a new era in biology, one in which we will finally have the tools and the data to understand how cells work, and what goes wrong in disease. Proteomics is at the forefront of this revolution.

Preface

Proteomics has come a long way. It was not that long ago that the mention of the term proteomics brought up images of a two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) separation and a mass spectrometer. Tim Veenstra can remember, as a post-doctoral researcher in a molecular and cellular biology laboratory, his supervisor, Dr. Rajiv Kumar, discussing with him a company that could take cell extracts from control and 1,25-dihydroxyvitamin D_3 neural cells and separate them out on 2D-PAGE gels. Protein spots that were stained differentially when comparing the two gels could then be identified. "Being somewhat obtuse, I never realized until several years later that Dr. Kumar was proposing a classical proteomics study before the term even became popular," relates Tim. Times have quickly changed, as scientists from various areas of life sciences have come to embrace the capabilities of the ideas and technologies that have been developed in the field of proteomics.

While there are many reasons, the diversity of chapters presented in this book best describes why proteomics has become so popular. Obviously, no book on proteomics would be complete without a chapter describing proteomic analysis using 2D-PAGE. The remaining chapters, however, span a diverse group of disciplines. These disciplines include gel-free proteomics to measure protein abundances, characterization of intact noncovalent protein complexes, three-dimensional protein structure determination, protein localization by imaging techniques, and the emerging field of protein arrays. Proteomics would not exist as it does today without three other critical elements described in this book: separations, bioinformatics, and automation.

The increasing number of fields that proteomics touches will not plateau any time soon. We are just beginning to realize the potential to look at biological systems as a whole instead of individualized parts. Proteomics, along with genomics, transcriptomics, and metabolomics, will play a major role in this next scientific venture. The impact of proteomics in clinical research is just in its infancy. Clinical research is a vast area in which proteomics can have a huge impact both in the

discovery of new diagnostics and therapeutics and also at the individual patient level by determining individually tailored courses of treatment. New ideas and technologies continue to be developed rapidly to meet these challenges.

The editors would like to thank the authors for the time and effort they put into their contributions. All authors were selected because they are recognized as leaders in their particular areas of proteomics and their special knowledge and experience are clearly reflected in each chapter. The editors are also grateful to the publisher, John Wiley & Sons, Inc., for patience in understanding the enormity of work required to put such a book together. We are confident that readers of this book will be enriched by insights provided by each of the authors.

Frederick, Maryland
La Jolla, California

TIMOTHY D. VEENSTRA
JOHN R. YATES

Part I

Foundations of Proteomics

1

Mass Spectrometry: The Foundation of Proteomics

Timothy D. Veenstra

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

1.1 INTRODUCTION

Scientific direction can be driven by many factors. Obviously, science is still primarily hypothesis driven; however, the continuing technology developments have enabled a greater focus on discovery driven science. Hypothesis driven science formulates a question and then uses whatever technology is available to acquire the information necessary to answer that question. In contrast, discovery driven science collects the information first and then determines the questions (or answers) that can be formulated from the available data. While it may seem to function through a "shot-in-the-dark" mentality, present technological developments make discovery approaches quite logical. Never before in the history of science has there been the capacity to acquire the wealth of data on biological molecules as exists today. A great example of this data gathering capability is reflected within the human genome project. It was inconceivable two decades ago that sequencing of the entire human genome could be accomplished; yet here we are today with the capability of sequencing genomes of other organisms as a routine procedure. Fortunately, science was not content with being able to sequence genomes and soon after the capability to obtain global measurements on the relative abundances of gene transcripts was established. This capability has naturally progressed to the development of technologies to perform discovery driven studies on entire proteomes. This stage does not even represent the end of development, as significant progress is being made in metabolomics.

Proteomics for Biological Discovery, edited by Timothy D. Veenstra and John R. Yates.

The term proteomics has evolved over the past few years to almost replace what was once referred to as protein chemistry. The original, and still classical, connotation of proteomics, however, is the characterization of the complete set of proteins encoded by the genome of a given organism (Wilkins et al., 1996). In the early history of proteomics, proteins were fractionated by two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) followed by visualization using protein stains such as Coomassie or silver stain (O'Farrell, 1975). To identify differences in the protein abundances of two distinct samples, each of their proteomes is fractionated and visualized on separate gels and those spots that reveal differences in their staining intensity are cored from the gel and identified, typically using mass spectrometry (MS). While it has been around for decades, the ability to use MS to characterize proteins has been the single largest force that has propelled proteomics. Many different facets of MS have led to its prominent position within the field of proteomics. The sensitivity of MS allows for the routine identification of proteins in the femtomole (fmol, 10^{-15} mol) to high attomole (amol, 10^{-18} mol) range (Moyer et al., 2003). The ability to identify proteins with confidence is aided by the mass measurement accuracy available using current MS technology. This accuracy is typically less than 50 parts per million (ppm) and is often routinely less than 5 ppm (Pasa-Tolic et al., 2004). The ability of tandem MS (MS/MS) to obtain partial sequence information in combination with on-line fractionation enables the confident identification of complex mixtures of peptides (Nesvizhskii and Aebersold, 2004). The throughput by which proteins can be identified by MS is unparalleled by any other biophysical technique—a critical parameter in the use of any technology to gather large datasets.

While used to characterize proteins, in reality it is peptides that MS is most adept at identifying. In a great majority of proteomics studies, the complex mixture of proteins is made even more complex by digesting the proteins into smaller peptides prior to MS analysis (Rappsilber and Mann, 2002). This digestion step is optimal for two main reasons. First, overall solubility of peptides in solution is much greater than that of intact proteins. Second, even though the mass measurement accuracy of MS is high, it is still not sufficient to confidently identify a protein *de novo* based solely on its molecular weight. Therefore, proteins are typically identified through peptides acting as surrogates for their parent protein of origin. One of the most common ways of identifying a protein is based on the mass spectrum of its peptide fragments that are produced by digestion using an enzyme such as trypsin. The resulting spectrum obtained from such a sample is referred to as a "peptide map" or a "peptide fingerprint" (Blackstock and Weir, 1999). To identify the protein, the collection of measured masses is compared to *in silico* peptide maps derived from a protein or genomic database (Figure 1.1). To identify a single protein within a simple mixture, peptide mapping works very well and it is quite easy to acquire the data necessary for obtaining the desired result. Peptide mapping of proteins within complex mixtures such as cell lysates is not possible since the peptide masses recorded in the mass spectrum will arise from a large number of different species and will not provide a conclusive identification. Fortunately the available instrumentation enables a greater depth of information to be obtained from peptide masses observed by MS. Instead of relying on the accurate mass of a specific peptide,

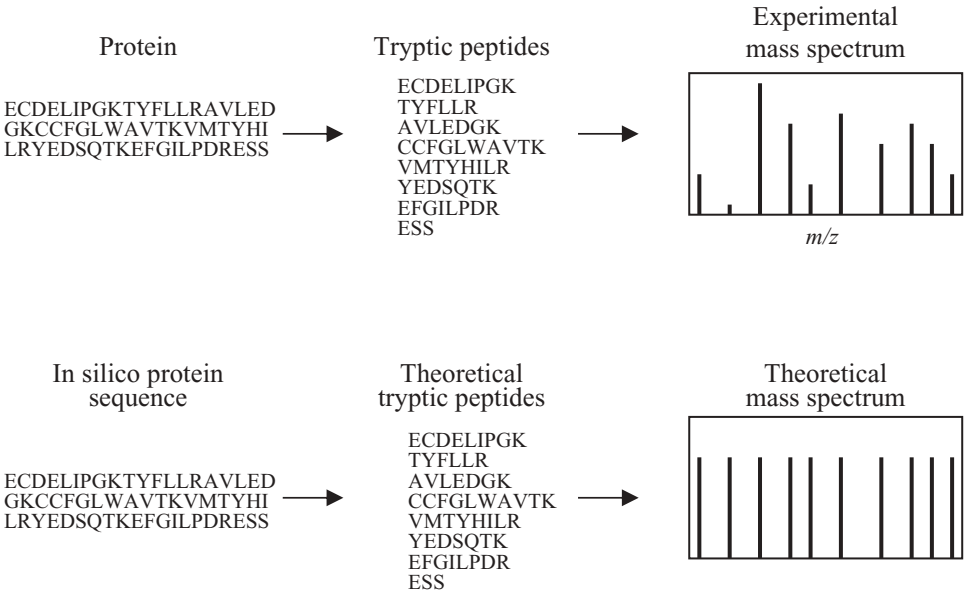

Figure 1.1. *Peptide mapping for protein identification. In peptide mapping, the protein of interest is proteolytically digested and the masses of the proteolytic peptides are measured using mass spectrometry (MS). To identify the correct protein, the sequences of all proteins within a specified database are digested in silico based on the specificity of the proteolytic enzyme used. The masses of the resulting peptides are calculated and theoretical mass spectra are constructed. The protein is identified based on the closest match between the experimental mass spectrum and the theoretical mass spectrum.*

individual peptide ions can be isolated and fragmented by collision induced dissociation (CID). After fragmentation of the peptide, the masses of the fragment ions are recorded and used to obtain partial or complete sequence information, as shown in Figure 1.2. This process is more commonly referred to as tandem MS or MS/MS (Martin et al., 1987). When peptides are subjected to MS/MS, they are not completely obliterated into their constituent amino acids, but instead an ensemble of fragments containing various lengths of the peptide is obtained. This information provides "sequence ladders" that enable partial primary sequence information of the peptide to be deduced. The raw data is then analyzed using software programs that can compare the experimental data to *in silico* MS/MS mass spectra calculated from the protein sequences in the database (Chamrad et al., 2004).

Proteomics is conducted for many different purposes and at many different levels. Fortunately there are several different types of spectrometers available depending on the focus of the research being conducted. Obviously, if an investigation is focused on identifying simple protein mixtures, the instrument requirements would be different than if entire cell or tissue lysates were the sample of interest. In the following, a description of the various types of MS instrumental platforms available will be discussed with a focus on their application and mode of operation.

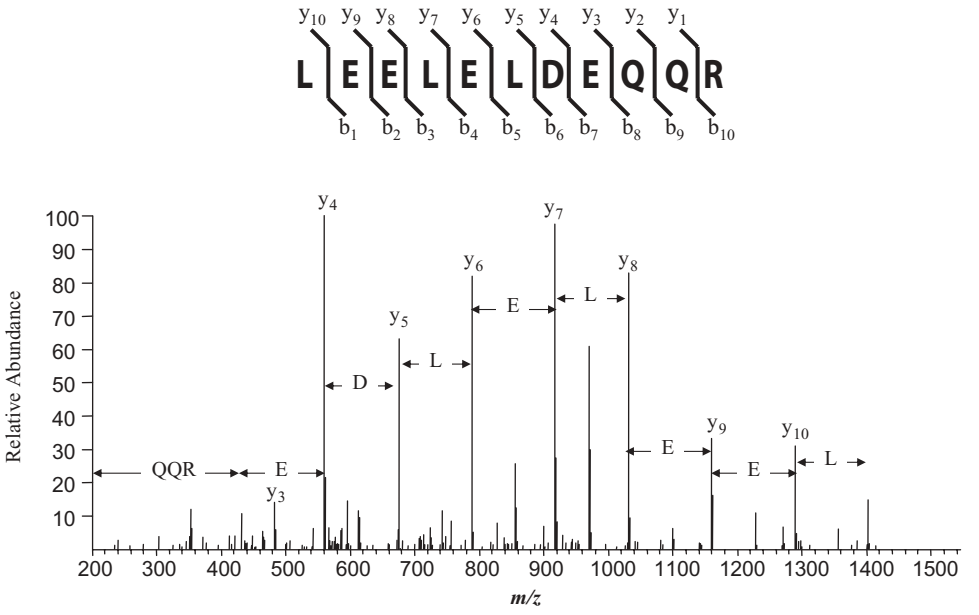

Figure 1.2. *Tandem mass spectrometry (MS/MS) spectrum of a peptide observed from a tryptic digest of mitogen activated protein kinase kinase (MAPKK). Partial primary sequence information is determined by comparing the differences between major peaks in the spectrum with the calculated molecular masses of the amino acid monomers within the peptide.*

1.2 IONIZATION METHODS

The mass spectrometer is made up of two major components: the ionization source and the mass analyzer. It is within the ionization source that the sample of interest is ionized and then desorbed into the gas phase. The mass analyzer acts to guide the gas-phase ions through the instrument to the detector. At the detector, the ions mass-to-charge (m/z) ratios are measured. While sometimes overlooked, many of the developments that have led to MS having a major impact on proteomics have been the invention of new ionization techniques.

The two most common methods to ionize biological molecules prior to their entrance into the analyzer region of the mass spectrometer are matrix-assisted laser desorption ionization (MALDI) (Karas et al., 1987) and electrospray ionization (ESI) (Fenn et al., 1989). While ESI and MALDI have enabled significantly larger proteins (i.e., greater than several hundred thousand daltons) to be analyzed, their greatest impact still remains in the analysis of peptides generated from proteolytic digests of larger species. One of the more significant advances enabled by ESI was the ability to interface separation methods such as liquid chromatography (LC) with MS. While separations are not discussed in this chapter, MS-based proteomics as it is practiced today would not be possible without the concurrent development of chromatographic and electrophoretic separation techniques.

1.2.1 Electrospray Ionization

ESI greatly enhanced the ability to characterize proteins and peptides by MS. Malcolm Dole, who conceived of using an electrospray process to produce intact high mass polymeric ions, provided the first description of ESI. He gained this insight from his knowledge of electrospraying automobile paint (Dole et al., 1968). These first experiments provided the basis of further studies by John Fenn (Fenn et al., 1989), who extended the use of ESI to measure biological molecules and was awarded the Nobel Prize in chemistry in 2002 for his discoveries.

The mechanism by which ESI works is relatively simple. ESI requires the sample of interest to be in solution so that it may flow into the ionization source region of the spectrometer (Figure 1.3A). Particulates or other insoluble entities in the

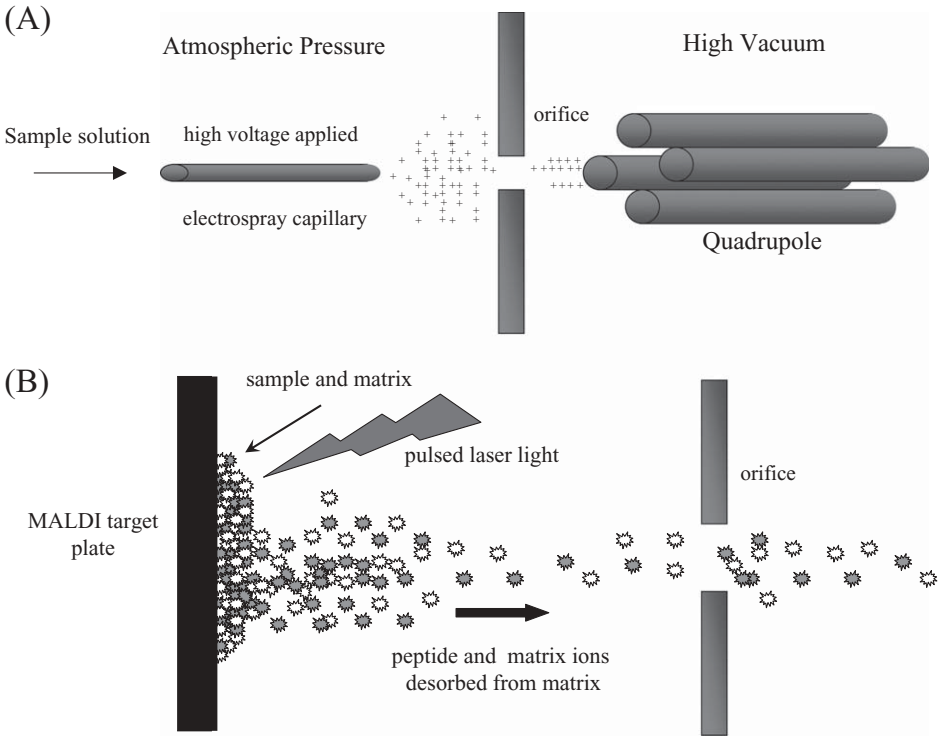

Figure 1.3. *(A) Electrospray ionization (ESI) of molecules for mass spectral characterization. The sample solution is passed through a stainless steel or other conductively coated needle. A high positive potential is applied to the capillary (cathode), causing positive ions to drift toward the tip with high voltage. The presence of a high electric field produces submicrometer-sized droplets upon the solution exiting the needle. The droplets travel toward the mass spectrometer orifice at atmospheric pressure and evaporate and eject charged analyte ions. The desolvated ions are drawn into the mass spectrometer by the relative low pressure maintained behind the orifice. (B) Principles of matrix-assisted laser desorption ionization (MALDI). The sample is cocrystallized with a large excess of matrix. Short pulses of laser light are focused onto the sample spot, causing the sample and matrix to volatilize. The matrix absorbs the laser energy, causing part of the illuminated substrate to vaporize. A rapidly expanding matrix plume carries some of the analyte into the vacuum with it and aids the sample ionization process. (See color insert.)*

sample will hamper ionization and cause the capillary through which the sample flows to become clogged. To ionize the sample, high voltage is applied to a stainless steel or other conductively coated needle through which the sample is flowing. The voltage results in charges being added to the sample, creating an ion that can be guided through the analyzer region of the instrument. The applied voltage can result in the sample becoming positively or negatively charged; however, positive ionization is used primarily in the analysis of proteins and peptides. As it exits the spray tip, the solution produces submicrometer-sized droplets containing both solute and analyte ions. The sample is then desorbed of solute prior to entering the analyzer region of the instrument. This desorption is achieved by evaporation of the solvent by passing the sample through a heated capillary or a curtain of drying gas, typically nitrogen. Since the desolvation of the ions occurs at atmospheric pressure and the mass analyzer region of the spectrometer is maintained at a lower pressure, the ions are drawn into the spectrometer based on this pressure differential.

What distinguished ESI from other ionization methods is its ability to produce multiply charged ions from large biological molecules. The number of charges that can be accepted by a particular molecule is dependent on many factors including its basicity and size. Depending on their size and the number of basic residues within, peptides typically exist as either singly, doubly, or triply charged ions. Since trypsin is the most commonly used protease in proteomics today, peptides are typically observed in both 1+ and 2+ charged states owing to the basic sites on the N terminus and the C-terminal lysine or arginine residues.

1.2.2 Matrix-Assisted Laser Desorption Ionization

Matrix-assisted laser desorption ionization (MALDI) is another "soft" ionization process that generates ions by irradiating a solid mixture with a pulsed laser beam. The solid mixture is comprised of the analyte of interest dissolved in an organic matrix compound. The laser pulse both indirectly ionizes and desorbs the analyte molecules from the solid mixture. A short-pulse (2–200 Hz) ultraviolet (UV) laser is typically used in MALDI; however, infrared irradiation has also been used (Tanaka et al., 1988; von Seggern et al., 2003). To prepare the solid mixture, an equal volume of the sample solution is combined with a saturated solution of matrix prepared in a solvent such as water, acetonitrile, acetone, or tetrahydrofuran. The matrix is a small, highly conjugated organic molecule (i.e., α-cyano-4-hydroxycinnamic acid (CHCA), 2,5-dihydroxybenzoic acid (DHB), and 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid)) that strongly absorbs energy in the (UV) region. A few microliters of the solid mixture is placed onto a MALDI target plate and allowed to dry. This drying procedure results in the incorporation of the peptides into a crystal lattice. The MALDI target plate is then inserted into the source region of the mass spectrometer followed by laser irradiation, as shown in Figure 1.3B. The MALDI source region of most spectrometers is maintained at a relatively high pressure, causing the ions to be drawn into the mass analyzer region of the instrument, which is maintained at a lower pressure. A recent development has been the design of MALDI sources that operate at atmospheric pressure (Moyer and Cotter, 2002). This ability to operate at atmospheric pressure enables MALDI sources to be interfaced to analyzers, such as ion traps and quadrupole

time-of-flight analyzers. Such instruments have historically been interfaced with ESI sources.

Similar to ESI, MALDI can produce both positive and negative ions. Positive ions, which are typically the species of interest in peptide analysis, are formed by the acceptance of a proton as the analyte leaves the matrix. While yet to be absolutely determined, the prevailing theory is that analyte ionization occurs within the dense gas cloud that forms and expands supersonically into the vacuum region of the spectrometer. The analytes are protonated (or deprotonated) through collisions between analyte neutrals, excited matrix ions, and protons and cations. In MALDI, most analytes accept a single protein; therefore peptide and large biomolecular ions are singly charged. This singly charged character results in some molecules having large m/z values and therefore MALDI is typically interfaced with mass analyzers with large m/z ranges, such as time-of-flight (TOF) spectrometers.

1.2.3 Desorption Electrospray Ionization

While not yet applied to proteomic technology, a new method of desorption ionization has recently been described that allows the direct analysis of surfaces by MS. This ionization technique, called desorption electrospray ionization (DESI), was developed in the laboratory of R. Graham Cooks (Takats et al., 2004) and is illustrated in Figure 1.4. In this technique, electrosprayed droplets are directed toward a surface to be analyzed. The droplets produce gaseous ions of the material on the surface and these ions are sampled with a mass analyzer. The mass analyzer is equipped with an atmospheric interface connected via a flexible and extended ion transfer line. This ionization technique, while extremely new, has shown the

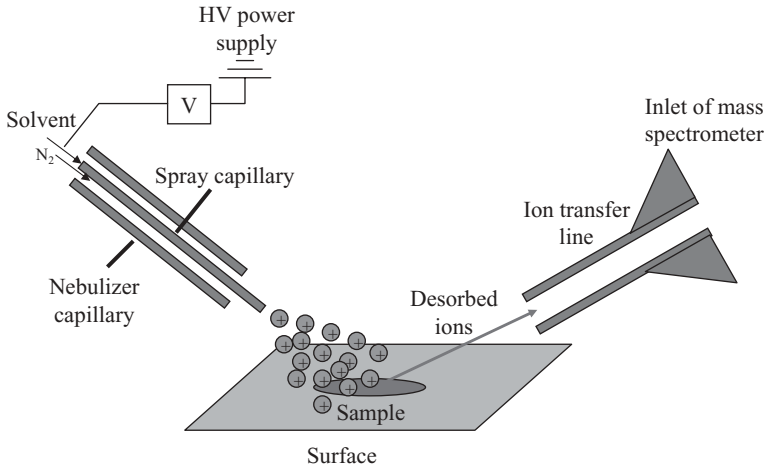

Figure 1.4. *Schematic of desorption electrospray ionization (DESI) instrument. In DESI, electrosprayed droplets are directed to a surface. The impact of the charged droplets produces gaseous ions from the sample on the surface, which can be sampled using a commercially available mass analyzer equipped with an ion transfer line. (See color insert.)*

capability of analyzing a range of compounds from nonpolar small molecules to polar peptides and proteins.

While the most fruitful uses of this new ionization technology are not clear, some of the demonstrated applications suggested a new exciting way to monitor things like drug distribution and surface analysis. In a novel experiment, 10 mg of loratadine, an over-the-counter antihistamine, was given to a patient and 40 minutes later DESI was able to detect the molecule directly from the skin surface and saliva of the individual. While the proteomic applications using this technology have not been clearly demonstrated, the potential exists for direct monitoring of proteins on the surface of cells in culture from tissue sections.

1.3 MASS ANALYZERS

1.3.1 Ion-Trap Mass Spectrometer

An ion-trap mass spectrometer functions just as its name implies: it traps ions. The ability to trap ions, however, does not explain the popularity of this mass analyzer for proteomic analysis. The popularity lies in the discovery and development of ways to manipulate the ions after they are trapped (Stafford et al., 1984). The first development was the mass-selective *instability* mode of operation. This mode allowed ions that were created and trapped over a given time period to be ejected sequentially into a conventional electron multiplier detector. Unlike the mass-selective *stability* mode of operation where only one *m/z* value could be stored, a wide range of *m/z* values could be stored. The second big development was showing that the addition of 1 mtorr of helium gas to the ion trap increased the mass resolution of the instrument. This increased resolution results from a reduction in the kinetic energy of the ions and causes the ion trajectories to contract to the center of the trap (Stafford et al., 1984). This phenomenon causes packets of ions of a given *m/z* to form, allowing them to be ejected faster and more efficiently than a diffuse cloud of ions.

The ion trap conducts repeated iterations of collecting, storing, and ejecting ions out of the trap. The true power of the ion-trap analyzer is its ability to isolate and fragment peptide ions (i.e., conduct MS/MS) from complex mixtures, such as found in many proteome analyses. To perform MS/MS analysis, specific ions are selected and the trapping voltages are adjusted to eject all other ions from the trap. The applied voltages are then increased to cause an increase in the energy of the remaining ions. These high-energy ions undergo collisions with He_2, causing them to fragment. These fragments are then trapped and scanned out according to their *m/z* values. Daughter ions resulting from the fragmentation of large ions can also be retained within the trap and subjected to further rounds of MS/MS (i.e., MS/MS/MS or MS^n), to obtain more structural information concerning the species of interest. While MS/MS/MS is not routinely used in the identification of peptides in a complex mixture, it has shown utility in the identification of phosphorylated peptides.

The ion-trap mass spectrometer enjoys a position of prominence as a true "workhorse" in global proteomic studies designed to characterize complex mixtures of proteins. When analyzing very complex mixtures, the ion trap operates

in a data-dependent MS/MS mode in which each full MS scan is followed by a specific number (usually 3–5, but sometimes >10 when using a linear ion trap) of MS/MS scans, where the three most abundant peptide molecular ions are dynamically selected for fragmentation. Using such a mode of operation will often allow for the identification of over 1000 peptides in a single LC/MS/MS experiment.

1.3.2 Time-of-Flight Mass Spectrometer

Time-of-flight mass spectrometers (Olthoff et al., 1988) are an extremely popular choice of mass analyzer for proteomic research. The major attributes that make TOF-MS attractive are their high throughput, sensitivity, and resolution. TOF spectrometers measure the *m/z* ratios of ions based on the time it takes for the ions generated in the source to fly the length of the analyzer and strike the detector. The speed, and therefore the time, at which the ions fly down the analyzer tube is proportional to their *m/z* value. Larger ions have a slower speed compared to smaller ions and therefore take a longer time to reach the detector.

TOF analyzers have been used primarily to generate peptide fingerprints for identifying individual proteins. The simplicity of operation and robustness of MALDI-TOF analyzers have made them an excellent choice for such applications. With the development of MALDI-TOF/TOF instruments, these analyzers have been able to do true tandem MS through the inclusion of a collision cell separated by two TOF tubes. MALDI-TOF/TOF instruments are characterized by high throughput and high resolution and mass accuracy for both the MS and MS/MS modes. In this configuration, peptide ions generated in the source region are accelerated through the first TOF tube and are dissociated by introducing an inert gas (i.e., air or nitrogen) into a collision cell. Collisions between the gas and peptide ions cause fragmentation of the peptide. These fragment ions are then accelerated through a second TOF tube to the detector. This combination allows proteins to be identified through peptide fingerprinting and identification is confirmed through MS/MS of selected peptide species. In addition, proteins within complex mixtures can now be identified solely through MS/MS of specific peptide signals. Many standard TOF instruments do not contain a true collision cell to provide MS/MS sequence data. Instruments equipped with a reflectron, however, can measure fragmentation products through a process called “post-source decay” (PSD) (Kaufmann, 1995). In PSD, the reflectron voltage is adjusted during the analysis so that fragment ions generated during the ionization and acceleration of the peptide are focused and detected. Specifically, PSD produces immonium ions, which are useful indicators of the presence of a specific amino acid within a peptide (Kaufmann et al., 1996). While PSD analysis can be relatively slow and will not meet the high-throughput demands necessary for proteomics, it does provide useful complementary information to substantiate the identification of an intact peptide.

1.3.3 Triple Quadrupole Mass Spectrometer

The quadrupole mass spectrometer has been the most commonly used mass analyzer with ESI (Yost and Boyd, 1990). As its name implies, a quadrupole consists of four metal rods arranged in parallel, as shown in Figure 1.3A. Direct current

and radiofrequency (rf) voltages are applied to these rods to guide and manipulate ions through the mass analyzer. Altering the voltage allows a specific *m/z* range of ions to pass through the quadrupole region of the analyzer and onto the detector. The two most common types of quadrupole mass spectrometers are single-stage and triple quadrupoles. Unfortunately single quadrupole analyzers have limited use in proteomics since they lack true MS/MS abilities, although in-source CID is possible. The triple quadrupole instrument, however, has true MS/MS capabilities since a collision cell is incorporated between two of the quadrupole regions.

To identify peptides using a triple quadrupole mass spectrometer requires switching the analyzer between two different scan modes. In the first "full-scan" mode, a broad *m/z* range of ions is allowed to pass through the first quadrupole. The ions that pass through are allowed to pass freely through the remaining two quadrupoles onto the detector. Essentially all of the ions produced in the source are measured. In the second scan mode, the first quadrupole is used as a mass filter and only a specific ion is allowed to pass through. The ion that is allowed through is then subjected to fragmentation within the second quadrupole by this region being filled with an inert gas. The resulting fragmentation ions then freely pass through the third quadrupole and are detected.

The versatility of the triple quadrupole analyzer is underscored by its ability to produce an ion, precursor ion, and neutral loss scanning. Triple quadrupoles have been used to identify proteins extracted from 2D-PAGE gels (Kuhn et al., 2004), in phosphopeptide characterization (Kocher et al., 2003), and in glycopeptide identification (Jiang et al., 2004). With a mass measurement accuracy of 0.5 amu, the fragment ions produced by MS/MS using triple quadrupole analyzers is sufficiently accurate to allow the identification of peptides by correlating the spectra with protein sequences obtained from biological databases.

1.3.4 Quadrupole Time-of-Flight Mass Spectrometer

The hybrid quadrupole time-of-flight mass spectrometer (QqTOF) is a versatile instrument that plays a key role in proteomic analysis (Chernushevich et al., 2001). The QqTOF mass spectrometer combines a mass-resolving quadrupole and collision cell with a TOF tube, as shown in Figure 1.5. This instrument combines the benefits of both types of mass analyzers: the ion selectivity and sensitivity of the quadrupole and the high mass resolution and mass accuracy of the TOF (Shevchenko et al., 2000). The high mass accuracy afforded with this configuration provides the potential for real *de novo* sequencing of peptides (Loboda et al., 2000). This instrument is also able to analyze samples using both ESI and MALDI methods, using sources that are readily interchangeable.

The usual QqTOF configuration is comprised of three quadrupoles, as in a triple quadrupole spectrometer, with the initial quadrupole acting as a rf-only quadrupole that serves to provide collisional damping. The following two quadrupoles perform as they would in a standard triple quadrupole mass analyzer; however, the third quadrupole is replaced by a reflecting TOF mass analyzer. For MS measurements, the mass filter is operated in a rf-only mode, permitting all of the ions to pass directly through onto the TOF tube. The resulting spectrum benefits from the high resolution and mass accuracy of the TOF tube. For MS/MS, the mass filter allows only the ion(s) of interest to pass to the collision cell, where it undergoes