Methods in Molecular Biology 1007

Springer Protocols

Rune Matthiesen *Editor*

Mass Spectrometry Data Analysis in Proteomics

Second Edition

Humana Press

Mass Spectrometry Data Analysis in Proteomics

Second Edition

Edited by

Rune Matthiesen

Institute of Molecular Pathology and Immunology
Universidade do Porto, Porto, Portugal

Humana Press

Editor
Rune Matthiesen
Institute of Molecular pathology
and Immunology
Universidade do Porto
Porto, Portugal

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-62703-391-6 ISBN 978-1-62703-392-3 (eBook)
DOI 10.1007/978-1-62703-392-3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013934714

Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book has been prepared to start out with a basic introduction to mass spectrometry-based proteomics. The introductory chapter is followed by chapters written by experts in specific subdomains of MS-based proteomics. The experts present their view on specific MS-based methods or data analysis strategies in proteomics. The book is written with the purpose of covering specific topics in detail rather than glossing over the topic. Detailed discussion is made possible since Springer allowed no page restriction on the experts' chapters.

This is the second edition of the book, and, in the 6 years since the publishing of the first edition, the instrumentation and methods have been considerably improved. My personal experience as editor has also been enhanced. I promise the reader that much effort and time have been devoted to making the book up-to-date, and, in fact, every single chapter is either new or has been completely rewritten. I have no doubt that this book is useful for everyone working in proteomics. I have to thank the many invited authors for their on-time delivery and detailed chapters. The authors have provided original manuscripts and even novel concepts in some cases. The book covers data analysis topics relevant for quantitative proteomics, posttranslational modification, HX-MS, glycomics, and data exchange standards.

Porto, Portugal *Rune Matthiesen*

Contributors

Conrad Bessant • *School of Biological and Chemical Sciences, Queen Mary University of London, London, UK*

Jakob Bunkenborg • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*

Ana Sofia Carvalho • *Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal*

Robert J. Chalkley • *Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, USA*

Susmita Datta • *Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, KY, USA*

Michael J. Eggertson • *Waters Corporation, Milford, MA, USA*

John R. Engen • *Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA*

Jun Fan • *School of Biological and Chemical Sciences, Queen Mary University of London, London, UK*

Catarina Fonseca • *Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal*

H.C. Harsha • *Institute of Bioinformatics, International Technology Park, Bangalore, India*

Alex A. Henneman • *Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands*

Rafael C. Jiménez • *EMBL-EBI, Hinxton, Cambridge, UK*

Zhenxin Lin • *Department of Chemistry, University of Michigan, Ann Arbor, MI, USA*

David M. Lubman • *Department of Surgery, University of Michigan, Ann Arbor, MI, USA*

Rune Matthiesen • *Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal*

Yassene Mohammed • *Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands*

Hugo Osório • *Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal; Faculty of Medicine, University of Porto, Porto, Portugal*

Magnus Palmblad • *Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands*

Akhilesh Pandey • *Departments of Biological Chemistry, Pathology, and Oncology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA*

Sneha M. Pinto • *Institute of Bioinformatics, International Technology Park, Bangalore, India*

Celso A. Reis • *Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal; Faculty of Medicine, Institute of Biomedical Sciences of Abel Salazar (ICBAS), University of Porto, Porto, Portugal*

Alan L. Rockwood • *Department of Pathology, University of Utah, Salt Lake City, UT, USA*
Juan Antonio Vizcaíno • *EMBL-EBI, Hinxton, Cambridge, UK*
Paula Voabil • *Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal*
Thomas E. Wales • *Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA*

Contents

Chapter 1

Introduction to Mass Spectrometry-Based Proteomics

Rune Matthiesen and Jakob Bunkenborg

Abstract

Mass spectrometry has been widely applied to study biomolecules and one rapidly developing field is the global analysis of proteins, proteomics. Understanding and handling mass spectrometry data is a multifaceted task that requires many decisions to be made to get the most comprehensive information from an experiment. Later chapters in this book deal in-depth with various aspects of the process and how different tools can be applied to the many analytical challenges. This introductory chapter is intended as a basic introduction to mass spectrometry (MS)-based proteomics to set the scene for newcomers and give pointers to reference material.

There are many applications of mass spectrometry in proteomics and each application is associated with some analytical choices, instrumental limitations and data processing steps that depend on the aim of the study and means of conducting it. Different aspects of the proteome can be explored by choosing the right combination of sample preparation, MS instrumentation and data processing. This chapter gives an outline for some of these commonly used setups and some of the key concepts, many of which are explored in greater depth in later chapters.

Key words Data formats, Proteomics, Mass spectrometry, Sample preparation

1 Introduction

1.1 What Is a Mass Spectrometer?

A mass spectrometer is a device for measuring the mass-to-charge ratio of ionized molecules. A wealth of qualitative and quantitative information can be extracted from this simple device ranging from elemental composition to detailed structural information. All mass spectrometers consist of three main parts; an ion source, a mass analyzer, and a detector (*see* Fig. 1). Analyte ions are produced in the ion source. The ion source generates ions by transferring molecules from the condensed (liquid or solid) phase to gas phase and ionizing them in the process (either positive or negative charge state *see* **Note 1**). The most commonly used ionization methods in proteomics are Electro Spray Ionization (ESI) and Matrix-Assisted Laser Desorption and Ionization (MALDI)—two soft ionization techniques that can deliver fragile biological

Rune Matthiesen (ed.), *Mass Spectrometry Data Analysis in Proteomics*, Methods in Molecular Biology, vol. 1007,
DOI 10.1007/978-1-62703-392-3_1,

Fig. 1 Outline of generic mass spectrometer components. (**a**) The molecules are converted to ionized species in the gas phase in the ion source. (**b**) The mass analyzer separates the ions according to their mass-to-charge ratio. (**c**) The detector records a signal that can be electronically amplified and stored

Table 1
Overview of different mass spectrometry instrument components and methods

Ion source	Mass analyzer	Fragmentation	Detector
ESI [3]	**Quadrupole** [4]	**CID/CAD** [5]	**Electron multiplier** [6]
MALDI [7]	**Time of flight** [4]	**HCD** [8]	**Inductive** [9]
SELDI [10]	**Ion traps**	**ETD** [11]	HED [6]
PD [12]	**3D quadrupole** [4]	**ECD** [13]	MCP [14]
ESSI [15]	**Linear quadrupole** [16]	PSD [17]	Faraday cup [6]
FAB [12]	**Orbitrap** [2]	IRMPD [18]	Scintillation counter [6]
LDI [19]	**FT-ICR** [9]	BIRD [20]	
	Magnetic sector [21]	SID [22]	

The different parts can be combined in many different ways. The references are suggested as entry points for further reading. The most common instrument components are highlighted in bold. Abbreviations are spelled out in Appendix A Chapter 18

macromolecules intact into the gas phase. The charged biomolecules in the gas phase can be controlled and analyzed by electric and magnetic fields in the mass analyzer. The ions that are produced in the ion source are transferred to the mass analyzer where they are separated according to their mass to charge ratio (m/z). There are a number of different mass analyzers operating by different principles and having different properties that are described later. Mass analyzers can be used in combination and tandem mass spectrometers are widely used because the composition of ions can be further examined by fragmenting the ions and analyzing the fragmentation pattern. The ions are finally recorded by a detector. The mass analyzer and detector are always within the high vacuum region [1]. As later described, ion sources can be combined with different mass analyzers giving mass spectrometers such as MALDI-TOF (time-of-flight), ESI-IT (ion trap), and ESI-Orbitrap [2]. A more complete overview of different mass spectrometry components is presented in Table 1.

Different mass analyzers operate by manipulating the ions using different principles but common for all is that the results

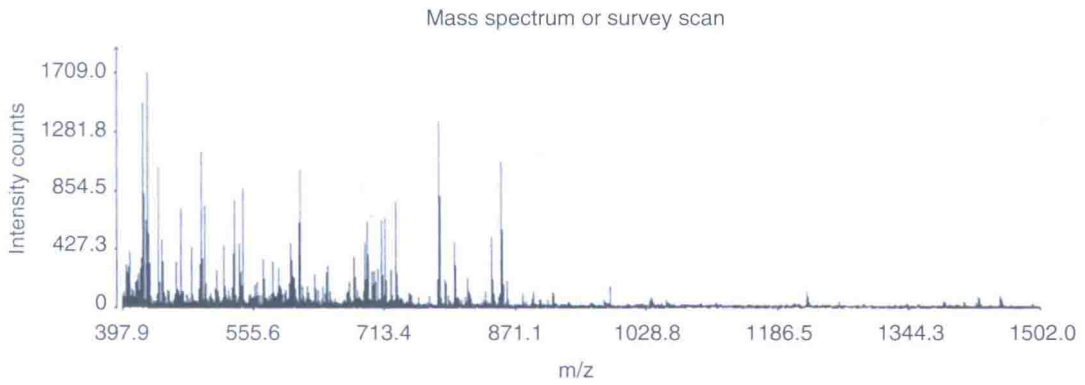

Fig. 2 The MS spectrum displays the abundance of ions as a function of mass-to-charge ratio. This spectrum displays ions in the *m/z* range from 350 to 1,500 Th and the *y*-axis displays the intensity of the signal from a given ion. The charge state of an ion can often be deduced from the isotope pattern and hence the mass of the molecules can be determined

can be transformed to intensities as a function of *m/z* (mass over charge) values. For example, Orbitrap and FT-ICR measure an AC image current induced by ions trapped in an electric or magnetic field that can be transformed to *m/z* values by Fourier Transformation. Time of flight instruments accelerate ions and measure the flight time between acceleration and hitting a detector. The flight time can again be transformed to *m/z* values. The output from the instrument is ion intensity at different *m/z* values. The result is visualized by an *m/z* versus intensity plot that is a mass spectrum (*see* Fig. 2).

By hyphenating liquid chromatography with MS (LC-MS) a whole series of MS spectra can be acquired as the molecules elute. Each MS spectrum in this array is also referred to as survey scans because ions can be selected from this spectrum and further interrogated by tandem mass spectrometry. In a typical MS setup for protein identifications the survey scan is analyzed on the fly by the instrument software to select ions that are isolated, fragmented and analyzed by a mass analyzer to generate an MS/MS spectrum (*see* Fig. 3). In the MS/MS spectrum of peptides sequence information is obtained by correlating mass differences between peaks with residue masses (*see* **Note 2**). Typically the most prominent features of the spectrum are extracted and used to query a protein database using different software tools.

There are many different instrument setups and fragmentation techniques that can be applied in the proteomics field. A glossary of some of the most commonly used is given in Table 1 and a selection of these is discussed in further detail later on in this chapter.

1.2 What Is MS-Based Proteomics?

Proteomics is the global analysis of all aspects of proteins and MS has in recent years become one of the most informative methods for studying proteins. Mass spectrometry offers complementary information to the detailed structural information obtained by

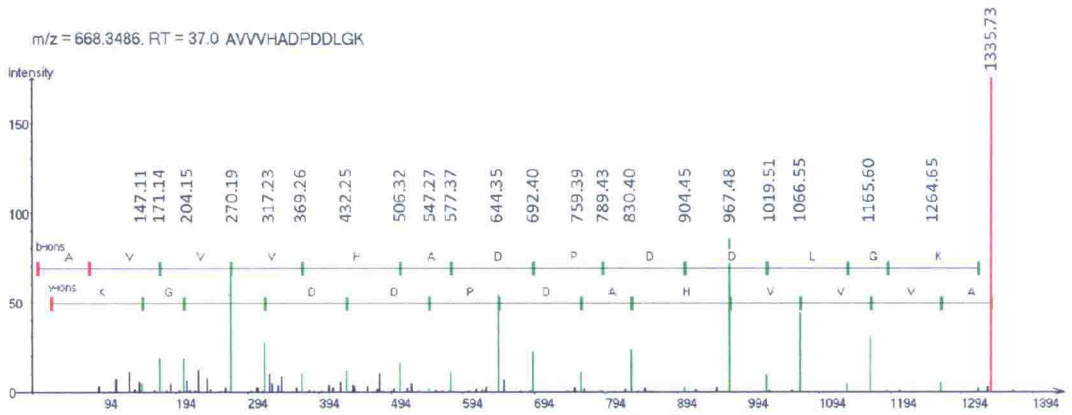

Fig. 3 Extracting information by tandem mass spectrometry. Using ESI-MS to analyze peptides, a doubly charged ion with observed *m/z* 668.3486 was isolated and fragmented by CID. The resulting MS/MS spectrum displays the intensities and *m/z* values of the resulting fragment ions. As described later the mass differences between peaks can be correlated to amino acid residue masses and the peptide sequence can be deduced to be AVVVHADPDDLGK from this information

condensed phase methods such as NMR and X-ray crystallography. Among the advantages are that mass spectrometry is much easier to automate, more sensitive, easy to hyphenate with different separation techniques and can be used on complex mixtures. Alternative methods such as protein arrays are more sensitive but also expensive and suffer from technical problems such as poor reproducibility. A discussion on some of the main applications of MS-based proteomics is provided in the following paragraphs.

One of the main aims of MS-based proteomics is to identify and quantify proteins and their post translational modifications in either a purified, enriched or complex protein mixture. MS-based proteomics for identification purposes can be divided into bottom-up [23], shotgun [24], and top-down approaches [25] (*see* Fig. 4). In bottom-up proteomics proteins are cut into peptides that serve as input to the MS equipment, whereas in top-down proteomics the mass spectrometer isolates a full length protein which can be subsequently fragmented inside the mass spectrometer and the masses of the fragments recorded as well. Top-down proteomics has traditionally been restricted to FTICR instruments but recently orbitrap instruments have been used as well [26, 27]. Shotgun proteomics is a special case of bottom-up proteomics where a complex mixture of proteins is digested into peptides, typically by trypsin, followed by multidimensional high performance liquid chromatography online coupled to the mass spectrometer. The difference between bottom-up and shotgun proteomics is that bottom-up strategies does not necessarily have LC separation of peptides prior to MS, whereas in shotgun strategy LC and typically multidimensional LC is always used to separate a complex mixture of peptides originating from many different proteins.

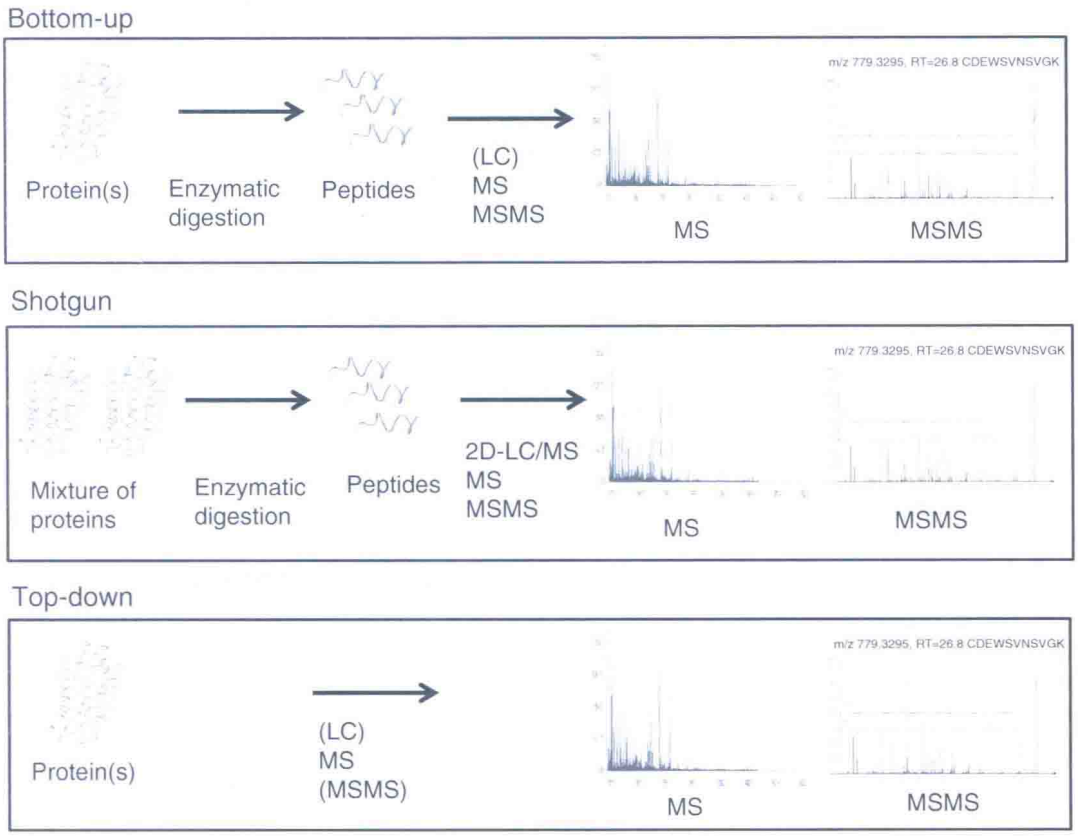

Fig. 4 Outline of the different analytical strategies named bottom-up, top-down, and shotgun proteomics. Intact proteins are analyzed in top-down experiments, whereas the proteins are processed to peptides prior to MS analysis in bottom-up and shot-gun experiments

Site identification of post translational modifications is another major use of MS-based proteomics. Amino acid residues can be covalently modified by various processes and many regulatory modifications, such as phosphorylation, acetylation, and methylation can be analyzed on a large scale by MS. Other modifications such as glycosylation, ubiquitinylation, and sumoylation are more complicated due to size and complexity and although the site mapping of some of these modifications can be done on a fairly large scale they are much harder to fully characterize. MS can be used for relative and absolute quantitation of peptide and proteins together with their associated post translational modifications (*see* Chapter 8 for more details on quantitative proteomics).

MS can be used to examine the primary structure of proteins and to map proteolytic cleavage sites in proteins by applying a strategy called N- and/or C-terminomics in which the N- and C-terminus of the proteins can be determined. By chemically modifying the C- and N-terminus at the protein level these termini become distinct from the C- and N-termini generated at later stages in the analysis. Recently a number of methods based on negative

enrichment have been proposed for enrichment of peptides that define the N- and C-terminus of proteins. These methods are referred to as N-TAILS [28] and C-TAILS [29], respectively. N-TAILS has also been combined with iTRAQ quantitation the so-called iTRAQ-TAILS [30]. N-terminomics can also be analyzed by combined fractional diagonal chromatography (COFRADIC) [31]. The lysine amines can also be blocked by guanidination followed by biotin tagging of terminal amines and positive enrichment on a streptavidin column [32]. NC-terminomics can be used to study proteolytic signaling and to identify protease labile sites in enzymes which can be used to improve enzyme stability.

MS can also be used for structural studies where conformational information can be extracted by mass changes. For example, hydrogen deuterium exchange can be used to examine if protein has folded correctly, to measure conformational changes after buffer exchange or modification of the protein, identification of binding sites of antibodies, other proteins and ligands, identification of site important for protein aggregation and to validate that different protein production strategies end up with the same protein fold (*see* Chapter 11 for more details on hydrogen deuterium exchange).

Mass spectrometry has found many related applications beyond identification and quantitation of proteins and has been pitted against macromolecular analytical challenges as microorganism classification [33] and studying protein complexes [34].

2 Introduction to Methodologies

The above short introduction intended to give a quick glimpse of MS-based proteomics and form a starting point for further reading. In the subsequent section more details of some of the methods mentioned above is provided.

2.1 The Basic Concepts of Proteomics

The term proteomics covers the analysis of all proteins expressed in an organism. One of the first tools used in proteomics was two-dimensional gel electrophoresis (2D-GE) introduced in 1969 [35] where proteins are separated by their isoelectric point (pI, the pH where the protein is uncharged) via isoelectric focusing and by size via their migration on an SDS-PAGE gel. Mass spectrometry (MS) techniques have been combined with two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) for direct and systematic identification of polypeptides. It has been shown that traditional 2D-PAGE can resolve up to 1,000 protein spots in a single gel [36]. This is an impressive number but compared to the number of expressed genes in various organisms which range typically from 5,000 to 40,000 it is clearly not good enough. Especially if one also considers that in eukaryotes each gene can have several splice forms and each protein can have an array of post-translational modifications [37]. Recently, efforts have been made to optimize the

standard 2D-PAGE technique by making larger 2D-PAGE [38]. The technique uses multiple narrow range isoelectric focusing gels to improve separation in the first dimension. In the second dimension, multiple long SDS-PAGE gels of different polyacrylamide concentrations are used. The large 2D-PAGE was claimed to resolve more than 11,000 protein spots. In general it is the low-abundance and hydrophobic proteins which are difficult to identify by the 2D-PAGE based method [39]. Proteins with extremes in pI and molecular mass will not be retained in the gel. In addition, 2D-PAGE has a low throughput of samples [40].

Protein spots resolved by 2D-PAGE are typically cut out and the proteins enzymatically digested in-gel [41, 42] or digested during blotting onto membranes containing immobilized trypsin [43]. This is a typical bottom-up experiment where each digest yields a peptide mixture that can be analyzed by MS (peptide mass fingerprinting, PMF), by MS/MS, or by a combination of the two.

Multidimensional protein identification technology (MudPIT) is a method where the entire complex protein sample is digested without prior separation on the protein level. The resulting highly complex mixture of peptides is fractionated by chromatography on the peptide level, originally using strong cation exchange chromatography, and analyzed by LC-MS. This type of approach is also referred to as Shotgun Proteomics [24] and has proven to be very efficient for identification of proteins. However, shotgun technology has some problems with the confidence of the identified peptides, correlating the identified peptides to the proteins they originated from and how to accurately quantify the proteins. One way to quantify is to integrate the UV absorbance or the intensity counts recorded by the mass spectrometer of a peptide in the chromatographic step. The method requires high reproducibility when two samples are compared. Reproducibility across several steps of chromatography is especially difficult to achieve if nano-liquid chromatography (nLC) columns are used. Especially frustrating are failures in the analysis due to partial blocking of columns they may occur to different extents during consecutive runs. A more precise method of quantification of peptides in LC-MS/MS experiments is to use stable isotope labeling. The quantitative methods can be divided into relative and absolute quantification methods (*see* Chapters 8 and 9). The principle is that two or more samples are labeled with different stable isotopes. The differential labeling can occur during the biosynthesis of the proteins in cultured cells (SILAC, *see* Chapter 8), by reacting residues with labels containing different stable isotopes (Chemical labeling, *see* Chapter 8), or by enzyme catalyzed incorporation of ^{18}O in the peptide from ^{18}O water during proteolysis [44].

The shotgun method gives an enormous amount of data, which requires automatic processing (*see* Chapter 5). Automatic computer based interpretation of data calls for high quality statistical testing

to evaluate the quality of the interpretation. In the traditional 2D-PAGE several peptides from the same protein normally confirm the identification, whereas the shotgun method often claims identification of proteins from two peptide sequence tags. This is problematic since the same tryptic peptides can occur in rather diverse protein sequences. Therefore, the shotgun experiments require that the significance of protein assignment must be more precisely evaluated than for the 2D-PAGE method.

2.2 Sample Preparation for MS

The quality outcome of MS-based proteomics is heavily dependent on sample complexity and purity. The main reason for this is ion suppression (*see* **Note 3**) where many different species compete for the charges during the ionization process. This can derive from peptides that overlap in liquid chromatography (LC) retention time and *m/z* value dimensions. The quality of the data is also heavily affected by contaminations from different sources. It is for example not uncommon when studying proteins from cell culture models to detect proteins from cell culture medium (typically bovine proteins from the serum used to supplement growth media). If proteins from the cell culture media are not included in the searched database then homologous proteins from the target organism get identified instead of the contaminating proteins from the cell culture media [45] (*see* **Note 4**). Even if the contaminating proteins are included in the searched database problems such as ion suppression and overlap with target peptides of interest may occur. During sample preparation human and sheep keratin from clothing can contaminate the sample and cause the same problems as protein contaminants from the cell culture medium. It is therefore import to work in a laminar flow hood, use gloves and rinse the gloves on regular basis. Measures to minimize chemical contaminants should also be taken. There are many different sources of chemical contaminants such as polyethylene glycol [46] from plastic tubes, volatile chemicals [47] in the oil used for mass spectrometer pumps and detergents used in buffers to solubilize proteins. It is therefore recommended to consider the use of detergents for cleaning glass ware such as gel electrophoresis plates and as a rule of thumb prepare all solutions freshly. A protein sample that has been contaminated by detergents can be cleaned by the FASP (Filter Aided Sample Preparation) protocol [48].

A proteomic project starts by generating a protein extract from tissues or from a cell culture. It is an advantage to chemically control the reactive cysteines at the earliest possible stage to prevent mixtures of different cysteine modifications. The reactive cysteines can for example become oxidized or react with non-polymerized acrylamide during electrophoresis [1]. Typically the protection is done by reduction with DTT and alkylation with iodoacetamide, but a discussion of the advantages of different cysteine modifications is presented in [49].

Table 2
Typical separation methods used in combination with MS setups

Pre-MS separation	1. Step	2. Step
2D-page	MALDI-MS	LC-MS
1D-PAGE	1D-LC-MS	
Chromatography	1D-LC-MS	
None (raw extract)	2D-LC-MS	

Notice that a common strategy, when 2D-PAGE separation is used, is to first attempt to identify the in gel digested protein by using half of the sample on MALDI-MS (both MS and MS/MS) and if this fails then use the remaining sample for LC-MS runs

In a MALDI-TOF MS setup peptide mass fingerprinting for protein identification, the protein must be almost 100 % purified before the proteolytic cleavage and subsequent MS analysis. The preferred method for partial protein purification in combination with MALDI-TOF MS is 2D-PAGE. LC-MS based approaches can both be used for simple protein mixtures purified by chromatography or 2D-PAGE or for more complex protein mixtures. An increase in protein sample complexity leads to more complex peptide mixtures requiring more efficient chromatography steps prior to MS analysis. In LC-MS the chromatography is directly coupled to the MS instrument. In the typical 1D-LC-MS a reverse phase C18 column is used for separating semi complex samples because the volatile buffers are well suited with MS analysis. In 1D-LC-MS the gradient and separation time can be varied according to the complexity of the sample. For directly analyzing whole cell extract 2D-LC-MS approaches are typically used. The first column is then frequently a strong cation exchanger (SCX) which is directly coupled to a reverse phase column which again is coupled to the MS instrument. An overview of typical strategies is provided in Table 2.

The proteolytic cleavage is most often done with trypsin in a 1:50 trypsin to protein ratio. Trypsin is a serine protease that specifically cleaves at the carboxylic side of lysine and arginine residues if these are not followed by proline (*see* **Note 5**). It is important to note that "in gel digestion" requires more trypsin than in solution digestion. The abundance and distribution of lysine and arginine residues in proteins are such that trypsin digestion yields peptides of molecular mass that are well suited for analysis by MS. The specificity of trypsin is of extreme importance. Native trypsin is subjected to autolysis, generating pseudotrypsin, which exhibits a broadened specificity including chymotrypsin-like activity [50]. Additionally, trypsin is often contaminated with chymotrypsin. For these reasons, trypsin which has reductively methylated lysine and has been treated with TPCK (*N*-tosyl-L-phenyl chloromethyl ketone) a chymotrypsin inhibitor, should be used. Such trypsin preparations can be