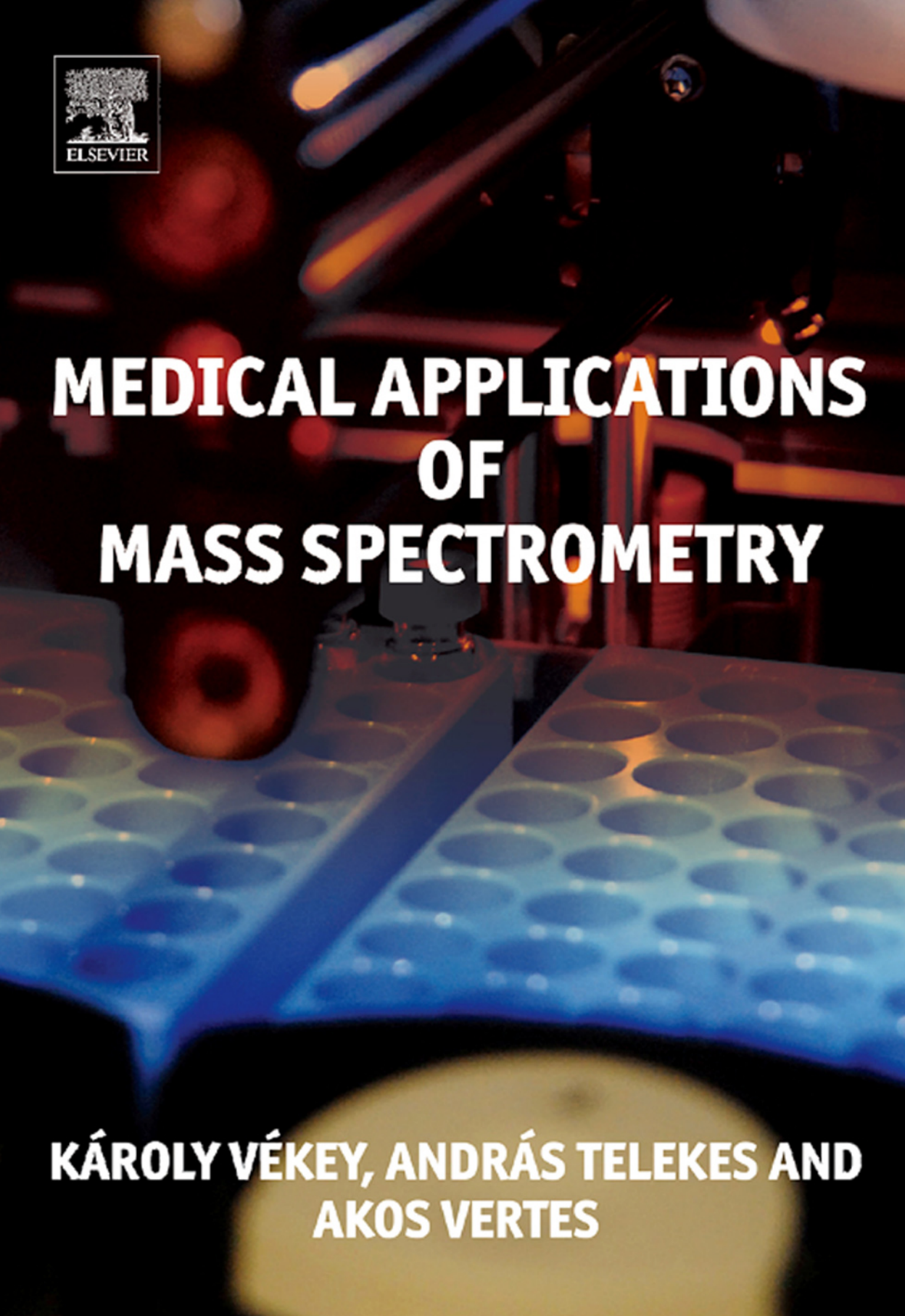

ELSEVIER
MEDICAL APPLICATIONS
OF
MASS SPECTROMETRY
KÁROLY VÉKEY, ANDRÁS TELEKES AND
AKOS VERTES

MEDICAL APPLICATIONS OF MASS SPECTROMETRY

This page intentionally left blank

MEDICAL APPLICATIONS OF MASS SPECTROMETRY

Edited by

Károly Vékey

Chemical Research Center, Hungarian Academy of Sciences
Budapest, Hungary

András Telekes

National Institute of Oncology,
Budapest, Hungary

Akos Vertes

W.M. Keck Institute for Proteomics Technology and Applications,
Department of Chemistry, The George Washington University
Washington, DC, USA

ELSEVIER

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

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
Linacre House, Jordan Hill, Oxford OX2 8DP, UK

First edition 2008

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-444-51980-1

Printed and bound in The Netherlands

08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

Contents

List of Contributors

Günter Allmaier
Institute of Chemical Technologies and Analytics, Vienna University of Technology, Vienna, Austria

Reem Berro
Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC 20037, USA

Olaf Bodamer
Division of Biochemical and Paediatric Genetics, University Children's Hospital, Vienna, Austria

Richard M. Caprioli
Department of Biochemistry and the Mass Spectrometry Research Center, Vanderbilt University School of Medicine, Nashville, TN 37232-8575, USA

Emilia Caputo
Institute of Genetics and Biophysics-IGB, CNR, Naples, Italy

Bruno Casetta
Applied Biosystems, Monza, Italy

Plamen A. Demirev
Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, MS 2-217, Laurel, MD 20723–6099, USA

Dominic M. Desiderio
Charles B. Stout Neuroscience Mass Spectrometry Laboratory, University of Tennessee Health Science Center, 847 Monroe Avenue, Memphis, TN 38163, USA
and
Department of Neurology, University of Tennessee Health Science Center, 847 Monroe Avenue, Memphis, TN 38163, USA

Jonathan A. Epstein
National Institute of Child Health and Human Development, NIH, Bethesda, MD, USA

Elisabeth Förster-Waldl
Department of Paediatrics and Juvenile Medicine, Medical University of Vienna, Vienna, Austria

Perttu Haimi
Institute of Biomedicine, Department of Biochemistry, University of Helsinki, 00014 Helsinki, Finland

Károly Héberger
Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary

Márta Hegedűs
National Institute of Oncology, Budapest, Hungary

Martin Hermansson
Institute of Biomedicine, Department of Biochemistry, University of Helsinki, 00014 Helsinki, Finland

Jasmin Hirschmann
Institute of Chemical Technologies and Analytics, Vienna University of Technology, Vienna, Austria

Hongzhan Huang
Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Reservoir Road, NW, Washington, DC 20057-1455, USA

Fatah Kashanchi
Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC 20037, USA
and
The Institute for Genomic Research, Rockville, MD 20850, USA

István Kiss
St. Imre Government Teaching Hospital, Budapest, Hungary

Martina Marchetti
Institute of Chemical Technologies and Analytics, Vienna University of Technology, Vienna, Austria

Peter McGarvey
Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Reservoir Road, NW, Washington, DC 20057-1455, USA

Kornél Nagy
Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary

Matthew T. Olson
National Institute of Child Health and Human Development, NIH, Bethesda, MD, USA

Laszlo Prokai
Department of Molecular Biology & Immunology, University of North Texas Health Science Center, Fort Worth, TX 76107, USA

Anne Pumfery
Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC 20037, USA

John Roboz
Division of Hematology/Oncology, Department of Medicine, Mount Sinai School of Medicine, 1 Gustave Levy Place, New York, NY 10029, USA

Harold Sacks
Division of Endocrinology, University of Tennessee Health Science Center, 847 Monroe Avenue, Memphis, TN 38163, USA

Sarah A. Schwartz
Midwest Research Institute, Kansas City, MO 64110-2241, USA

Pentti Somerharju
Institute of Biomedicine, Department of Biochemistry, University of Helsinki, 00014 Helsinki, Finland

Árpád Somogyi
Department of Chemistry, University of Arizona, Tucson, AZ, USA

András Telekes
National Institute of Oncology, Budapest, Hungary

Bradley J. Thatcher
Ciphergen, Italy

Andreas Uphoff
Institute of Biomedicine, Department of Biochemistry, University of Helsinki, 00014 Helsinki, Finland

György Vas
Cordis Corporation, Analytical Technologies, Pharmaceutical & Package Development, Welsh & McKean Roads, P.O. Box 776, Spring House, PA 19477-0776, USA

Károly Vékey
Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary
and
Hungarian Academy of Sciences, Chemical Research Center, Hungarian Academy of Sciences, Pusztaszeri ut 59-67, Budapest 1025, Hungary

Akos Vertes
W.M. Keck Institute for Proteomics Technology and Applications, Department of Chemistry, The George Washington University, Washington, DC, USA

Cathy H. Wu
Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Reservoir Road, NW, Washington, DC 20057-1455, USA

Alfred L. Yergey
National Institute of Child Health and Human Development, NIH, Bethesda, MD, USA

Xianquan Zhan
Charles B. Stout Neuroscience Mass Spectrometry Laboratory, University of Tennessee Health Science Center, 847 Monroe Avenue, Memphis, TN 38163, USA
and
Department of Neurology, University of Tennessee Health Science Center, 847 Monroe Avenue, Memphis, TN 38163, USA

Preface

The next frontier for mass spectrometry (MS) lies in medicine. This book provides evidence for this proposition and will help to realize it. Over the past 25 years, MS and its accompanying technology, has been driven to a significant degree by the aim of achieving successful application to all classes of biological molecules. It is worthwhile to consider this objective and the methods used to achieve it, in part because it embraces many of the results recounted in this text. Such a retrospective also provides guidance for the future as to the likely course of developments in MS as it engages ever more directly with the medical sciences and with clinical practice.

The main objective which has driven MS over the past quarter century was refreshingly clear-cut . . . it was the desire to ionize any type of molecule and to obtain characteristic molecular mass and structural information with which to achieve identification. The result of this focused effort was the development of ionization methods applicable to an immense variety of chemical and biochemical molecular types, present in samples encompassing an array of physical states. Complementary technology was developed to allow the dissociation of particular ions so as to provide structural information from the characteristic fragmentation processes. The successes in ionization are evident from the large amount of space devoted to electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) the standard methods for analyzing biomolecules in solution and in the condensed phase, respectively. Chapter 6 includes basic coverage of the ionization methods; their applications are to be found in many other chapters throughout the text. The same chapter introduces the different types of mass analyzers used in mass spectrometry; these devices are based on various physical principles and have complementary advantages. Each has legions of supporters. The successes achieved in developing methods of producing characteristic fragments from specific compounds are dependent on the ability to carry out tandem mass spectrometry, (MS/MS), that is, the ability to perform experiments on specific mass-selected ions. These experiments often involve collisions of ions with neutral atoms or molecules (collision induced dissociation) but there is strong interest in alternatives such as those in which dissociation is a consequence of electron capture (electron capture dissociation, ECD) or electron transfer (electron transfer dissociation, ETD). These techniques have developed rapidly in the past few years and are widely applied to the characterization of proteins.

The technologies described here have had major effects in developing new paradigms in biology. The mass spectrometry developments, in conjunction with

chromatographic methods which achieve sample separation and automated introduction into the mass spectrometer (described in Chapter 5), have led to great success is characterizing and quantifying proteins (the topic of Chapter 8). The method of protein sequencing in which they are degraded to peptides and the peptides are sequenced by MS/MS (the "bottom up" methodology) is one of the outstanding achievements of modern mass spectrometry and biology. This, and other contributions from mass spectrometry, has played a key role in the birth of the field of proteomics. The subject is taken up in detail in several chapters (Ch. 8–10, 15) with appropriate emphasis on the need for greatly enhanced methods of automated data handling and interpretation.

The related topics of metabolomics and lipidomics (Chapter 11) are also, in significant part, outgrowths of research and developments in mass spectrometry. This text contains fascinating chapters on the applications of mass spectrometry to a variety of problems including for example drug and drug metabolite monitoring, (Chapter 13), a classic field in which chromatography and mass spectrometry are used in combination for quantitation of trace amounts of specific compounds in complex biofluids. Similarly the treatment of infectious pathogens (Chapter 14) presents the range of application of mass spectrometry and its growing potential to contribute to clinical diagnostics. There are few more striking examples of this latter application to neonatal screening (Chapter 16), an application that relies on MS/MS methods.

In considering these and other successful applications of MS to biological samples it is worth noting that some objectives have not been fully realized. This means that there is considerable room for future advances. Notable among unrealized objectives are:

(i) Ionization is inefficient, never more than 0.1%; (ii) The dynamic range of MS is limited in real (complex) sample analysis; (iii) The application of MS to chiral and other stereoisomers has been limited; (iv) Quantitative analysis is achieved by methods that are strongly dependent on solution chemistry and which are slow and relatively expensive. In spite of the strong progress in applying mass spectrometry in some areas of medicine and biochemistry, there are other areas in which much more progress can be and is likely to be made in the future. Areas ripe for progress include (i) Nucleic acids, a subject in which extensions of the molecular weight range has been far less successful than in the protein area; (ii) Protein complexes, currently an emerging area as instrumentation and methods capable of providing high quality data at high mass become available; (iii) Lipids, where the complex structure/fragmentation patterns have been incompletely elucidated; (iv) Glycoproteomics and (v) Quantitative proteomics, especially for low copy number proteins.

The retrospective discussion which this Preface has followed provides a vantage point for attempting to discern likely significant future developments. The trends and achievements just noted refer to the application of mass spectrometry to

traditional qualitative and quantitative analysis of biomolecules, albeit biomolecules in complex solutions. There are other, quite different ways in which mass spectrometry might in future be useful in medicine. The driving forces for the next stage of development of MS and its applications to medicine include the following:

Imaging mass spectrometry
In situ mass spectrometry
In vivo mass spectrometry

These new tools will allow applications of MS in medicine which go far beyond biochemistry (and far deeper into biochemistry) to include pathology and forensics and clinical diagnosis. Brief consideration of each of these topics is worthwhile.

The use of mass spectrometry to create molecular images of the distribution of compounds in biological material, discussed in Chapter 24, is an experiment that has rapidly come to the fore in the past decade. There are (as is so often the case in mass spectrometry) several different ways to do the experiment, including MALDI imaging and secondary ion mass spectrometry. These are not rapid experiments but they provide remarkable spatial and chemical resolution and are beginning to contribute significantly to the discovery of biomarkers for disease. In respect to the second item, mass spectrometers have generally been designed for the lab environment, not the bedside or operating room. However, a new generation of miniature mass spectrometers is emerging with capabilities for biomolecule analysis; such in situ instruments may well be major drivers of future progress in clinical practice. The third of these capabilities—in vivo experiments using mass spectrometry—is yet to be realized. However, the conjunction of new ionization experiments in which the sample is in the ambient environment—especially the desorption electrospray ionization (DESI) method—with the emergence of miniature mass spectrometers makes this a credible objective.

Graham Cooks
Henry B. Hass Distinguished Professor
Department of Chemistry
Purdue University

This page intentionally left blank

Part I
Motivation and Essentials

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 1

Introduction

AKOS VERTES[a,*] and KÁROLY VÉKEY[b,1]

[a]*W. M. Keck Institute for Proteomics Technology and Applications, Department of Chemistry, The George Washington University, Washington, DC, USA*
[b]*Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary*

Mass spectrometry took the biomedical field by storm. The cross-fertilization of these fields was sparked by the confluence of technological development in novel ion sources during the late 1980s and the mounting needs for accurate molecular analysis in biology and medicine. Although the existing technologies at the time, e.g., gel electrophoresis and high-performance liquid chromatography, were simple and ubiquitous, the accuracy of the obtained information was insufficient and the data were slow in coming. For example, gel-based separations could determine the molecular weight of unknown proteins, but the results were reported in kilodaltons. Identifying a protein as a 10-kDa molecule through gel electrophoresis left ~10% or up to 1 kDa uncertainty in its size. Thus, exploring crucial posttranslational modifications, key regulators of protein function, was not a simple matter.

At the same time, mass spectrometry offered exquisite details on the mass and structure of small (<5000 Da) molecules but was unable to efficiently ionize larger ones. The dilemma of the mid-1980s is illustrated in Fig. 1. The results of a two-dimensional gel electrophoresis separation of kidney proteins showed a wealth of information in the >5000 Da range. Results from mass spectrometry, however, left off all molecular species in this region. The lack of efficient ion sources for these molecules started a decade-long race to produce gas-phase ions from ever larger molecules. This quest culminated in the discovery of electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) by the end of the decade. Almost overnight, molecules with masses in excess of 100 kDa could be studied by

*Corresponding author. Tel.: +1-202-994-2717; Fax: +1-202-994-5873.
E-mail: vertes@gwu.edu (A. Vertes).
[1]E-mail: vekey@chemres.hu.

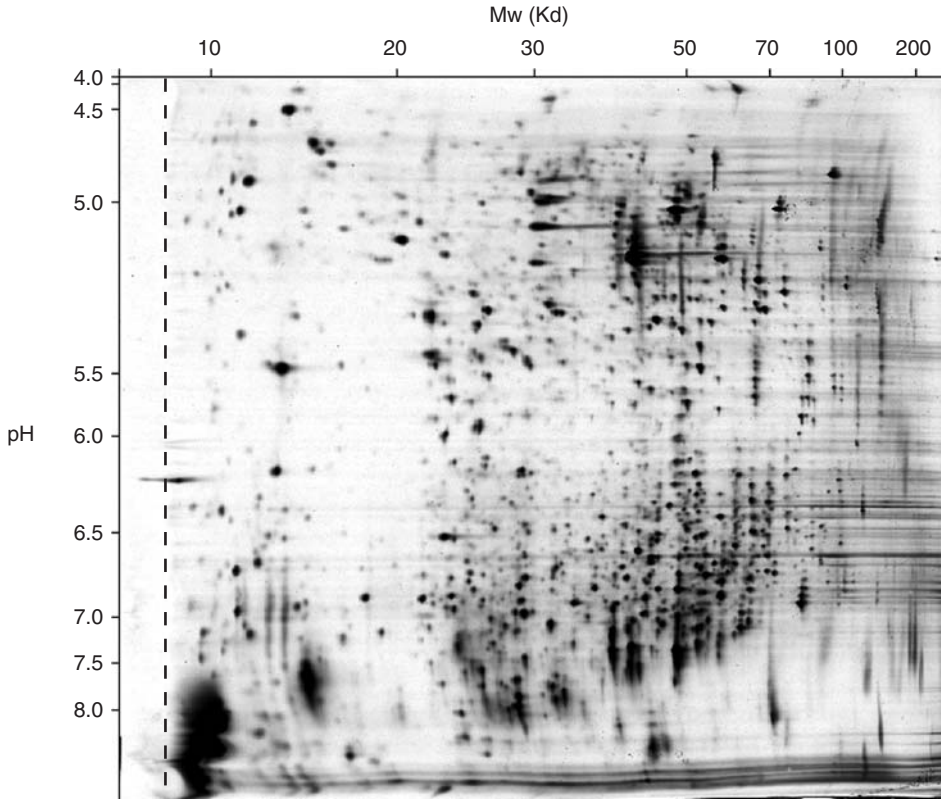


Fig. 1. Gel-separated proteins in human kidney from the SWISS-2DPAGE database (http://www.expasy.org/swiss-2dpage/). In the early 1980s, detecting any species above ~5000 Da (right to the dashed line) was an insurmountable challenge for mass spectrometry. Yet gel electrophoresis showed that a wealth of information was available on crucial biomedical species in this region. The lack of efficient ion sources for these molecules started a decade-long race to produce gas-phase ions from ever larger molecules.

mass spectrometry. The ensuing interest reordered the landscape of mass spectrometry and laid the foundations of new scientific disciplines (e.g., proteomics).

In the wake of these discoveries, established instrument manufacturers (producing sector instruments) became marginalized and others that were quick to embrace the new technology rose to prominence. The opportunity to explore large biomolecules attracted the attention of academia, government, and industry alike. On the scholarly level, the new insight promised a vastly improved understanding of the molecules of life. On a practical level, it enabled the design of smart drugs that specifically targeted the cellular processes related to a particular disease.

It is anticipated that in a few years mass spectrometers will be routinely used in clinical settings. With the availability of dedicated instrumentation and the expanding discovery of disease biomarkers, diagnostic laboratories will increasingly turn

to mass spectrometric methods. Assessment of treatment efficacy and monitoring of patient recovery will also be aided by this technology. At the research level, mass spectrometry is fast becoming an indispensable tool for the biomedical professional. The current generation of medical students and biologists are being trained through their regular degree education in this highly technical field. Training workshops, certificate courses, and continuing education are trying to fill the gap between the increasingly sophisticated new techniques and the limitations of traditional training in bioanalysis.

Fuelled by the emergence of new disciplines, e.g., proteomics and bioinformatics, there is a rapidly increasing demand for advanced information both in laboratory and in classroom settings. Publishers are scrambling to fill these needs. For example, in 2001 and 2002, there were ~19 new volumes published in the field of proteomics alone. Although some of these publications are excellent in conveying the latest information and techniques, most medical professionals and biologists need a more introductory treatise. Indeed, based on surveying the general field of mass spectrometry in the life sciences, this seems to be a significantly underserved niche in these publications. With a few exceptions, there are no mass spectrometry books published specifically dedicated to biomedical professionals.

The structure and the content of this book targets readers at the full spectrum of the advanced student–professional–specialist level. For example, parts of the book were successfully adopted as a high-level text in the Genomics and Bioinformatics Masters Program at the George Washington University (e.g., in the course "Fundamentals of Genomics and Proteomics") and in our doctoral programs. In addition, various courses offered by the Department of Chemistry (e.g., "Ions: Wet and Dry" and "Mass Spectrometry in Life Sciences") capitalized on the text. Although online learning technologies enhance the student experience, the availability of a comprehensive text is of great help. Owing to its broad scope, the book can also serve as a desk reference for professionals and specialists.

Producing this volume amidst the vigorous development of a continuously evolving field, even with our excellent group of contributors, was a challenge. Like in any emerging field, in biomedical mass spectrometry there are "childhood diseases" associated with the employed tools and the methods themselves. Sometimes inappropriate technology is being developed or legitimate approaches end up in inefficient combinations. For example, biomarker discovery with low-resolution mass spectrometers can produce less than convincing data. In other cases—no names are named here—enthusiastic investigators overinterpret their data and "discover" long-sought biomarkers. Although these cases can be embarrassing, the natural evolution of the discipline is sure to correct such blunders. These problems are common in emerging and fast-growing fields everywhere and cannot subtract from the tremendous value produced by the interaction of biomedical fields and mass spectrometry. Therefore, we ask the reader not to look at this book as a finished picture but as the beginning of a long movie.

We made sure that the areas that are mature, such as the foundations of mass spectrometry and its application as a research tool in the medical fields, are thoroughly and accurately discussed. The clinical applications relevant for the practicing physician at the bedside (with the exception of pediatrics) are still in their infancy with only tentative and fragmented information available. Some of the related chapters were written by medical professionals who summarized the available information and lent their unique perspective to these chapters.

The book is built of five main parts. In the first part, essential information on analytical concepts and mass spectrometry is summarized. Specifics of the ethical, legal, and safety aspects of medical research are also included here. The second part focuses on four essential tools of the trade: biomedical sampling, separation methods for complex mixtures, a broad foundation in mass spectrometry, and the chemoinformatics principles used in data analysis. The next part demonstrates how to use mass spectrometry for select classes of biomolecules. Here we mainly focus on peptides and proteins, as these are the molecules that have primarily driven the field. Following short introductions to proteomics, *de novo* sequencing, and the related bioinformatics, the application of mass spectrometry in lipid research is discussed. Clearly, there are numerous other compound groups that could have been included here. Metabolomics, the systematic study of small molecules in living organisms, or glycomics, the field specialized on oligosaccharides, would also deserve their chapters. Unfortunately, we failed to convince the experts in these fields to contribute, so these chapters have to wait for future editions.

The main body of the book in part four is devoted to selected medical applications. Here too, originally we wanted to secure a chapter for every major medical discipline. However, this did not quite work out. Some medical fields are slower to adapt new technologies than others. At the outset, some are less amenable to the application of mass spectrometry. Most fields are still researching the utility of mass spectrometry, i.e., the methods are not yet in the hands of clinicians. Eventually, we worked out a tradeoff. We asked some top scientists to write about their fields of specialization and some practicing doctors to summarize the various medical applications from their point of view. The concluding part of the book gives a glimpse of some emerging areas including biomarker discovery and molecular imaging by mass spectrometry. These exciting applications promise to revolutionize medical diagnostics and drug development.

We envision that in the not-too-distant future clinical laboratories will augment their microscopes, centrifuges, and Coulter counters with oligonucleotide microarray readers and mass spectrometers. As genetics and proteomics are making headways into the decision making of practicing physicians, we hope that this volume can be of help along the way.

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 2

Basics of analytical chemistry and mass spectrometry for medical professionals

KÁROLY VÉKEY[a,*] and ANDRÁS TELEKES[b]

[a]*Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary*
[b]*National Institute of Oncology, Budapest, Hungary*

1. Introduction

In medical sciences, emphasis is increasingly placed on instrumental techniques and accurate, quantitative measurements. It is especially apparent in diagnosis, where imaging techniques and laboratory results have became invaluable and compulsory. Breakthroughs in biochemistry made it possible to characterize physiological processes and living organisms at the molecular level. This led to a proliferation of, e.g., DNA tests and the use of biomarkers in daily clinical practice. Characterization of molecular structure and determination of the composition of a mixture are the fields of analytical chemistry and analytical biochemistry. There is no clear borderline between them; in the following discussion both will be indicated as analytical chemistry. In a medical environment, this shows a large overlap with laboratory analysis.

The objective of analytical chemistry is to *determine the composition of a sample. It means the identity, molecular structure, quantity, and concentration of in*

*Corresponding author. Tel.: +361-438-1158; Fax: +361-438-1157. E-mail: vekey@chemres.hu.

principle all, but in practice some, components of the sample. In most cases we are dealing with complex mixtures, such as blood, urine, or tissue. Complete chemical analysis (identifying and quantifying all components) in such a case is not required, but is not even possible with current technologies. Typically either a few target compounds or a wide range of a given class of compounds (e.g., proteins) are detected, identified, and possibly also quantified. Classes of compounds (e.g., total protein content) may also be measured, while in other cases minor structural deviations (such as single nucleotide polymorphism) are characterized. In application fields, like in most pharmaceutical analyses, all compounds above a certain threshold (e.g., 0.1 or 0.01%) need to be accurately characterized.

In chemical and biochemical analysis first a given compound needs to be identified and its structure determined. Structural studies are most often performed by spectroscopy (mostly nuclear magnetic resonance (NMR) but also IR or UV), X-ray diffraction, or mass spectrometry, although a large number of other techniques are used as well. There are techniques (notably NMR and X-ray diffraction) capable of determining the structure of molecules with no or minimal prior information (up to approximately 1000 Da molecular mass), but these typically require a relatively large amount of pure compound (e.g., 1 mg). Other methods for structure determination, such as IR, UV spectroscopy, or mass spectrometry, also yield valuable structural information, e.g., mass spectrometry is excellent for protein sequencing. These latter techniques have the advantage of requiring less sample (even 10^{-9} or 10^{-12} g may be sufficient) and are well adapted to deal with complex materials (e.g., plasma). Structure determination of macromolecules is more challenging, usually requiring the use of several different methods in combination.

Identification of known compounds is less demanding than the structure determination of an unknown. It is also based on molecular characterization, e.g., spectral features (as discussed earlier), chromatographic retention time, and comparison with standards of known structure. The reliability of identification is a critical issue. Several decades ago the chromatographic retention time itself was often accepted as proof of identification of a compound. It is no longer the case, as various examples of false-positive and false-negative results were found. The current trend is to require more and more detailed and specific information before identification of a compound is accepted. For example, besides retention time, mass spectra and/or accurate mass measurements are also needed.

Following identification of a given compound, its amount (or concentration) needs to be determined as well (quantitation). As a given sample may contain thousands of different compounds in widely differing amounts, this is not a trivial task. Instead of structure identification and quantitation, often the biological effect (such as enzyme activity) is measured in the biomedical field. In many cases measurement of biochemical activity and chemical analysis are performed in parallel.

Analytical techniques yield information on *sample composition at a given time*—usually at the time of sampling (e.g., taking of blood). *Time dependence of molecular concentrations* can also be followed, like in pharmacokinetics, where changes in plasma concentration of a given drug are determined. These are usually performed as a series of measurements on samples taken at different times. Alternatively, continuous monitoring of molecular concentrations may also be performed. In most cases homogenized samples are studied, where spatial information is lost. Occasionally the sample may relate to a particular position (like the central or outer part of tumor growth), but modern analytical techniques are capable of delivering *molecular imaging* as well (a usually two-dimensional distribution of a given molecule in a slice of tissue). These are particularly important to characterize physiological and metabolic processes. Time-dependent studies and molecular imaging not only can yield information on the state of health of a given person but also may shed light on the development of disease and on physiological processes.

2. Terms and definitions

Various terms, definitions, and concepts are needed to discuss analytical results. The *amount of material* is measured in weight (grams, milligrams, etc.) or in molar amounts (e.g., micromole or μmol). *Concentration* is also significant, measured in weight/weight, weight/volume, mole/weight, or mole/volume units (e.g., mg/g, μg/l, μmol/g, or nmol/l). Concentration can also be specified as molar solutions (e.g., millimolar, indicated as mM), which indicates *x* millimoles of sample per liter of solution. Concentration may also be given as parts per million (ppm), parts per billion (ppb, $1{:}10^9$), or parts per trillion (ppt, $1{:}10^{12}$) values (this usually refers to weight/weight, but depending on context it may also mean mole/mole ratios).

Among the most important characteristics of an analytical process are sensitivity and selectivity (or specificity). *Sensitivity* means how large signal is obtained from a given amount of material and what is the signal intensity compared to the noise (S/N ratio). Noise may be due to imperfect instrumentation (*"instrument" noise* due to the noise of electric circuits, scattered light, etc.) and due to *"chemical" noise*. The latter is due to a background of signals originating from various molecules present in the mixture, which interfere with analysis of the target compound(s). Even in very clean samples there are usually a large number of compounds in low concentration, e.g., an "ultra-pure" solvent also contains trace-level impurities. Improving instrumentation reduces instrument noise significantly, but does not reduce the chemical noise, which is becoming the major obstacle while improving sensitivity. The chemical noise can be reduced by increasing the selectivity, as discussed below.

Sensitivity can be related to sample amount but more often it relates to concentration. It is closely connected to the *limit of detection* (LOD), e.g., 10 pmol LOD means that we need at least this amount of compound for detection. "Detectable" is usually defined as a given (typically 3:1) signal-to-noise ratio. *The limit of quantitation* (LOQ) is a similar term, meaning the minimum amount of compound that can be accurately quantified (usually at least 10–20% accuracy is required). LOQ is always larger than LOD, and is often defined as a 10:1 S/N ratio. Sensitivity depends not only on the analytical process and instrumentation but also on the matrix (i.e., whether the target compound measured is dissolved in a pure solvent or plasma). Sensitivity often deteriorates when a complex matrix is used; a 100-fold decrease in sensitivity due to matrix effects is not uncommon.

This brings us to another topic, *selectivity (or specificity)*. This characterizes how well a compound can be measured in the presence of other compounds or in a complex matrix. The signal of various compounds interfering with analysis can be separated from that of the studied compound by increasing the selectivity. In an analogous way, increasing selectivity typically reduces the chemical noise (and therefore decreases detection limits). The specificity needed depends on the problem studied and also on the matrix used (e.g., plasma or tissue). Selectivity is a particularly critical issue in studying isomers (e.g., measuring lathosterol in the presence of cholesterol). To increase selectivity, the sample often needs to be separated into several fractions or specific detectors must be used. Increasing selectivity may require the use of expensive and time-consuming analytical methodology, and can be increased often only at the expense of sensitivity. Most often a compromise is necessary among sensitivity, specificity, and the cost of analysis.

The quality and reliability of the obtained result are always of prime interest. In research, one has to establish (and maintain and prove) the reliability of analysis; in many cases (in the majority of clinical and pharmaceutical applications) one has to comply with regulative and administrative requirements as well. The latter requirements are often in the form of good laboratory practice (GLP) requirements, analogous to good clinical practice in a hospital environment.

Quality of analysis is characterized by accuracy, precision, reproducibility, and repeatability. *Accuracy* is the degree of agreement of a measured quantity to its actual (true) value. Unfortunately, in the biomedical field, the "true" values are often not known. To overcome this problem, a "consensus" value is often used. This does not necessarily represent the "true" value (of a given property of, e.g., a well-defined standard sample), but is an estimate of the "true" value accepted by the scientific community. In such a case accuracy is defined as the degree of agreement of a measured quantity to its accepted "consensus" value. The object is to make results obtained by diverse techniques, methodologies, and laboratories comparable. *Precision* characterizes the degree of mutual agreement among a series of individual measurements under the same conditions. *Repeatability and*

reproducibility are similar to precision, but are more narrowly defined. All are statistical parameters, usually expressed in units of *standard deviation* or relative standard deviation. Repeatability relates to the standard deviation of a series of replicate measurements performed by the same person, using the same instrument, and under the same conditions. Reproducibility is also the standard deviation of a series of replicate measurements, but in this case different persons and different equipments may be involved. *Validation* is also a commonly used term (most often used in chromatography and in the pharmaceutical industry). This refers to establishing evidence that a given analytical process, when operated within established parameters (i.e., using solvent composition with 1% reproducibility), will yield results within a specified reproducibility. *Robustness* is a related term indicating the resilience of a method when confronted with changing conditions.

Speed is another characteristic of the analytical process. One aspect is *sample throughput,* which may vary from one sample per week to thousands or millions of daily analysis. Another aspect is the *time delay* between sampling and obtaining the result of analysis. Chemical and biochemical analyses are usually fast and typically require seconds or hours to perform. This is in contrast to several biological tests, which often need time (days) for growing bacterial cultures. This time delay may be a significant factor for selecting proper treatment in serious illnesses.

The *cost of analysis* is also of critical importance, which is closely related to the number of samples analyzed. Development of analytical techniques is always expensive and time-consuming. This is the major part of the total cost when analysis of only a "small" number of samples is required. Analysis of 10–100 samples or 10 samples per month is usually considered a "small" number compared to more than 1000 samples or 100 samples per month ("high throughput"), although analysis of 100,000 samples per year is also not uncommon. When analyzing large number of samples, the major part of the cost comprises labor, consumables, and instrument time, usually in this order. For high-throughput experiments, therefore, it is always worth investing money and effort to simplify sample preparation and to speed up analysis, even if this would necessitate using expensive instrumentation and strict quality control.

3. The analytical procedure

All projects require proper strategy and careful planning to be successful. This relates to analytical chemistry as well, so all steps need to be carefully considered. The analytical procedure consists of several distinct steps; the most important are (1) sampling, (2) preparation, (3) separation, (4) analysis, and (5) evaluation. Some of these may not be needed and some may be performed in a single step. For example, when sample preparation is efficient, separation may not be necessary, while separation and analysis may be performed in one step using online combination of,

e.g., gas chromatography and mass spectrometry (GC–MS). Here we give a short outline of the analytical process; a more detailed discussion will be presented in Part II of the book (Tools of the Trade), while particular issues will be discussed in subsequent chapters.

(1) *Sampling:* This is the first step in any analysis. In clinical studies it is most often performed by an MD or by a nurse (e.g., taking blood or tissue samples). Sampling may seem to be straightforward, but it is a critical step, which has to be designed carefully and executed accurately. The sample taken should be representative—easy for biological fluids and not trivial for solid samples such as tissues. Analysis often uses internal standards; these should be added to the sample as soon as possible, preferably immediately after sampling. Samples may be changed or contaminated during sampling, which should be taken into account and minimized. For example, blood samples are typically taken into vacuum tubes, but these (and often the syringes used) may contain heparin or other substances to prevent clotting. Although this may be necessary, it will contaminate the sample, which has to be taken into account. The samples often are stored before analysis, occasionally even for years. Sample composition may change during storage, which has to be minimized (and/or taken into account). The simplest and usually safest way to store samples is freezing them: most samples can be stored for days or weeks at −20°C. Storage at −80°C is safer; most samples can be stored under such conditions for several years without change.

(2) *Sample preparation:* The aim is to make the sample more amenable for subsequent analysis. This often means removing part of the sample, e.g., by centrifugation (to remove cells and aggregates from blood) or by extraction (which removes or enriches certain types of molecules). Note that sample preparation always changes sample composition and this has to be taken into account in the evaluation phase. Often several preparation steps are performed in succession, such as centrifugation, filtration, extraction, derivatization, another extraction, etc., to sufficiently simplify the complexity of the sample and to ensure the success of analysis. Sample preparation is time-consuming and often labor intensive. The current trend in biochemical analysis is to use a complex, high-quality (and expensive) instrumentation to allow simplification of the sample preparation process.

(3) *Separation:* The classical approach to analysis is first to separate mixtures into its individual components (compounds) and then proceed with identification, structure determination, and quantitation. High-quality analytical methods are now often capable of dealing with mixtures of compounds, so complete separation of mixtures into individual components is no longer necessary. Many modern analytical instruments consist of a combination of separation and structure characterization

methods, such as high-performance liquid chromatography–mass spectrometry (HPLC–MS; HPLC to separate the sample and MS to characterize or identify the separated compound). Separation methods most often mean chromatography, and these two terms are often (but inaccurately) used as synonyms. Prerequisite of chromatography is that the sample needs to be soluble (or vaporizable). Insoluble and nonvolatile particles cannot be separated by chromatography. The most common chromatographic methods are the following:

(a) Gas chromatography (GC) is very efficient for separating volatile compounds. Volatility of some compounds may be increased by derivatization. As most molecules of biochemical or clinical interest are nonvolatile, and derivatization has many drawbacks and is not always possible, GC has a limited (but important) scope in the biomedical field.
(b) Liquid chromatography (LC) is widespread, has many different versions, and can be used to solve a variety of problems. These are well suited to analyze most samples including polar and ionic compounds. The most common chromatographic method is HPLC.
(c) The methods of choice to separate macromolecules such as proteins and nucleic acids are gel-based electrophoretic methods. These can be performed in one or two dimensions (in a tube or on a chip or on a 2D plate). These form the basis of most DNA and RNA diagnostics.

(4) *Analysis and detection:* It is the high point of an analytical process. The simplest and probably oldest version is densitometry or spectrophotometry, which measures light absorbance at a particular wavelength. Signal intensity characterizes the sample amount. Most samples absorb UV light, so it is typical to use a UV lamp for spectrophotometry (e.g., at 254 nm). It is also possible to scan over a range of wavelengths, which yields the UV spectrum, which in turn characterizes the molecular structure. Spectrophotometry can be performed after separating a mixture using chromatography. The time necessary for a sample to pass through the HPLC system (called retention time) depends on the molecular structure, and can also be used for compound identification. Signal intensity (like in conventional spectrophotometry) characterizes the sample amount.

UV detection is quite common, but in many cases it is not sufficiently selective: even combined with chromatography, it often leads to false-positive or false-negative results. For this reason many other types of detectors are used in analytical chemistry, to increase selectivity, specificity, or sensitivity. To identify or determine the molecular structure, the use of spectroscopic techniques is common. Mass spectrometry, the main topic of this book, is among the most commonly used and highest performance methods. Infrared spectroscopy (IR) and NMR are also often used, although the relatively low sensitivity of NMR restricts its use in the biomedical field.

(5) *Data evaluation:* First, the signal detected during analysis needs to be evaluated in terms of structure determination of unknown components, and identification and quantitation of various known (or presumably present) components—this is an integral part of the analytical process. Second, the results obtained this way have to be evaluated in terms of biomedical relevance. The latter involves mathematical or statistical procedures, often referred to as *"chemometrics."* To be efficient, a joint effort of chemists, biochemists, analytical specialists, statisticians, and medical doctors is required. It is highly advantageous that these specialists communicate efficiently and have at least a superficial knowledge of each other's specialty.

4. A case study: analysis of plasma sterol profile

The analytical process discussed above can be illustrated by an example of determining plasma sterol concentrations, which was published recently (see ref. [1]). The purpose is not to go into detail but to illustrate the various aspects of analytical work. Before starting the analytical procedure, the study needs to be carefully planned: the objective was to develop a method capable of determining plasma concentration of various sterols to study cholesterol metabolism and related diseases. It was decided to determine plasma level of desmosterol, lathosterol (precursors of cholesterol synthesis in the liver), cholestanol, and β-sitosterol (sterols present in food but not synthesized in the human body). The analytical challenge was that these sterols need to be separated from cholesterol (in fact, lathosterol and cholesterol are closely related isomers) at a 1000-times lower concentration in plasma than that of cholesterol (μmol/l vs. mmol/l). It was determined what patient and control groups were needed and what was the minimum number of people for a meaningful pilot study (10 in each group but would be much higher in a full-scale project).

Analytical chemistry starts only after this phase. As it is a multiple-step procedure, first it is decided that the main strategy is to use a *relatively* simple sample treatment and follow it by a highly efficient HPLC–MS analysis. This offers the possibility of relatively high throughput (hundreds of samples analyzed) with medium time and cost requirement. In the present example, the following analytical procedure was developed: *sampling* consisted of taking 5-ml blood samples from each individual. *Sample preparation* started by centrifugation to obtain plasma, which was stored at −80°C until utilized. (Note that obtaining plasma from blood is often considered part of the sampling process, as it is typically done in the same laboratory.) Plasma samples were thawed, and 50-μl aliquots were used for analysis. First, the *protein content was precipitated* (by adding methanol, centrifuging and pipetting the supernatant clear liquid, and finally diluting it with water). This is one of several well-established procedures to separate macromolecular components from plasma. Further sample cleanup was performed by *solid-phase extraction (SPE)*. It can be viewed as a simplified chromatographic

separation and is one of the most common sample preparation methods. For solid-phase extraction of sterols from plasma the following method was developed: C18ec-type cartridges were used, which were preconditioned (first by MeOH, then by MeOH/water mixture). Diluted plasma samples were applied to the cartridges, washed with MeOH/water mixture, and briefly dried in vacuum. The sterols were then eluted with a mixture of MeOH/acetone/*n*-hexane (of course, selection of solvents and solvent ratios are of critical importance and were optimized). The eluted substances were dried and the residue was dissolved in MeOH. This SPE process resulted in a clear liquid, which did not contain macromolecules and was enriched in sterols. Efficiency of the sample preparation process was controlled using various standards (i.e., to check that the sterols were not lost during preparation, but the amount of interferences was reduced).

Separation and analysis were performed in one step using an online coupled HPLC–MS instrument. Both chromatographic and mass spectrometric methods needed to be developed. To separate various sterols from each other and from various other compounds present in the prepared sample, a novel, *reverse-phase HPLC* method was developed. This involved using an RP-18e column of 3 μm particle size. Initial solvent composition was methanol/water, which was changed (in two fast steps) to methanol/acetone/*n*-hexane. As it is typical in the biomedical field, HPLC alone was not sufficient to separate all compounds completely, even after the sample preparation discussed earlier. To increase selectivity, *mass spectrometry* was used and likewise optimized. Best results were obtained by atmospheric pressure chemical ionization in positive ion mode; the most characteristic ion for sterols was formed by water loss from the protonated molecule, which was used for quantitation. Using mass spectrometry signals of the various sterols were separated from each other and from that of interfering compounds (not resolved by chromatography). Cholesterol and its isomer lathosterol gave identical spectra (even in tandem mass spectrometry). Separation of these isomers was the main reason to develop the novel chromatographic method discussed earlier.

The first phase of *data evaluation* was to determine plasma concentration of sterols analyzed as described earlier. The standard addition method was used, calibration curves were constructed, and plasma sterol concentrations were determined. The second phase of data evaluation was to look for characteristic biomarkers and separate patient groups based on sterol concentrations. This was done by applying chemometrics (e.g., linear discriminant analysis) with sufficient validation. It was found that sterol concentration ratios are much more characteristic disease markers than the individual concentrations. For example, the concentration ratio of desmosterol to sitosterol was a much better marker of cholesterol-related disorders than the cholesterol concentration itself, and the concentration ratio of lathosterol to total plasma cholesterol was an excellent marker of statin treatment. Application of these analytical results by biochemists and medical doctors will hopefully result in better treatment of patients.

Analytical chemistry is, however, time-consuming, and method development needs highly trained personnel. The study discussed above required original ideas and a significant amount of method development: about two months of work for two scientists at the PhD level. Application of the method, however, is much more straightforward. Using high-quality conventional equipment (not designed for high throughput), a good technician or PhD student can prepare 20 samples per day, which can easily be measured in a day using a good-quality HPLC–MS instrument—that is deemed perfectly adequate for the present purpose. This type of research requires expensive instrumentation; the cost of an HPLC–MS instrument is in the range of $100,000–500,000 depending on its capabilities. Sample preparation requires less investment; $100,000 is a reasonable figure for purchasing the various small instruments needed in such a lab.

When high throughput is desired, suitable equipments are needed, but this way 100 or 200 samples may be prepared in a day, and this process can even be performed by robots (further improving throughput). Measurements by HPLC–MS can also be automated and accelerated. Throughput in this case mainly depends on the length of chromatography (this is the reason for the current trend of trying to substitute HPLC–MS by MS/MS, whenever possible).

5. Mass spectrometry

A mass spectrometer is a very special kind of balance, which measures the mass of molecules and their subunits. It can be used to *characterize and identify compounds,* to *detect trace-level components,* and to *measure their concentration* in complex matrices. Mass spectrometry will be described in detail in Chapter 6; here only a very brief introduction is presented.

Mass spectrometry yields a *mass spectrum* (or spectra) of a compound, which establishes its *molecular mass* and the characteristics of the *molecular structure*. It is among the *most sensitive* molecular probes, which can detect compounds in femtomol, attomol, or even zeptomol amounts (10^{-15}, 10^{-18}, 10^{-21}). Peak intensities are proportional to the amount of the material or concentration of the compound present (such as light absorption in photometry). This is the basis of *quantitative measurements*. Mass spectrometry is also *very selective,* so trace components may be analyzed in the presence of a large amount of matrix. *Tandem mass spectrometry (MS/MS)* is also commonly used. This increases the amount of structural information obtained and the specificity of analysis. As a consequence, the chemical noise decreases, which improves detection limits. Using high-specificity MS/MS techniques often allows simplification of sample preparation procedures. *High-resolution* mass spectrometry also increases specificity of analysis and allows determination of the accurate mass of a molecule. This establishes the *elemental formula* of an unknown molecule.

Solids, liquids, and gaseous samples can all be analyzed by mass spectrometry. The sample can be inserted into the mass spectrometer as they are or after sample preparation. This way of sample insertion is best suited to study pure samples, but mixtures can also be analyzed in this way. *For studying mixtures* (e.g., biological fluids) it has become a common practice to couple chromatography and mass spectrometry online. In such a case the sample is fractionated by chromatography and the individual components eluting from the chromatographic column pass directly into the mass spectrometer, where detection and structure analysis are performed. The first such successful combination has been *gas chromatography–mass spectrometry* (GC/MS or GC–MS), which is widely used in the biomedical field for at least 20 or 30 years. GC–MS is well suited to study relatively volatile compounds (not ionic and not very polar compounds up to approximately 500 Da molecular mass). Most biologically important molecules are polar, so derivatization of the sample is often necessary to make them amenable for GC–MS. To overcome this problem, the use of another combination *HPLC–MS* has become most common, and is still gaining ground. It is an excellent method to study polar and even ionic molecules and requires less sample preparation than GC–MS.

Its high sensitivity, high specificity, and straightforward coupling to chromatography make mass spectrometry one of the best and most widely used techniques in analytical chemistry. It is the method of choice for analyzing minor components in complex matrices, both for qualitative analysis and for quantitation. It is often used when chromatography is not sufficiently selective (there are too many peaks or the chemical background is too high) or yields equivocal results. Mass spectrometry is among the highest performance analytical tools in the biomedical field, and mass spectrometry-based methodologies are often considered as "gold standards." Mass spectrometry is widely used in the pharmaceutical and biomedical fields.

To help orient the reader, a few typical applications are listed below:

(a) Determination of the *impurity profile,* i.e., detection and quantitation of impurities. It is a typical problem in the pharmaceutical field, but also in many other areas. All impurities (typically down to 0.01%) need to be identified and often quantified. Mass spectrometry (GC–MS, HPLC–MS, HPLC–MS/MS) is the method of choice, especially at low concentrations.

(b) *Quantitation of impurities,* usually at the trace level. It is similar to that discussed above. Only selected (predefined) target compounds are studied, but their concentration may be much lower than 0.01%. A typical case is doping control; another application field is forensic analysis. Mass spectrometry (GC–MS, HPLC–MS, HPLC–MS/MS) is the method of choice, especially at low concentrations.

(c) *Studies on metabolism*. The structure and amount of drug metabolites (in blood, urine, and faces) need to be determined prior to phase I clinical

studies, and later on in human volunteers as well. Mass spectrometry is one of the key techniques in this field.

(d) *Pharmacokinetic studies*. The object is to monitor the time dependence and clearance of drug concentrations usually in plasma. In simple cases this is performed by chromatography, the more challenging problems are usually solved by GC–MS, HPLC–MS, or HPLC–MS/MS techniques.

(e) *Therapeutic drug monitoring*. Plasma level of various drugs is monitored in patients. This is more and more often being used in the clinical field, especially in cases where the therapeutically necessary and toxic concentrations are close to each other.

(f) *Neonatal screening,* i.e., studies on metabolic disorders. Various small molecules (amino acids, fatty acids, steroids, etc., commonly called metabolites) are determined in biological fluids, usually in blood. Some of these molecules may have abnormally high or abnormally low concentration, indicative of an inherited metabolic disorder.

(g) *Proteomics* is a popular and fast-developing field. Mass spectrometry combined with chromatography (most typically 2D gels) is the prime analytical method to identify proteins, and to study protein expression and *posttranslational modifications*. The proteome (all proteins present in a sample, e.g., tissue or cell culture) reflects the current state of the organism and yields valuable information on the physiological state, disease progression, etc.

(h) Analogous to proteomics, all metabolites (i.e., practically the assembly of all small molecules in a cell or tissue) represent the "metabolome" and are studied by *"metabolomics."* These also reflect the state of the organism, and one of the prime techniques in these studies is mass spectrometry, most usually HPLC–MS.

(i) There are other analogous applications, studying the assembly of a given class of molecules in an organism, and these are often called *"–omics"* (such as lipidomics, glycomics, etc.). Mass spectrometry plays an important role here as well.

Reference

1. Nagy, K., Jakab, A., Pollreisz, F., Bongiorno, D., Ceraulo, L., Averna, R.M., Noto, D. and Vekey, K., Analysis of sterols by high-performance liquid chromatography/mass spectrometry combined with chemometrics. *Rapid Commun. Mass Spectrom.,* **20**, 2433–2440 (2006).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 3

Ethical, legal, safety, and scientific aspects of medical research

ANDRÁS TELEKES[a,*] and KÁROLY VÉKEY[b,1]

[a]*National Institute of Oncology, Budapest, Hungary*
[b]*Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary*

Regardless of whether it is a clinical trial or analytical investigation of a sample (e.g., plasma), the subjects of medical research are human beings. Ethical considerations are therefore important; they influence decision making and are well regulated in the medical field.

The basic aim of medical research is to improve clinical practice, and this should be evidence based, if possible. Clinically relevant research evidence may relate to basic medical science, but it especially relates to patient-centered clinical research, e.g., to the study of the accuracy and precision of diagnostic tests, the power of prognostic markers, and the efficacy and safety of therapeutic, rehabilitative, and preventive regimens. New evidence from clinical research at the same time invalidates previously accepted diagnostic tests and treatments and replaces them with new ones that are more powerful, more accurate, more efficacious, and safer [1]. Ethical decision making is based on the Declaration of Helsinki,

*Corresponding author. Tel.: +36-204-119080; Fax: +36-139-52835.
E-mail: andras.telekes@gmail.com (A. Telekes).
[1]E-mail: vekey@chemres.hu (K. Vékey).

although research methodology should usually be conducted according to Good Clinical Practice (GCP). Laboratories contributing to medical research should comply with rules and regulations relating to human samples, the aspects of which are well regulated in most countries (although the respective laws may be different in various countries).

Despite significant advances in medicine and the fast improvement of technology, clinical decision making is still the cornerstone of medical practice. Medical decision making is challenging since it involves problem identification, selection, and evaluation of diagnostic information and a choice among various possible interventions. Note that medical decisions are sometimes based on ambiguous background since our knowledge quickly changes, data are often contradictory or may not be available, and the validity and reliability of even published data may be uncertain. Medical decision making is further complicated by biological variation of diseases and by differences in preferences and values among various patients. Uncertainty of clinical decision making is an inherent part of clinical practice and a possible source of bias in clinical trials.

In the present chapter we summarize fundamental ethical, legal, and safety-related aspects of medical (and especially clinical) research. These are well known for medical professionals, but chemists and biologists may be less familiar with these aspects. Nevertheless, it is essential that all persons working in studies related to medical research should be familiar with the basic concepts and rules that apply.

1. Ethical aspects

The intersection of ethics and evidence and the context of scientific uncertainty relate to the problem of ethical decision making [2]. Since uncertainty is an inherent part of nature, one can never be sure to prevent harm from occurring. Medical practice, typically and unfortunately, requires judgments under uncertainty. This is the reason why ethical aspects compete (and sometimes override) scientific points of view. The ethical aspect of medical science requires that every possible step should be made to prevent harm from occurring, which includes careful consideration of all available data. A study is considered unethical if the potential harm overwhelms the true benefit to patients or healthy volunteers. This also means that initiation of a clinical study without sufficient preclinical data (e.g., short- and long-term toxicity profile, dose–effect and dose–toxicity relationships, dose-limiting toxicity, pharmacokinetics, etc.) is unethical since the potential harm cannot be properly estimated. Therefore, every clinical trial requires a detailed trial design, including careful assessment of all preclinical evidence and whether it is ethically acceptable for patients or healthy volunteers to participate in it in the proposed fashion. It is of great importance from the ethical point of view to avoid any unnecessary suffering or other inconvenience of the involved participants.

The balance between achieving medical progress and ensuring individual patient care and safety is an ethical dilemma of medical research [3]. Thus, clinical trials require a delicate balance between individual and collective ethics. On the one hand, individual ethics means that each patient should receive the treatment, which is believed to be the most appropriate for his condition. On the other hand, collective ethics is concerned with achieving medical progress in the most efficient way to provide superior therapy for the future patients. It is of importance that the real interest of the participating individual should never be sacrificed for possible benefits of future patients! As stated in the Declaration of Helsinki: *"In medical research on human subjects, considerations related to the well-being of the human subject should take precedence over the interests of science and society"* [4]. In order to provide medical progress, one needs collaborating patients. To settle this ethical paradox usually two principles should be met. Patients can only be involved in therapeutic trials if the efficacy of available treatments is insufficient (e.g., the patient is incurable), and enrollment always should be voluntary, based on the free will of the informed patient. Any pressure put on the patient to obtain his or her consent is unethical (even if well meaning and true, e.g., referring to her children who might benefit from the result of the trial).

Because of the complex nature of these issues, there are well-established ethical guidelines and statements even for special situations [5–9]. To assure compliance with these guidelines, ethical committees are formed in most countries, which have to approve and might have the right to control clinical trials. Ethical committees are usually made up of clinicians (who are not involved in the trial), other professionals (such as spiritual counselors, lawyers, psychologists, statisticians, etc.), and laypeople. Thus, all protocols are subjected to profound social control and sound judgment by the committee members representing different aspects of society. In the committee, clinicians explain clinical implications and technical aspects of each protocol. Ethical committees may be local (i.e., at the hospital where the trial is to be carried out), regional, or national. All clinical trials (and other types of clinical research projects) need to have their protocol approved by such a committee before the trial is started. In the case of a multicenter trial, either the regional or the national committee grants the permission, or each collaborating partner must have approval from its local ethical committee. Note that not only the trial's design but also all details of conducting the trial need to be approved, since these may affect the individual patients. Maintenance of high ethical standards cannot be achieved by purely administrative procedures, so it is the job of all clinical investigators to make sure that his or her patients do not suffer as a consequence of clinical research. There are ethical implications of substandard research as well [10]. For example, it is unethical to misuse patients by exposing them to unjustified risk and inconvenience, or to publish misleading results that may promote further unnecessary work.

2. Legal aspects

The ethical issues discussed above are, in most countries, codified in legal form. Treatment of human beings and human samples depends on the legal environment. Research objectives may also be subject to legal issues, like in cloning or stem cell research.

Tissue research (including bone marrow, blood, urine, sputum, etc.) is currently regulated by distinct and sometimes contradictory laws and regulations. The legal edicts that determine the research on human specimens are clearly affected by many factors including policy decisions, cultural, religious and moral issues, jurisprudence, etc. In a recent review, a comparison of the laws in the U.S. and Europe regarding the use of human biological samples in research was presented [11]. Since there are a wide variety of laws to be applied, international collaborative research should take these differences into account, especially those affecting how to obtain, transfer, and investigate the human samples. In all cases researchers should be alert to implement all the local laws and regulations.

In most countries there are offices, institutes, and legal bodies relating to medical research. For example, in the U.S., the Office for Human Research Protections of the Department of Health and Human Services is responsible for the federal policy of human subject protection, and the Food and Drug Administration for research on products. In case of any legal or ethical doubt, it is often worthwhile asking for their advice or approval.

3. Safety aspects

Handling biological material always raises the issue of safety for the personnel involved. Chemical hazards and safety procedures relating to these are well known for chemists. Regarding biological hazards, first the staff participating in the research should be aware of them, and second, adequate precautions should be taken and the personnel should be trained on how to handle human samples safely.

As a principle, all human specimens should be regarded as potentially infectious. Detailed tests for pathogen profiles are rarely carried out (e.g., on hepatitis, HIV, etc.), and they can never be complete. As a general precaution, all personnel working with human samples (e.g., blood or body fluids) should be immunized against hepatitis B (and sometimes against hepatitis A as well). The application of other vaccines is optional (e.g., against typhoid fever and tetanus) depending on the circumstances. It is important to note that immunization must never be considered as a substitute for safe working practices.

Laboratory personnel should avoid any direct contact of skin and mucous membranes with the human specimens including blood or blood products, excretions, secretions, tissues, or other biological materials. Note that most accidental personnel

Table 1
Classification of laboratories according to safety hazards

Chemical safety levels (CSL) [12]	Biological safety levels (BSL) [13]
CSL 1 Use of chemicals is highly restricted; hazardous chemicals cannot be used or stored. Low risk of exposure due to the strong restriction for using hazardous chemicals.	*BSL 1* The agents used are well characterized and are not associated with disease in healthy adult humans. The potential hazard is minimal to laboratory personnel and the environment.
CSL 2 Use and storage of hazardous chemicals is restricted. Moderate risk of exposure, controlled by limiting the use of hazardous chemicals.	*BSL 2* The agents used can cause human disease. The potential hazard is moderate to personnel and the environment. Treatment or prophylaxis is available. Risk of spread is limited.
CSL 3 Use of chemicals is generally unrestricted; the use of hazardous chemicals is restricted to closed environment. Substantial risk of exposure, controlled by stringent engineering controls, by minimizing the use and storage of hazardous chemicals, and by carefully reviewing work practices.	*BSL 3* The agents used are indigenous or exotic, which may cause serious disease. The potential hazard to personnel and the community (in case of spreading) is serious. Usually there is effective treatment or prophylaxis available. The infection usually does not spread by casual contact.
CSL 4 Use of chemicals is unrestricted, hazardous chemicals are frequently used. There is possibility of high risk of exposure and contamination during operations, controlled by stringent engineering controls and design requirements of such facilities and by carefully defining and monitoring work practices.	*BSL 4* Dangerous and exotic agents can be used that can cause severe or lethal human disease. The individual risk of aerosol-transmitted laboratory infections and life-threatening disease is significant. Usually there is no effective treatment or prophylaxis available. The infection could be transmitted directly from one individual to another or from animals to humans.

contaminations are due to sharp items such as needles. Laboratories working with chemical and/or biological materials are classified according to the level of hazard (Table 1).

The classification of laboratories using radioactive material is usually based on the relative radiotoxicity per unit activity. One of such classification is presented in Table 2.

Low-level laboratory correspond to a CSL 1/CSL 2 chemical laboratory (e.g., normal ventilation is usually sufficient). Intermediate-level laboratory is specially designed for the use of radioisotopes. High-level laboratory is engineered for handling radionuclide materials with high activities. High-level laboratories must be kept at a slightly negative pressure, and personnel working in it should be monitored.

The least strict is safety level 1, and the most dangerous level is safety level 4. Working in different safety level laboratories requires special rules, safety

Table 2
Classification of laboratories according to limitation on activities [14]

Type of laboratory	The limit of use (mCi)			
	Low hazard (examples[a])	Intermediate hazard (examples[b])	High hazard (examples[c])	Very high hazard (examples[d])
Low-level radioactive materials laboratory	<5	<0.5	<0.05	<0.01
Intermediate-level radioactive materials laboratory	5–50	0.5–5	0.05–0.5	0.01–0.1
High-level radioactive materials laboratory	>50	>5	>0.5	>0.1

[a] Co_m-58; Cs_m-134; Cs-135; Ge-71; H-3; I-125; In_m-113; Kr-85; Nb-97; Ni-59; O-15; Pt_m-193; Pt_m-197; Rb-87; Re-187; S_m-147; Sr_m-85; Tc_m-96; Tc_m-99; U-238; U-Nat; Y_m-91; Zn-69; Zr-93.
[b] As-73; Be-7; Ba-133; C-14; Ca-47; Cd-109; Co-57; Co-58; Cr-51; Cs-137; Cu-64; Fe-55; Fe-59; Gd-153; Hg_m-197; I-129; I-133; Ir-190; K-42; Kr-85; Mn-54; Mo-99; Na-24; Ni-63; P-32; P-33; Pm-147; S-35; Se-75; Sr-85; Sr-89; Tc-99; Xe-133; Y-90; W-181; Zn-65; Zn_m-69.
[c] Ba-140; Bi-207; Ca-45; Cd_m-115; Ce-144; Cl-36; Co-60; Cs-134; Eu-152; Eu-154; Ge-68; I-125; I-131; Ir-192; Mn-54; Na-22; Ru-106; Sb-124; Sr-90; Th-232; Th-Nat; Te_m-127; TI-204; U-236; Zr-95; Y91.
[d] Ac-227; Am-241; Cf-252; Cm-243; Cm-244; Np-237; Pb-210; Po-210; Pu-236; Pu-238; Pu-239; Pu-242; Ra-223; Ra-226; Ra-228; Th-228; Th-230; U-232; U-233; U-235.

equipments, and various permits. The operations of chemical and biomedical laboratories are usually strictly controlled by internal regulations and external authorities, but discussing these is outside the scope of the present chapter. Here we list some general advice for working in biomedical laboratories:

(1) Used needles and other sharp objects should not be sheared, bent, broken, recapped, etc., by hand.
(2) All needles and sharps objects should be discarded in rigid, puncture-proof containers.
(3) Safety gloves, coats, gowns, or uniforms should be worn while working with potentially infectious materials. Never wear these outside the working area, and these should be changed at least once per week. If they are obviously contaminated, they should be decontaminated (by autoclaving) prior to laundry, or disposed off as hazardous waste.
(4) Gloves should always be worn for those manipulations that might lead to direct contact with potentially infectious specimens. Never leave the work area in these gloves. Disposable gloves should be collected and disposed off as hazardous material.
(5) All procedures with potentially infectious materials should possibly be carried out without creation of aerosols. Face shields and masks should be worn during all manipulation where a “splash” hazard exists. Eye

protection is required for all personnel in all locations where chemicals are used or stored.

(6) Never use mouth pipetting; mechanical pipetting devices should be used for the manipulation of all liquids.

(7) Appropriate disinfectant should be applied following any spill of potentially infectious materials or at the end of daily work to decontaminate laboratory work surfaces. Spill kits (containing absorbent pads and/or neutralizing agents) should be prepared for all frequently used chemicals or biological waste.

(8) Before disposal, laboratory waste should always be decontaminated by autoclaving and all waste should be identified by unambiguous labeling whether or not the decontamination had been carried out. Remains of biological material (e.g., blood or tissue samples) should be regarded as hazardous material. Chemicals should be treated as hazardous if they are ignitable, corrosive, reactive, or toxic.

(9) All chemicals should be stored properly and according to compatibility (i.e., acids and solvents must be stored separately). Chemical waste should never be mixed (i.e., mixing acids and alkaline liquids could lead to heat generation and violent reaction).

(10) In each laboratory using chemicals, safety shower, emergency eyewash, fire blanket, and extinguishers should be provided. All emergency equipments should be maintained in proper working order and their access should not be obstructed.

(11) All personnel should be trained in emergency procedures and they must be informed about the locations of emergency equipments. Telephone numbers of emergency contacts should be available.

(12) There should be adequate space and shielding for radioactive materials used in the laboratory as well as for storing radioactive waste.

(13) The access to the laboratory should be limited during operations.

(14) Before leaving the laboratory each staff member should wash their hands with soap and water. Personnel working with radioactive materials are required to survey themselves when leaving the laboratory.

(15) In laboratories, storing food, eating, drinking, smoking, or applying cosmetics is prohibited.

(16) Signs should always be posted at the entrance of the laboratory identifying biological/chemical hazards.

4. Handling biological materials

Sample handling is always a crucial part of analysis. In the biomedical field, there are several issues not commonly encountered in general analytical work, and these will be described in this section. Biological materials are potentially hazardous;

this aspect has been discussed above. Using human samples for research is an ethical issue as well; thus, operational and ethical guidelines apply [15].

To obtain human samples (e.g., blood), first the patient should be informed and voluntarily consent to sample collection (as discussed earlier concerning ethical and legal aspects). The possibility to obtain repeated samples is limited in many cases since condition of the patient may change or a given volunteer may no longer be available. Every precaution should be made to process all samples in as similar manner as possible, including the samples taken from different study groups. Sample collection and analysis are often separated in space and time, which make issues regarding sample handling, labeling, transporting, and storage very critical, and sample flow needs to be organized. Samples often need pretreatment before storage (e.g., centrifuging blood to obtain plasma), and the pretreated sample is stored until further analysis. Samples are often stored for a very long time (even years) before analysis; avoiding sample deterioration is therefore a significant issue. To avoid errors, it is essential that issues relating to sample collection and sample handling should be described in detail for any clinical trial and a standard operating procedure (SOP) should also be developed describing all particulars both for volunteers and for staff. Note also that hospital nurses are often overworked, and may deviate from the given SOP. For this reason it is often advantageous to assign a study-nurse for a given project, whose only job is sample collection. In complex or long-term studies, it is often helpful to run a pilot study to assess the best conditions for sample collection, storage, and organization of the workflow.

Univocal communication between study subjects, medical staff, and researchers are essential for reliable and consistent sample collection. Clear-cut instructions regarding the timing of collection, specific containers to be used, sample volumes required, sample handling (e.g., place on ice until it is transferred to the study personnel), etc., are the foundation of reliable results. The study scientists must also deliver easily understandable instructions to patients and healthy volunteers on how to prepare for the trial (e.g., fasting before blood withdrawal). It is a good practice to give all information in writing to everyone participating in the study, such as nurses, physicians, research staff, and sometimes the subjects as well. It is critical to explain the importance of precisely following sample collection protocols to the personnel who are responsible for it since deviation from the protocol could ruin the whole study.

Sample collection may be noninvasive (e.g., urine, feces, sputum, collection of exfoliated cells with buccal swab) or invasive (e.g., blood taking and biopsy). The measured parameter may show time dependence; in such cases the sample should be taken at various intervals and the time course of the parameter (such as drug clearance) should be determined. Biological fluids are (in most cases) homogenous. For nonhomogenous samples (such as tissues), it is important to establish that the sample taken is representative and for all persons involved in the study the same type (or same fraction) of sample is studied.

For biological samples their stability is a critical issue. There are various factors influencing stability, and these should be carefully controlled: (1) presence of anticoagulants; (2) endogenous degrading factors, such as proteases or other enzymes; (3) stabilizing agents, such as protease inhibitors; (4) sterility; (5) temperature; (6) time before preprocessing (such as centrifuging); and (7) storage time [16]. To decrease sample deterioration often various additives are added and the sample is cooled down or frozen. Although sample pretreatment is nearly always essential, these factors do change and may deteriorate the sample in some manner. Defining sample treatment and storage conditions is therefore an essential part of study design. Many cases are known in which large and expensive trials went amiss due to sample deterioration. To complicate matters, optimum sample pretreatment and storage often depends on the type of analysis desired. Often (as mentioned earlier) a pilot study is helpful to define optimum conditions.

To cite a few examples, anticoagulants must nearly always be added to the blood sample. There are various anticoagulants, for example, heparin, citrates, or EDTA. Best quality of RNA and DNA samples may be obtained from citrate-stabilized blood, but it may lead to a higher yield of lymphocytes for culture. On the contrary, heparin-stabilized blood could influence T-cell proliferation, and moreover heparin binds to many proteins and may therefore compromise proteomic studies. EDTA is suitable for both DNA assays and proteomics, but it affects Mg^{++} concentration causing problems for cytogenetic analyses (e.g., decreases mitotic index).

The time between sample collection and analysis is called holding time. This is the sum of the transportation time (from the location where the sample was obtained to the laboratory) and the storage time (keeping the sample in the laboratory prior to analysis). Quickly reducing the temperature of the biological samples in order to minimize deterioration is frequently required. Since it is difficult to control temperature outside the laboratory, transportation time (e.g., from operating theatre to laboratory) may be critical. In an optimal case, the sample is separated immediately after collection into different components (e.g., plasma and cells) and each of them is kept at the most appropriate temperature. For proteomic studies, freezing the plasma to −80°C is ideal, whereas for DNA and RNA profiling freezing the cell should be avoided, if possible.

The intended storage time also influences what temperature should be considered optimal. Isolated DNA may be stored at 4°C for several weeks, at −20°C for several months, and at −80°C for several years. On the contrary, isolated RNA should always be kept at −80°C. Live cells are stable at room temperature up to 48 h, but after that they must be either cultured or cryopreserved in liquid nitrogen at −150°C in order to remain alive. Since serum and plasma contain a large amount of soluble molecules, it should be kept at very low temperature (−80°C) to remain intact. At this temperature, plasma is regarded stable (in most respects) for several years.

Proteins are sensitive to degradation by proteases; thus, if the cells are damaged the result of the assay may be misleading. To avoid this problem the proteins

should be protected by applying commercially available protease inhibitors (such as 1 μg/ml pepstatin, etc.) On the contrary, protease inhibitors are toxic to cells; thus, they must not be added if cell viability is prerequisite for the assay.

There are several other precautions to consider. Cooling/freezing the samples may be necessary not only during storage but also during sample preparation. For example, centrifugation should be performed in a cool environment, for which special equipment (chilled centrifuge) is available. Light also degrades the sample, so it is often recommended to store/transport samples in brown or amber glass containers. Some compounds may be absorbed on surfaces (like glassware), so containers need to be made of plastic (usually polypropylene). Some studies/samples are sensitive to microbes, so the use of sterile equipment may be needed.

All aspects of sample collection and handling need to be carefully considered and described in detail before starting the study. All these should be part of the study (or trial) protocol, to be described next.

5. Clinical trials and protocols

The previous sections described various aspects of studies in a clinical environment. Owing to the complexities and interrelation of the different aspects, establishing a well-defined protocol for a clinical trial is probably even more important than in other fields of science. Here, both bureaucratic and scientific aspects need to be studied and accommodated. The methodology of clinical research is well established, and many possible sources of bias are recognized [17], which need to be eliminated. In this section, the most important terminology and basic concepts related to clinical trials will be explained.

Clinical trials comprise research that is designed and evaluated to provide reliable information for preventing, detecting, or treating certain diseases or for improving quality of life of patients. Common types of clinical trials are listed in Table 3.

Table 3
Different types of clinical trials

Prevention trials	The aim is to identify interventions that can prevent the particular disease.
Early detection trials	The aim is to identify the methods that can reveal or recognize the particular disease early in its development.
Treatment trials	The aim is to identify interventions that are effective in inverting, stopping, or slowing down the progression of the selected disease.
Quality-of-life trials	The aim is to identify strategies or interventions that improve the quality of life (QoL) of patients during and/or after treatment of the particular disease.
Symptom management trials	The aim is to identify the interventions that ease or prevent the symptoms of the particular disease and/or its treatment.

The objective of a clinical trial is to determine the effectiveness of a planned intervention in achieving its stated goals. Three phases of clinical trials are distinguished; each represents a certain level of knowledge on the question to be answered. In *Phase I* trials a given intervention is tested in human beings for the first time. The aim of these studies is to find the optimum dose, the method of administration (e.g., intravenous injection), and/or possible side effects of the intervention. The aim of *Phase II* trials is to determine the efficacy against the disease specified in the protocol. The objective of *Phase III* trials is to compare the new treatment against the gold standard treatment(s) on a larger number of patients. Most clinical trials require a major effort. Often before starting a full-scale project, it is worthwhile performing a *pilot study* to assess feasibility, potential error sources, and various practical aspects of the whole study.

It is important to emphasize that the patients participating in clinical trials may have personal benefits, but there is no guarantee for a therapeutic advantage. Moreover, the chance for a negative outcome can never be completely ruled out. When the treatment of patients is based on the results of well-designed large-scale clinical trials, *evidence-based medicine* is practiced.

A *prospective trial* means preplanned study in which the question to be answered is formulated before the trial is initiated. Trials should be conducted according to the *study protocols,* which are formal documents specifying all relevant aspects of the planned investigation including which patients are eligible, which treatments are to be evaluated (what questions should be answered), how each patient's response is to be assessed, etc. Protocols should be followed by all investigators to ensure comparability of results. Each protocol should contain an *operation manual* and a *scientific study design*. The operation manual should contain detailed specification of the trial procedure relating to each individual patient (e.g., including the patient selection criteria, treatment schedule and procedure, evaluation of the results, method of data collection, case report form, etc.). Scientific design should include a description of the trial's motivation, its theoretical background, data on which questions were formulated, specific aims to be achieved, and explaining how the results might be utilized. It should also outline the rationale behind the chosen study design, statistical aspects, randomization procedures, and ethical considerations (e.g., the procedures for obtaining informed patient consent prior to commencement of treatment). The main features of a study protocol are shown in Table 4.

Careful documentation is always very important since the report on a trial is usually written a long time, sometimes years, after the first patient is entered. The study protocol therefore serves not only as a basis of decisions made during the trial but also as the source of those decisions. The protocol is also important since clinical investigators may change or new ones join the study. In fact, a protocol should be written clearly in order to help researchers to repeat the trial elsewhere. All documents (including patient data collecting forms) should be kept for a long time to provide the possibility to reanalyze the trial if needed.

Table 4
Main features of a study protocol

(1)	Introduction (background and general aims)
(2)	Specific objectives (questions to be answered, aims to be achieved)
(3)	Patient selection criteria (inclusion and exclusion criteria)
(4)	Treatment schedules (drug formulation, route of administration, amount and frequency of each dose, treatment duration, possible side effects, and their treatment, etc.)
(5)	Methods of patient evaluation (assessment of the treatment, criteria for response, side effects checklist, followup. It must include all intervention, e.g., the frequency and amount of blood samples taken)
(6)	Trial design (choice of control group, procedures for avoiding bias, criteria for interim analysis, etc.)
(7)	Registration and randomization procedures (method of registering a patient to the trial, e.g., telephone, fax, e-mail. Method of randomization, e.g., randomization table, balanced randomization, etc.)
(8)	Informed consent (according to legal requirements)
(9)	The required size of study (patient number per group to be able to detect prespecified differences between groups)
(10)	Monitoring trial progress (usually carried out by independent monitors)
(11)	Case report forms (CRF) and data handling (codes to preserve patient anonymity, etc.)
(12)	Protocol deviations (dose modifications, checks on patient compliance, patient withdrawal, etc.)
(13)	Plans for statistical analysis (statistical test(s) to be used, level of significance, statistical power, etc.)
(14)	Administrative responsibilities (who should file the CRF, how long the documentation should be kept, etc.)
(15)	Funding (who is the sponsor, what kind of research grant or financial support is to be used, etc.)
(16)	Reporting (to sponsor, publications)
(17)	Summary of protocol (the general outline and the flow chart of the study)

The importance of *study design* cannot be overemphasized since subsequent analysis is unable to compensate for major design errors. The choice for a given trial depends on many aspects of the study including the question to be answered, the seriousness of the disease to be treated, the type of treatment to be given, the time course of response to be measured, the endpoint to be evaluated, etc. The term “design” encompasses all the structural aspects of a trial. An important aim of trial methodology is to obtain a bias-free meaningful result by using the least possible resources.

An essential part of any clinical trial is establishing two or more groups of patients (or healthy individuals), which (or responses in the respective groups) are compared. In the simplest case, two groups are compared in some respect. One is the *treatment group;* the other is the *control group.* One group of patients is treated by the drug (or treatment method) under evaluation, and response of these patients is compared to that of the untreated control or control group receiving the standard treatment of the

time. The efficiency of the treatment is determined by some preestablished criterion (such as how many subjects are completely cured after a certain time). Success of this comparison rests on correct selection of the treatment and control groups: Both should be identical regarding a number of parameters such as distribution of sex, age, physical condition, degree of illness, etc., and (in case of placebo trial) should be treated in exactly the same manner but administering the drug in question. Evaluation of the result (e.g., "complete response") should be objective, unbiased by, e.g., preconception of the doctor believing in the treatment.

To provide unbiased results, a number of concepts have been established in the medical community. The concept of *random allocation* was developed about 70 years ago [18]. This means that patients will be randomly assigned to treatment and control groups. Randomization is expected to produce groups that are comparable on all important characteristics, so there is no significant difference between the two groups. This protects against preconception, systematic arrangement, or accidental bias, which can distort the groups. Randomization does not automatically guarantee balance in every aspect (due to the groups having a restricted size, statistical fluctuations may be significant), so the investigator should check whether a satisfactory balance has emerged. Random allocation has the further advantage that it allows using standard statistical methods (such as significance tests) for data evaluation.

To most correctly evaluate results, it is important that both groups should get exactly the same treatment (but the drug in question). To obtain this condition, patients in the control group should get *placebo* (if possible), which is identical in all respect to the active drug except that the active ingredient is absent. This is important, as it is known that a treated patient's attitude will change since something is being done. (Note that many patients could be treated effectively by placebo.) It is also important that the study should be *double blind,* meaning that neither doctor nor patient is informed if a given patient belongs to the treatment or control group. This prevents biased evaluation of the results. If the result of a trial is subjective (e.g., pain relief), double-blind treatment is especially significant.

Although above simple cases have been discussed, in real life there are various complicating factors. Parameters (variables) in clinical studies may have only two categories (e.g., male/female), or several (such as mild, moderate, severe), or could be objectively measurable (such as age or cholesterol level) or not measurable (such as pain). In some studies there are only two groups (treatment and control), in other cases several treatment groups are compared. Sometimes *restricted randomization* is carried out if investigators want to ensure that the numbers of patients allocated to each treatment or important subgroups of patients are approximately equal in number. The method of *random permuted blocks* is often used when there are more than two groups of patients; *stratification* is used to protect against random allocation producing imbalance between groups regarding important variables such as stage of the disease or age. Selecting the control group is not always easy,

as it is unethical not to treat patients. In such a case the control group may receive the standard treatment against the illness.

Those discussed above relate to conventional treatment trials, but (as listed in Table 3) there are other types of clinical trials as well. The main concepts are analogous: A well-selected control group is essential, and unbiased statistical data evaluation should be provided. For example, in the case of early detection trials usually the level of a potential biomarker is determined in a group of patients and it is compared to that of the control group. In such a case it is important that the only difference between the two groups should be that of the presence of a given illness. False results may easily be obtained (and are often even reported) if, e.g., melanoma patients are compared to a group of healthy individuals. In such a case the found biomarker may be representative of very ill persons and not that of melanoma.

In case of *observational studies* it is essential that the data obtained should be as representative of the population as possible. If the sample is not representative enough, the results will be unreliable and of dubious value. It is often useful to sample several subgroups (e.g., by age, sex, etc.). Defining suitable control groups are often difficult in such a case as well. A *cohort study* is a prospective, observational study that follows a group (cohort) over a period of time and investigates the effect of a treatment or a risk factor. *Historical (retrospective) controls* are used sometimes, but these are very prone to errors. In such a case data are collected initially on patients receiving a different treatment considered as the control group. Possibly the worst case is comparison with published results because a publication is strongly biased toward positive results.

Information regarding the relative value of treatments is often accumulated slowly. Hence, *interim analysis* is important. This means that results of a clinical trial are analyzed while the study is still in progress. When the results become statistically relevant (which may happen much earlier than predicted in the study protocol), the new treatment may prove beneficial or undesirable. In such a case ethical considerations require the trial to be concluded, in order that all patients should receive the more efficient treatment.

In clinical routine it is often observed that some patients do not stick to their treatment, so *patient compliance* is a critical aspect of clinical trials. Noncompliance may be reduced by careful explanation regarding the treatment schedule and the trial's objectives, by handing overwritten instructions to the patients, and by regular checkups (e.g., counting the number of remaining tablets, by blood analysis through measuring plasma level of drugs, etc.). It is important to differentiate between lack of cooperation and misunderstanding. The latter can and should be avoided by better planning.

All trials require precise definition which patients can or cannot participate, which is defined by *patients' eligibility criteria* and often controlled by an *eligibility checklist*. In each trial the number of ineligible patients and the reasons for ineligibility should always be reported. Even when using careful planning, a small

proportion of ineligible patients are often included by mistake, which is a case of *protocol deviation*. If the proportion of ineligible patients goes above a certain threshold value (e.g., 10%), it ruins the whole trial. *Protocol violation* is a more serious event which greatly influences the study results and which is often caused by or could have been prevented by the investigator. *Patient withdrawal* from treatment, for whatever reason, should not preclude a patient from subsequent evaluation; in fact, they should be followed for reporting of morbidity and mortality. Patient withdrawal could occur in case of serious noncompliance of a patient, because of patient refusal for further participation, and due to clinical judgment (i.e., due to severe side effect or disease progression). If such evidence is lacking, it should be regarded as protocol violation.

Regardless of the type of statistical design *sample size* is of high importance in a clinical trial. The required sample size often depends on trial design. In small-scale trials or pilot studies approximately 10, in medium-size trials typically 50–100, and in large-scale trials several hundred subjects participate in each patient group. When the required number of patients is too large, several institutions may participate, so *multicenter* or *multinational* trials are organized. Slow *patient recruitment* is a serious concern in all trials.

6. Administrative procedures

Biomedical research and, in particular, clinical trials are very complex and well regulated. Administrative procedures have two objectives: to provide help in obtaining correct and unambiguous results, and to prove that experiments were performed as described, according to high scientific and ethical standards, and therefore to increase confidence in the conclusions obtained. Administrative procedures also involve various controls (inside and outside the institute performing the study).

Quality assurance/quality control should be a standard part of research practice, and this includes sample handling. Since it is inevitable that several individuals and often several laboratories will collaborate, strict adherence to SOP is essential. Inappropriate handling of the samples could endanger the result of the whole project. It also helps the technical personnel to avoid misunderstanding or misinterpretation.

Labeling, retrieval, and storage of samples should be regulated in detail, just as sample flow. Equipment should be maintained and calibrated regularly; the personnel operating them should be trained properly. Note that these not just have to be performed adequately, but have to be documented in detail. Clinical work should typically be performed according to GCP: an international ethical and scientific quality standard. Its main objective is to ensure that the data and reported results are credible and accurate, and that the rights of trial subjects are adequately protected.

References

1. Sackett, D.L., Straus, S.E., Richardson, W.S., Rosenberg, W. and Haynes, R.B., *Evidence-based Medicine: How to Practice and Teach EBM* (2nd edition). Churchill Livingstone, Edinburgh, 2000.
2. Goodman, K.W., *Ethics and Evidence-Based Medicine*. Cambridge University Press, Cambridge, 2003.
3. Pocock, S.J., *Clinical Trials. A Practical Approach*. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1983.
4. World Medical Association, The Declaration of Helsinki. www.wma.net. Copyright World Medical Association. All Rights Reserved, 2006.
5. Lötjönen, S., Medical research in clinical emergency settings in Europe. *J. Med. Ethics*, **28**, 183–187 (2002).
6. Fulford, K.W. and Howse, K., Ethics of research with psychiatric patients: principles, problems and the primary responsibilities of researchers. *J. Med. Ethics*, **19**, 85–91 (1993).
7. Lo, B., Zettler, P., Cedars, M.I., Gates, E., Kriegstein, A.R., Oberman, M., Reijo, P.R., Wagner, R.M., Wuerth, M.T., Wolf, L.E. and Yamamoto, K.R., A new era in the ethics of human embryonic stem cell research. *Stem Cells*, **23**, 1454–1459 (2005).
8. McIntosh, N., Bates, P., Brykczynska, G., Dunstan, G., Goldman, A., Harvey, D., Larcher, V., McCrae, D., McKinnon, A., Patton, M., Saunders, J. and Shelley, P., Guidelines for the ethical conduct of medical research involving children. Royal College of Paediatrics and Child Health: Ethics Advisory Committee. *Arch. Dis. Child.*, **82**, 177–182 (2000).
9. Obermeyer, C.M., Ethical guidelines for HIV research: a contextual implementation process. *J. Int. Bioethique*, **15**, 134–135 (2004).
10. Altman, D.G., Statistics and ethics in medical research. In: *Statistics in Practice*. British Medical Associstion, Tavistock Square, London, 1994, pp. 1–24.
11. Baeyens, A.J., Hakimian, R., Aamodt, R. and Spatz, A., The use of human biological samples in research: a comparison of the laws in the United States and Europe. *Biosci. Law Rev.*, **5(5)**, 155–160 (2002).
12. Hill, H.R. Jr., Gaunce, J.A. and Whithead, P., Chemical safety levels (CSLs): a proposal for chemical safety practices in microbiological and biomedical laboratories. Office of Health and Safety Centers for Disease Control and Prevention, 1999. http://www.cdc.gov/od/ohs/CSL%20article.htm.
13. Centers for Disease Control and Prevention/National Institutes of Health, *Biosafety in Microbiological and Biomedical Laboratories* (4th edition). In: McKinney, R.W. and Richmond, J.Y. (Eds.), HHS Publication No. (CDC) 93-8395; U.S. Government Printing Office, Washington DC, 1999.
14. Georgia Tech Radiological Laboratory Classification, 1999. http://www.ors.gatech.edu/labclass.htm.
15. Auray-Blais, C. and Patenaude, J., A biobank management model applicable to biomedical research. *BMC Medical Ethics*, **7**, E4 (2006). http://www.biomedcentral.com/1472-6939/7/4.
16. Holland, N.T., Smith, M.T., Eskenazi, B. and Bastaki, M., Biological sample collection and processing for molecular epidemiological studies. *Mutat. Res.*, **543**, 217–234 (2003).
17. Sackett, D.L., Bias in analytic research. *J. Chronic Dis.*, **32**, 51–63 (1979).
18. Fischer, R.A., *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Part II
Tools of the Trade

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 4

Biomedical sampling

GYÖRGY VAS[a,*], KORNÉL NAGY[b], and KÁROLY VÉKEY[b]

[a]*Cordis Corporation, Analytical Technologies, Pharmaceutical & Package Development, Welsh & McKean Roads, P.O. Box 776, Spring House, PA 19477-0776, USA*
[b]*Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary*

1. Sampling

Mass spectrometry is a highly selective analytical technique, which can provide reliable information about the molecular composition of a biological sample. The

*Corresponding author. Tel.: +1-215-628-5841; Fax: +1-215-540-4601.
E-mail: gvas@crdus.jnj.com (G. Vas).

latest improvements in mass spectrometry can cope with more and more challenging bioanalytical problems, thereby dramatically increasing the utility of mass spectrometric applications. Nevertheless, the success does depend not only on the actually applied analytical technique but also on other steps of the whole analytical protocol, from sampling to data evaluation. Among the most crucial and time-consuming steps are sampling and sample preparation, which may comprise over 80% of total analysis time [1]. The quality of these steps is a key factor in determining the success of analysis [2]. The majority of analytical processes consist of four primary steps: sampling, sample preparation, analysis, and data evaluation, as discussed briefly in Chapter 2.

Even for a highly efficient analytical method, such as mass spectrometry, the sampling is of critical importance. It ensures that the analyzed sample is representative and reflects the condition of the biological object. Clinical and biomedical aspects of sampling have been discussed in Chapter 3. Improper sample handling may introduce drastic errors in the process, making the analysis useless or misleading. Sample pretreatment is generally also required to avoid interferences and improve the performance of the analytical protocol, especially when complex matrices such as blood or tissue are studied.

In the sampling step, the material to be analyzed is collected, for example, blood is drawn into heparin-containing vacuum tubes. The objective of any sampling strategy is to obtain a homogenous and representative sample that is a prerequisite for obtaining meaningful results. Homogeneity is generally not a critical issue for gaseous and fluid samples, but it really matters for solid samples.

Sample collection includes a decision on where and when to get samples so that it properly represents the biological objects being analyzed. For instance, what time span must be spent between administration of a drug and drawing of the blood sample, or shall the sampling happen before or after alimentation, etc. Sampling also includes the selection of a method that obtains samples in the appropriate amounts (e.g., is the blood needed only for dried blood spots or for methods requiring several milliliters?). For trace analysis, sampling is a very critical issue. If not properly planned and performed by using appropriate sampling tools with care and expertise, the total error caused by sampling can increase from the usually expected few percentages to several orders of magnitude.

In bioanalytical studies it is always necessary to collect appropriate blank samples. These blank samples are the matrices that have no measurable amount of the analyte of interest. The ideal blank will be collected from the same source as the samples, but will be free of analyte. All the conditions related to the collection of the blank sample—storage, pretreatment, extraction, concentration, and analysis—have to be the same as for the actual samples. Such an ideal blank sample is not always available, so often a compromise is necessary. For example, when an

endogenous analyte is studied (which is always present in the given biological matrix), a well-defined standard sample should be used. This should contain the studied analyte in a well-defined (not variable) amount.

In most cases, standards are also needed for analysis, and these should be added to the sample as soon as possible, preferably at the time of sampling. The standard is often an isotope-labeled compound. A big advantage of mass spectrometry-based methods is that they can detect stable isotope labels and are not radioactive; therefore, they do not pose a health hazard and can be freely used. Stable isotope labeling is also called "isotope labeled affinity tags," especially in the field of proteomics.

Biological matrices may be liquid or solid, and contain a variety of different molecules and particles. Various biological samples are used for chemical analysis. Most commonly blood and urine are used, but saliva, milk, sweat, feces, and various tissues (liver, kidney, brain, etc.) are also studied. The properties of these matrices for sampling and sample preparation are described in books [3,4] and reviews [5]; here only a brief description is given.

Blood is a fluid connective tissue in which the blood cells are suspended in a fluid matrix called plasma. The blood transports oxygen, the products of digestion, hormones, enzymes, and many other chemical substances including the waste products of metabolism. If fresh blood is placed in a centrifuge tube and rotated rapidly, it separates into its three basic components. The upper layer, about 55% of whole blood, is a light-amber fluid called plasma. The remaining 45% is a mixture containing the formed elements, mostly red blood cells (lower layer) and approximately 1% white blood cells (middle layer, also called buffy coat). Serum is also studied; this is a yellowish liquid obtained after clotting. Clotting is an important property of blood, and is usually undesirable for analytical work. Most often clotting is inhibited by anticoagulants like heparin (a mucopolysaccharide) or ethylene diamine tetraacetic acid (EDTA). In most cases EDTA is considered more appropriate, as heparin may interact with some analytes changing sample composition. However, EDTA may not be used for studying metals or organometallics.

Molecular composition of blood is most often analyzed by studying plasma or serum. For a long period these should be stored at $-80^{\circ}C$; for a few weeks storage at $-20^{\circ}C$ is usually acceptable. Note that storage plasma is a critical issue; the type of analysis to be performed may dictate different conditions. For example, and arguably, RNA profiling may give meaningful results only when fresh samples are used. Blood contains both small molecules and a large amount of proteins. After obtaining the plasma (or serum), the next step is usually separation of macromolecules (mostly proteins) and small molecules, e.g., by protein precipitation (see the following text). Analysis of these two molecular fractions requires different methodologies. Note that small molecules may bind to proteins, and this possibility has to be taken into account.

Urine is also a very commonly studied biological matrix. It is much less complex than blood, and contains only a small amount of macromolecules. It has a high salt content and both organic and inorganic constituents. Its main components include NaCl (10 g/l), K (1.5 g/l), sulfate (0.8 g/l), phosphate (0.8 g/l), Ca and Mg (0.15 g/l), urea (20 g/l), creatinine (1 g/l), and uric acid (0.5 g/l). Urine should be protected from bacterial degradation, which is mostly accomplished by freezing the samples until analysis.

Other body fluids such as *saliva* or *milk* are also sometimes analyzed. Saliva contains approximately 0.3% protein (mostly enzymes), 0.3% mucin, and some salts. It can often be analyzed directly, without sample preparation or extraction. Human milk contains mainly proteins (3%), lipids (3%), and carbohydrates (mainly lactose, 6.8%). Lipids are suspended in the form of droplets, so homogeneity of the studied sample must be ensured. From the third week from the start of lactation, composition of human milk is quite constant, but that of the initially secreted colostrum is significantly different. Milk samples are commonly used for the trace analysis of pesticides, heavy metals, antibiotics, and some drugs.

Various tissues are also analyzed, although much less frequently than blood or urine. This is partly because these are far more difficult to obtain (especially human samples), partly because sample preparation is much more challenging. Obtaining a representative sample is important (and often difficult) and requires homogenization. Tissues often have high fat content, which also complicates sample preparation. Combination of various sample preparation and extraction methods is often required, and precise protocols are indispensable. These protocols depend significantly on the type of tissue studied.

Hair is a special case, and is an attractive target for chemical analysis: It is noninvasive to collect, requires relatively simple sample preparation protocols, and provides a historical record of exposure to various chemicals and drugs. Hair is usually collected from the area at the back of the head, and to provide a representative sample, at least 200 mg should be collected. Hair analysis is being often used, especially in forensic applications.

2. Sample preparation

Biological matrices are mostly highly complex aqueous (except the fat or lipid tissues) mixtures, usually having high protein and salt content. Most are not adequate for direct chromatographic or mass spectrometric analysis but require sample preparation including cleanup, extraction, and/or derivatization. The main objective of sample preparation is to convert a real biological matrix into a form suitable for analysis by the desired analytical technique [6]. The theory and implementation of sample cleanup and extraction are based on similar physicochemical principles as chromatographic methods, so the present chapter and the

following one on "separation methods" are strongly related and complement each other.

The first aim of sample preparation is the removal of potential interferences. For example, inorganic salts need to be removed from the sample before analysis by mass spectrometry, as they suppress ionization of organic analytes and reduce sensitivity. For analyzing small molecules, such as drugs, fatty acids, and sugar phosphates, proteins and glycoproteins need to be removed.

The second aim of sample preparation is to increase the concentration of analytes to achieve adequate signal intensities. Enrichment is usually performed by extraction methods, such as liquid–liquid (LLE) or solid-phase (SPE) extraction. The simplest form of enrichment is drying the sample and reconstituting it in a smaller solvent volume. Extraction is often combined with a change of solvent (e.g., an aqueous sample, after extraction, will be reconstituted in an organic solvent).

There may be several other reasons for using sample preparation. Analytes are often changed chemically (i.e., are derivatized) to become better suited for analysis or detection. Furthermore, sample preparation should be robust to provide reproducible samples independent of variations in the sample matrix.

A large variety of different sample preparation methods are available. In most cases one technique is rarely sufficient; usually several are used in combination. Here only the most commonly used ones will be discussed. Sample preparation usually starts with separating the sample into various fractions. First, particles (such as cells, fibers, etc.) are separated, using centrifugation and/or filtration to provide a homogenous solution. In the next step, small molecules are separated from macromolecules. Commonly this is done by protein precipitation, ultrafiltration, or dialysis. Salt content may also be reduced by dialysis. Following these preliminary steps, proteins and other macromolecules are usually digested. Derivatization is also commonly used to convert the sample into a form more amenable for analysis. Other simple sample preparation methods include removal of the solvent, often done using vacuum evaporation or lyophilization. Sample preparation also includes various extraction methods that will be discussed in the next section.

2.1. Centrifugation

Centrifugation is a very common technique to separate solid particles dispersed in liquid medium, e.g., blood cells and plasma. The liquid sample is placed in a special vial or holder, which is rotated very fast. Sample components are separated due to the centrifugal force, based on their density difference. Centrifugation is commonly used in combination with a variety of sample preparation techniques. Centrifugation can also be used to separate emulsions (such as milk) and immiscible solvents (e.g., in combination with LLE).

Laboratory centrifuges usually work with 20–40 cm diameter rotors holding 10–100 sample vials or two to four microtiter plates. Efficiency depends on rotational velocity; typical laboratory centrifuges work with 100–20,000 rotations per minute, allowing separations in a few minutes time.

Ultracentrifuges are different specialized equipment, working at higher velocities. These are mostly used to separate macromolecules based on molecular mass. *Vacuum centrifuges* are also common; their purpose is evaporating solvents—centrifugation is used to help in keeping the solution at the bottom of the vial.

2.2. Filtration

Filtration is another common method to separate solid particles dispersed in a liquid. The simplest case is filter paper, although mostly polymer-based filter membranes are used. The most important quality of filters is the size (diameter) of the particles filtered out, corresponding to the pore size of the filter. In the case of ultrafiltration, macromolecules can be filtered out—this is discussed in Section 2.4. Filters come in various sizes, depending on the quantity of sample to be analyzed. They may be incorporated into the tips of pipettes, which make it easy to remove small solid particles from a solution, for example, before injecting it onto a chromatographic column.

2.3. Protein precipitation

Probably the simplest way to separate proteins from small molecules is protein precipitation. It is needed for studying low-molecular-weight compounds (MW below 2–5 kDa), as the presence of macromolecules typically deteriorates analytical performance. In chromatography they lift the baseline, cause noise, and may even ruin chromatographic columns. In MS, they deteriorate ionization and may block the ion source.

Precipitation is performed by adding organic solvent (acetonitrile, methanol), inorganic acid (perchloric acid), or salt (zinc sulfate) to the sample. After mixing, the proteins aggregate, and after centrifugation, they form a pellet at the bottom of the sample vial. This pellet can be easily removed from the remaining liquid, making separation of proteins and small molecules easy and quick. However, the disadvantage of protein precipitation is that various proteins precipitate under different conditions, so protein removal is not perfect. The precipitated proteins may bind various small molecules and remove them from the solution. This may influence quantitation, which has to be taken into account.

2.4. Ultrafiltration

Ultrafiltration is another common way of separating small and large molecules (e.g., sodium vs. albumin). The liquid sample is dispensed into an ultrafiltration

tube. The bottom of this tube is a membrane, usually made of regenerated cellulose. After the tube is put into a centrifuge, the centrifugal force pushes the solvent and the small molecules through the membrane. The macromolecules (in the present example albumin) are retained on the membrane. After ultrafiltration they can be washed from the membrane. Both the macromolecules and the solution containing the small molecules may be used for further analysis. Care must be taken with ultrafiltration because the membrane material might bind some analytes. This must be checked before applying ultrafiltration.

The most important characteristic of an ultrafiltration tube is its MW cutoff value, usually expressed in kilodaltons. A tube with 10 kDa cutoff retains the molecules with molecular mass higher than approximately 10 kDa. This cutoff value is not very accurate; in the present example, a small fraction of compounds with 5–10 kDa may be retained, while some of 15–20 kDa may pass through the filter. There are various filters, with cutoff values in the range 3–100 kDa.

2.5. Dialysis and electrodialysis

Like the other methods, the main purpose of dialysis and electrodialysis is the separation of small and large molecules; it is often used for desalting purposes. These are based on the phenomenon that certain compounds can diffuse through a semipermeable membrane, while others cannot. This differentiation is mainly based on molecular size. The principle of dialysis is, in fact, quite similar to ultrafiltration; the driving force is not only gravity (assisted by centrifugation) but also osmotic pressure.

In a typical dialysis experiment, a membrane separates two liquid phases, one of which is the sample (see Fig. 1) and the other is a clean washing liquid. The membrane is permeable for small molecules but retains large ones. Small molecules can therefore diffuse through the membrane into the other liquid phase. This diffusion process goes on until equilibration is reached. In practice, a large amount of washing liquid and a small amount of sample solution are used, so

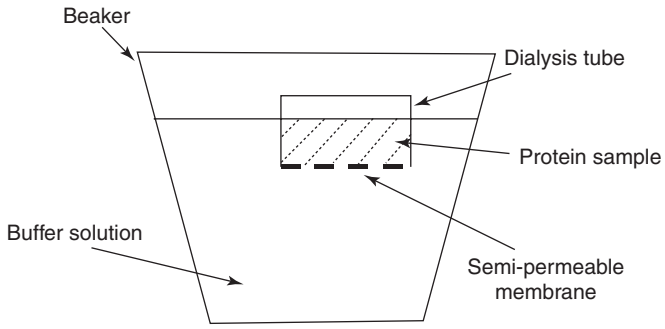


Fig. 1. Schematic diagram of dialysis.

concentration of the (unwanted) small molecules in the sample is significantly reduced.

A typical example of dialysis is desalting of proteins. About 500 μl protein solution is put into a dialysis tube, which is immersed in 500 ml buffer solution (see Fig. 1). The salts diffuse from the sample into the buffer solution, while the buffer (since diffusion can occur in the opposite direction too) diffuses into the sample and maintains the pH. This process not only desalts the protein but also can be used to exchange the buffer.

Various dialysis membranes are used; those of 10–15 kDa molecular weight cutoff are most common. Dialysis may also be used to clean small molecules from unwanted macromolecules. It is easy to miniaturize; sample volumes as small as a microliter can be used (e.g., one drop of sample placed onto a small filter floating on pure water). Various parameters may influence the efficiency of dialysis, such as the type of the membrane, temperature, the volume of the sample, extractant volume, etc. Efficiency of the dialysis may significantly be decreased if the analytes bind to the membrane either by electrostatic or by hydrophobic interaction. The use of a low-concentration surfactant may decrease this effect.

In electrodialysis, diffusion of charged compounds through the membrane is aided by an electric potential difference. Naturally, this potential difference acts only on charged species, so in electrodialysis the charge on the analyte has key importance.

2.6. Digestion

Digestion is among the most commonly used preparative steps for studying macromolecules. Most analytical techniques yield only limited information on macromolecules, so breaking them into smaller fragments is often necessary. These small fragments are then studied by "conventional" methods, such as chromatography and mass spectrometry. Tryptic digestion of proteins is probably most common, but many other digestion procedures are also used. These break up proteins and other macromolecules, and are invaluable tools for studying the structure and function of macromolecules. Many enzymes have special selectivity. These may be used not only to break up macromolecules but also to study special structural features.

In the case of trypsine, proteins are cleaved at basic sites (lysine and arginine). Using this enzyme is very popular as it has high specificity, it is easy to use and the result is well reproducible. Tryptic digestion is commonly accompanied with other chemical treatment, such as unfolding the protein and cleavage of the sulfur bridges. Various experimental protocols can be found in the literature for using trypsine; most of them are quite effective. The obtained digest, which is a complex mixture, is then analyzed by MALDI–TOF or HPLC–MS–MS experiments. The

results are evaluated by using bioinformatics and identifying the protein; it may also yield other structural information (e.g., on posttranslational modifications). Tryptic digestion is the most common tool in proteomics, and will be discussed in detail later in the book.

2.7. Chemical derivatization

Chemical derivatization is often applied when the properties of target analytes are not compatible with the analytical procedure, when detection is not sufficiently sensitive, to label a given analyte, or to increase selectivity. While derivatization may increase analytical performance, it is time-consuming and often labor intensive. The most critical issue is, however, that derivatization changes sample composition, may not be quantitative, may lead to by-products, and may lead to the loss of the analyte (especially if done at a very small scale). These problems need to be carefully considered (increasing the time and cost of method development) and may compromise reproducibility and robustness of the analytical procedure. For these reasons although derivatization is often indispensable, if possible, it is often avoided.

A popular application of chemical derivatization in the biomedical field is amino acid and carnitine analysis using tandem mass spectrometry. Free amino acids in dried blood spots are butylated, which increases both selectivity and ionization efficiency. This way the time-consuming chromatography–mass spectrometry analysis can be substituted by very fast tandem mass spectrometry (requiring approximately 1 min). This opens up the possibility of population-wide screening for inherited metabolic disorders, which is applied in many countries. Another typical example is volatilization of polar compounds to make them amenable for gas chromatographic analysis. Many such procedures are known, e.g., methylation or silylation. One such application in the biomedical field is methylation of very long chain fatty acids in plasma to screen peroxisomal disorders. Although time-consuming, this makes very long chain fatty acid analysis possible using gas chromatography (GC).

2.8. Lyophilization

The purpose of lyophilization is to evaporate the solvent under very gentle conditions. First, the sample is frozen and then the solvent is sublimed away using vacuum. The remaining solid sample forms a very light structure with high surface area. In practice, it means that it is easy to collect and/or re-solvate the sample, even if the quantity is very small. Note that conventional, complete solvent evaporation (using heating and often also using vacuum evaporation) often results in a very compact solid material, partly sticking to the wall of the vial, which is difficult to collect or re-solvate. It is a particularly important aspect of handling

small amounts of material—complete solvent evaporation often results in a significant loss of sample. Lyophilization is therefore an efficient method for concentrating and handling small amounts of samples.

3. Extraction techniques

Extraction methods form an integral part of sample preparation. They are grouped together, as their main purpose is to increase the concentration of the analyte and they are strongly connected to chromatography. In fact, these can be regarded as a very simplified form of chromatography. The oldest, but efficient and still used version is LLE. SPE is probably most common; it has many versions and has become an indispensable tool in the biomedical field. Its further simplified form is "ZipTip" preparation. The principle of solid-phase microextraction (SPME) is more similar to LLE than to SPE; it is commonly used sample preparation method for analyzing volatile compounds.

3.1. Liquid–liquid extraction (LLE)

LLE is also called solvent extraction. It is used for both sample cleanup and concentration of the analyte. LLE is based on the phenomenon that a compound will distribute between two nonmiscible liquid phases. The equilibrium is strongly determined by the physicochemical parameters of the two liquids and can be advantageously used to concentrate some while dilute other components of the sample.

In a typical LLE experiment, an aqueous sample is mixed with an apolar, nonmiscible solvent (like *n*-hexane). This may be performed in a simple vial or in a special separatory funnel. After combining the two liquids the vial (or separatory funnel) is shaken vigorously to aid mixing of the two liquids. Once the shaking is over liquid droplets are formed (as the two solvents are not miscible); these are allowed to coalesce (possibly aided by centrifugation) and the two bulk phases are separated from each other. In this experiment, polar and apolar analytes are separated to a large degree: The polar ones are concentrated in the aqueous phase and the apolar ones in the organic phase. The phase in which the analyte is dissolved is collected (e.g., by pipetting that phase into another vial).

The success of LLE is mainly determined by the choice of the solvents, the use of additives (e.g., by adjusting the pH, which strongly determines solubility in water), and the type of impurities which needed to be separated. When the interferences are similar to the analytes, LLE cannot be applied successfully. Important advantages of LLE are its large sample capacity and negligible memory effects. Disadvantages are that it uses large amounts of organic solvents (which are expensive and toxic), is very labor intensive, and is difficult to automate.

A common application of LLE is extraction of drugs from aqueous matrices using volatile organic solvents (e.g., dichloromethane), which is easy to concentrate by evaporation of the solvent and can be directly injected into a GC or GC/MS instrument for analysis. A very efficient method is the so-called back-extraction, which can be used for various drugs that can be ionized in a certain pH range. First, the pH of the aqueous phase is adjusted so that the analyte will be in neutral form when it migrates into the organic phase, where it is well soluble. This may be aided by salting the aqueous phase. In the next step (back-extraction) the organic phase is mixed with an aqueous phase, in which the pH is adjusted so that the analyte will be ionized. This strongly favors solubility in the aqueous phase, so the analyte will be "back-extracted." This method is able to separate the analyte from both apolar and many polar impurities, resulting in an efficient and easy sample cleanup procedure.

The conventional LLE technique requires large sample size (10–100 ml at least), which is usually not available in the biomedical field (a possible exception is urine analysis). To overcome this difficulty (and also that of the cost of large volumes of organic solvents), *liquid-phase microextraction* techniques have been developed—these rapidly gain popularity. Easiest and the most common among these is the "single liquid drop" technique, which utilizes a microliter or smaller size droplet of organic solvent suspended in a large volume of aqueous phase [7], as shown in Fig. 2. Analyte distribution and equilibration between the two phases

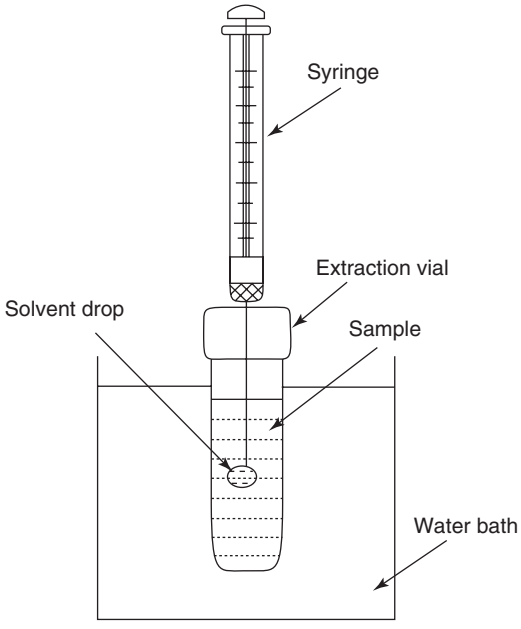


Fig. 2. Schematic diagram of the "single liquid drop" LLE technique.

occur similarly to classical LLE. When the extraction is terminated, the drop can be withdrawn into the syringe and injected directly into an analyzer system (usually GC or HPLC) [7].

Unsupported liquid membrane techniques with three phases involve an aqueous sample phase separated from another aqueous phase (called as receiver phase) by a layer of organic solvent (e.g., octane). Analyte components first diffuse from the sample into the organic liquid membrane and then back-extract out of the membrane into the receiving phase. At the same time, interferences do not diffuse into the organic membrane layer but stay in the original sample phase.

Supported liquid membrane extraction techniques employ either two or three phases, with simultaneous forward- and back-extraction in the latter configuration. The aqueous sample phase is separated from the bulk organic or an aqueous receiver phase by a porous polymer membrane, in the form of either a flat sheet or a hollow fiber that has been impregnated with the organic solvent phase. The sample phase is continuously pumped, the receiver phase may be stagnant or pumped, and the organic phase in the membrane pores is stagnant and reusable [8–10].

3.2. Solid-phase extraction (SPE)

The principal goals of SPE are analyte enrichment, purification, and medium exchange (e.g., from aqueous to organic) [11]. SPE is very similar to liquid chromatography and uses the same physicochemical principles, solvents, and stationary (solid) phases. It has become a very successful and widespread method; most biomedical laboratories use it in everyday practice.

A typical SPE experiment includes several SPE cartridges placed onto a vacuum manifold, as shown in Fig. 3. The SPE cartridge is a short column, resembling an open syringe barrel, containing sorbent material (the solid or stationary phase) packed between porous metal or plastic frits. First, the cartridge is treated with a solvent (to wet the surface) and then the sample solution is placed (pipetted or poured) into the open tube. The solvent passes through the column material and drops into the container below. To speed up the process, vacuum is applied to the bottom of the column. Using proper solvents and cartridges, the analytes will be absorbed on the sorbent material, while the impurities will not be retained and pass through the column with the solvent. In the next step the analytes will be eluted using another solvent and collected into Eppendorf tubes. In typical SPE applications, approximately 1 ml sample size is used, the cartridges are washed with a few millimeters of solvent, and elution may require 10–20 min. A typical vacuum manifold accommodates about 10 SPE cartridges (which are easy to manage manually). In order to improve reproducibility and avoid cross-contamination, SPE cartridges are used only once; then they are discarded.

SPE, like other sample preparation procedures, requires careful and accurate execution. As it is very common, various steps of the SPE procedure will be

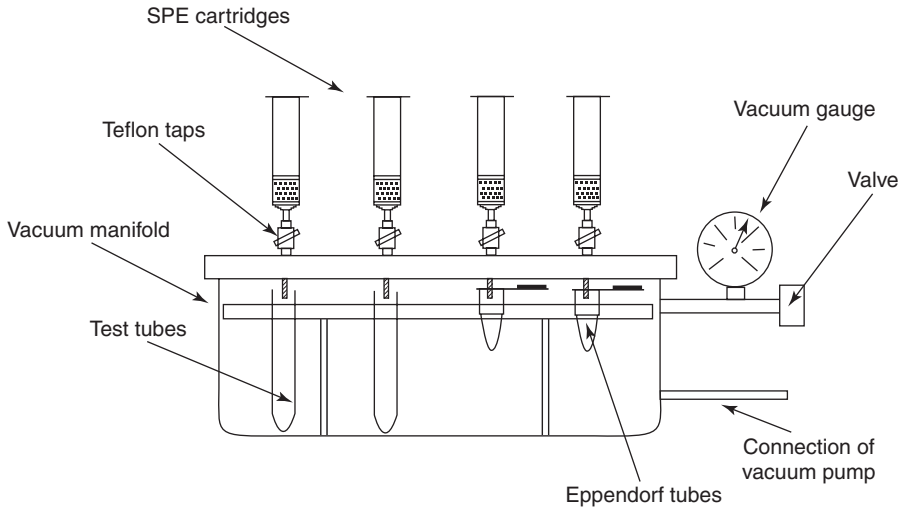


Fig. 3. Schematic setup of an SPE experiment.

described in detail. Viscous samples often need to be diluted before SPE. As a typical example, 200 μl serum is studied, which is first diluted four times with water to 800 μl, and it is studied by reverse-phase (RP) SPE, using a C18 cartridge of 500 mg capacity (or bed volume).

(1) First, the packing material in the SPE cartridge must be conditioned and equilibrated. The role of conditioning is to solvate the functional groups in the sorbent material. Equilibration maximizes the efficiency and reproducibility of retention and also reduces the amount of sorbent impurities washed off at the elution stage. Conditioning is particularly important for processing aqueous samples. Conditioning and equilibration occur at the same time by flushing the packing material with the same solvent (i.e., water) as used for the sample. This and all other washing steps should take place at a controlled flow rate, typically one to two drops per second, using approximately 2 ml of water. Solvent flow can be adjusted by the vacuum pressure applied. After conditioning and equilibration the SPE cartridge is ready for the sample.
(2) The sample is applied to the cartridge, maintaining the flow rate in order to allow efficient binding onto the phase.
(3) After the sample is applied to the SPE cartridge, a washing step is typically included to complete elution of the interferences. This washing step can be performed with either the same or a different (stronger) solvent than used for the sample, using again approximately 2 ml of water. In some cases this washing step is followed by drying, when traces of the washing solvent (typically an aqueous phase) should be completely eliminated.

(4) Finally, the analytes of interest are eluted from the sorbent in a small volume of strong solvent, in the present case 0.5–1 ml of methanol or acetonitrile. The result of this final elution is an effluent that contains the purified analyte. This SPE preparation usually requires about 20 min time; preparing 10 samples in parallel is typical. Note that sometimes SPE can be used in the "opposite" direction, when the analyte to be purified is eluted and the impurities are bound to the cartridge.

There are various practical aspects to consider when performing SPE operation. The high surface tension and high polarity of water often result in a slow and uneven flow through the packing, leading to low analyte recovery. To overcome this problem, addition of a small amount of organic solvent to aqueous phases is suggested, which helps to maintain constant flow even if large sample volumes are used. Viscous samples often exhibit irreproducible results, mainly because they do not allow a stable flow through the cartridge. In such cases dilution (especially with relatively low viscosity solvent) helps to overcome this problem. In the typical case of plasma samples, it is diluted three to five times using water containing 5–10% methanol, which often solves both problems discussed earlier. Accurate adjustment of pH and ion strength is often necessary both to ensure efficient separation and to obtain good reproducibility, particularly when ionic compounds are studied. In some cases proteins and other macromolecules are removed from the biological fluid before SPE, but this is not always necessary. The most typical practical problem with SPE is nonuniform flow of the liquid through the extraction bed. This should be carefully controlled by the operator. Changing the vacuum pressure and degree of dilution and not allowing the extraction bed to dry are needed to avoid this problem.

Development of SPE methods requires a sound knowledge of liquid chromatography. The most important parameters are the SPE cartridge (type of the sorbent) and the type of solvents used. The size of the cartridge (which determines the sample amount) is also important: Too small sorbent size is easily overloaded, while too large sorbent size may bind the analyte, thereby decreasing recovery. Like in chromatography, various additives, especially buffers, may be used to improve performance.

The three most important types of SPE are RP, normal phase (NP), and ion exchange (IE). RP-SPE is best to clean up polar samples from an aqueous phase; NP-SPE is best used for apolar compounds dissolved in an organic matrix, while ionic compounds are best retained on IE-SPE cartridges. Fig. 4 provides a useful guide for selecting the SPE method. Retention mechanism on SPE cartridges is essentially the same as that in chromatography and this is discussed in more detail in the next chapter.

Reverse-phase SPE separations, like RP chromatography, uses a polar (aqueous) sample, an apolar stationary phase, and an organic solvent to elute the analyte. Commonly used stationary phases are alkyl-bonded silicas such as C18, C8, and C4,

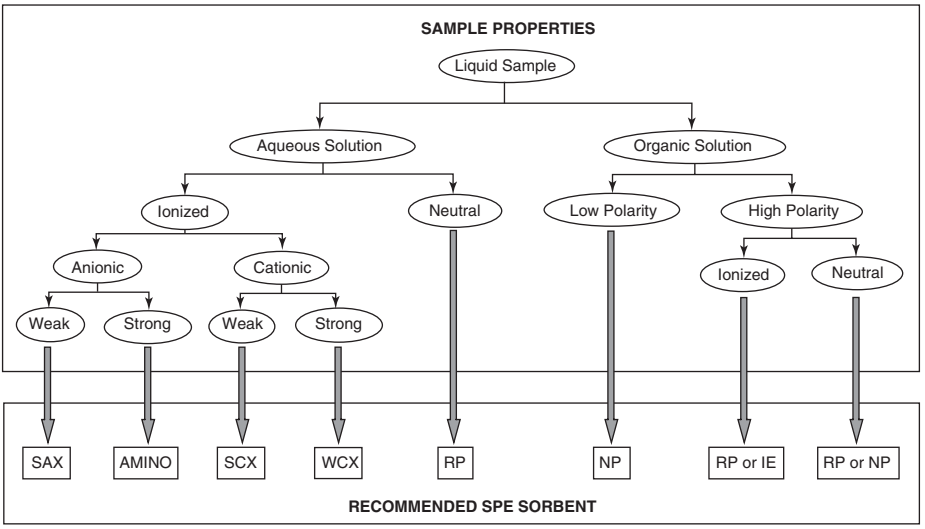


Fig. 4. General guidelines for selecting the type of SPE cartridge to be used. SAX, strong anion exchange; amino, amino column; SCX, strong cation exchange; WCX, weak cation exchange; RP, reverse phase; NP, normal phase; IE, ion exchange.

but occasionally polymers (such as styrene/divinylbenzene [12]) may also be used. Retention of organic analytes is primarily due to the attractive (van der Waals) forces. A typical example of RP-SPE extraction of drugs from plasma was described earlier. Slightly acidic or basic compounds can be also purified using this method, if one adjusts the solvent pH to the value at which the analytes are present in their nondissociated form. A different, somewhat more complex, application of RP SPE cartridges is the extraction of very polar components from aqueous matrices by using ion-pair SPE [13]. The highly polar compounds have poor retention on RP media, but with the help of an ion-pairing reagent (such as triethanolamine) they can also be retained by the apolar reverse phase. First, the ion-pairing reagent binds to the SPE surface and then the polar analyte binds to the ion-pairing reagent. If SPE is followed by mass spectrometric analysis, volatile ion-pairing reagent should be used at a relatively low (maximum 10 mM) concentration to avoid suppression effects.

Normal-phase SPE methods use apolar solvents and a polar stationary phase (cartridge packing). These are mostly applied to clean up and concentrate polar analytes in mid- to nonpolar matrices (e.g., acetone, chlorinated solvents, and hexane). The most widely used NP packings are pure or occasionally functionalized silica (cyano, amine, and diol phases). Retention of the analytes is primarily due to interactions between polar functional groups between the analyte and silica packing. Compounds absorbed on the cartridge are then eluted using a polar solvent, such as methanol or water. A typical example to extract slightly polar drugs

from biological matrices such as plasma is the following. The first step is LLE of plasma with hexane:ethylacetate 95:5 mixture. A silica SPE cartridge is conditioned with two bed volumes of hexane, followed by equilibration with two bed volumes of hexane:ethylacetate 95:5 mixture. The hexane:ethylacetate extract of the sample is diluted with hexane:ethylacetate 95:5 to about one bed volume and is applied to the cartridge. It is washed with one bed volume of hexane:ethylacetate 95:5. Finally, the analytes are eluted with 0.5–3 ml of hexane:ethylacetate 2:1 mixture.

Ion-exchange SPE is best used for the extraction of ionized compounds or compounds which may be ionic by shifting the pH. Negatively charged compounds can be retained with strong anion exchange (SAX) or weak anion-exchange (silica-based amine) cartridges. Positively charged compounds can be retained by strong cation exchange (SCX) or weak cation exchange (WCX) phases. Retention mechanism is based on electrostatic attraction between the charged functional group on the compound and the charged group that is bonded to the silica surface. In the case of SAX, SPE packing material contains aliphatic quaternary amine groups bound to the silica surface. This is a strong base in the form of a permanent cation (pK_a of a quaternary amine is very high, greater than 14) that attracts anionic species present in the solution. Likewise, strong or weak anion-exchange phases may be used to extract positively charged analytes.

More sophisticated SPE methodologies are also in use, for example, involving a mixture of different cartridge packings (so-called mixed phases). Several SPE steps can also be performed in sequence, resulting in highly efficient purification. SPE is also well adapted to high-throughput operations, using 96-well sample plates, which may be used to collect samples in a conventional 96-well plate. Method optimization is made easier by using special plates where different SPE packings are placed, so the result of parallel experiments can be evaluated and the method (SPE cartridge–solvent combination) giving the best results can be determined easily.

Application of SPE offers an opportunity to obtain an exceptionally clean, concentrated fraction of analytes from very complex matrices. This approach is attractive in biological sciences, since the samples are nearly always highly complex mixtures. The SPE process is based on physicochemical sorption processes and it does not involve chemical treatments, so it is less prone to introduce artifacts. SPE is well suited to a sample size small enough to be easily available (less or much less than a milliliter of body fluid or a gram of tissue), but large enough that even minor components could be identified in the subsequent analysis step. Disadvantages of SPE include that it is a multistep, labor-intensive process. Automation is possible but expensive. In some cases, irreversible adsorption of the analytes can occur on the SPE cartridge, leading to recovery problems, especially when very small sample size is used.

3.3. ZipTip® sampling

To overcome problems with automation and recovery, ZipTip sampling has been developed, which is best regarded as miniaturized and simplified SPE equipment. It is frequently used in the field of proteomics, mostly as the last sample preparation step. Sample amount is only a few (3–30) μl, containing less than a picomole protein digest. Commonly it is used to remove salts and detergents from the sample, e.g., after tryptic digest, just before mass spectrometric analysis.

The ZipTip equipment is similar to a conventional pipette tip, packed with a small amount of sorbent. Usually 10 μl pipette tips are used with 0.2–0.6 μl bed volume; the sorbent is packed into the tip region. Similar types of sorbents are used as in SPE, although with less variety. Most common are C18, C4, occasionally SCX, and metal chelate stationary phases. Operation is similar to SPE: First, the tip is conditioned by aspirating and dispensing a few microliters of clean solvent. Then, the sample is aspirated and dispensed, followed by washing with the same solvent (usually water). The sample is bound to the sorbent, while contaminants (salts, detergents, etc.) are washed away. Last, the sample is eluted in a few microliters of stronger solvent, e.g., 0.1% formic acid:75% methanol. The recovered samples are directly transferred to the MALDI target or injected into the mass spectrometer or loaded directly into a nanospray needle [14]. For MALDI analysis, it is also common to elute the sample with the MALDI matrix, spotting it directly onto the target.

There are different resins available. C18 and C4 packings are often used for desalting and concentrating peptides and proteins, SCX phases for removing detergents, and metal chelate packings for enriching phosphopeptides. The main advantages of using ZipTip are that it is very simple and fast (requires less than a minute), and recovery problems are minimized (due to the use of only very small sorbent size).

3.4. Solid-phase microextraction

A recent and very successful approach to sample preparation is SPME invented by Pawliszyn and coworkers [15], and reviewed recently [16]. SPME integrates sampling, extraction, concentration, and sample introduction into a single, solvent-free step. It is excellent as a sampling tool for GC and gas chromatography–mass spectrometry (GC–MS). It is routinely used for extraction of volatile and semivolatile organics, mostly as headspace (HS) analysis.

The SPME apparatus looks like a modified syringe (see Fig. 5) consisting of a fiber holder needle and a fiber assembly, the latter equipped with a 1–2 cm long retractable SPME fiber. The fiber itself is a thin fused-silica optical fiber, coated with a thin polymer film (such as polydimethylsiloxane, PDMS), as shown

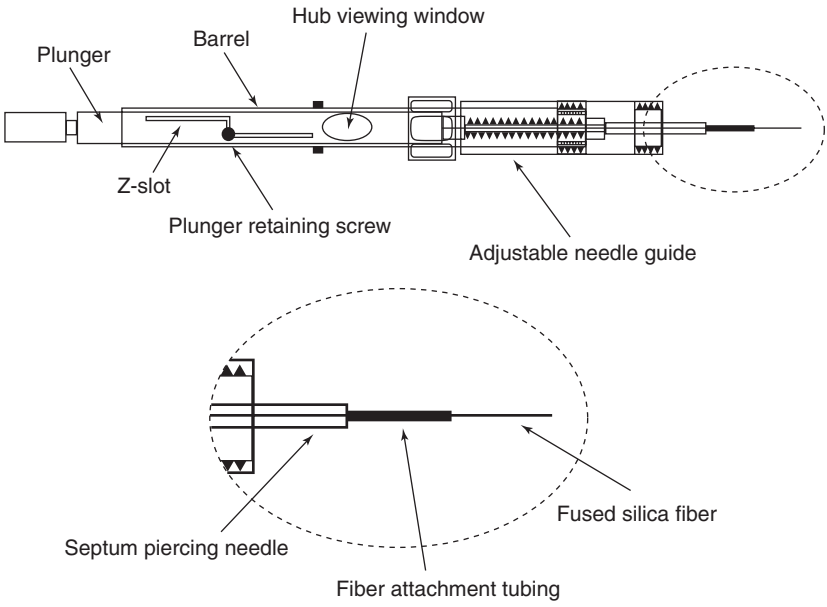


Fig. 5. Schematic diagram of an SPME needle.

in Fig. 5. The polymer coating on the fiber acts as a sponge, concentrating the analytes by absorption/adsorption processes. The principle of extraction is analogous to that of GC, based on a partitioning process [17]. There are various types of fibers; the choice depends mainly on the polarity and volatility of analytes. Extraction usually takes place in the gas phase (HS sampling), though occasionally the fiber may be immersed into a liquid sample.

During sampling, the SPME needle is first introduced into the sample vial usually by piercing a septum, as shown in Fig. 6. Then the extraction fiber is pushed out of the needle, either into a gas, into the HS of a sample, or immersed into a liquid sample (direct immersion (DI) or DI-SPME analysis). Agitation of the sample (by stirring or by sonication) improves transport of analytes from the bulk phase, accelerating equilibration. After equilibrium is reached, the fiber is withdrawn into the needle, taken out of the sample vial, and introduced into the GC injector. The fiber is exposed; analytes are desorbed and carried onto the separation column by the carrier gas. The GC injector is usually at a high temperature, so desorption is fast. As there is no solvent, splitless injection can usually be performed, making the analysis very sensitive. Finally, the SPME device is withdrawn from the GC injector. The SPME fibers can easily be cleaned by heating. This is usually performed by keeping the fiber in the GC injector for some time (switching to split mode after injection) or using a special syringe cleaner. In the case of HS analysis, fibers can be reused hundreds of times, so SPME operation is relatively inexpensive.

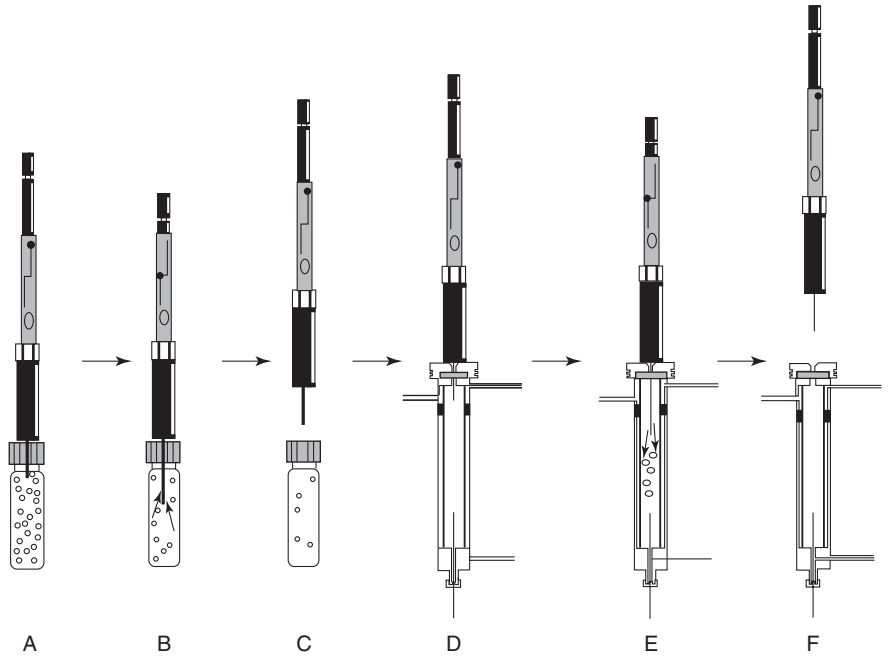


Fig. 6. Illustration of the SPME process.

The sampling process in SPME depends on a number of parameters. Probably the most important is temperature—in the case of both traditional HS analysis and SPME. Polymer coating has a similar influence as the stationary phase in GC. The types of SPME fibers are analogous to the types of GC columns available. Probably, most commonly the apolar PDMS coating is used. Film thickness not only relates to sample capacity and volatility of the analyte but also has an influence on the time needed to establish equilibrium. Establishing equilibrium conditions depends on many parameters (volatility of the sample, volume of head space, intensity of stirring, etc.) and may require a few minutes or several hours. The SPME fiber has to be immersed into the HS either till equilibrium is established or, in order to obtain good reproducibility, for a well-defined and precisely controlled period. Note that SPME fibers should be carefully handled as they are fragile and the fiber coating can be easily damaged.

The most widespread SPME applications utilize injection to a GC (or GC–MS) system. Thermal desorption in the GC injection port depends on the temperature, exposure time, analyte volatility, and the type and thickness of the fiber coating. To ensure a high linear flow, a narrow-bore GC injector insert is required. The fiber needs to be inserted to a depth corresponding to the hot injector zone. This is important because the temperature varies along the length of the injector, and desorption of analytes is very sensitive to the temperature.

Desorption time is generally in the 10–100 s range, but it needs to be optimized. To ensure high sensitivity, the injector is usually operated in splitless mode—this is possible as no solvent is used in SPME. A frequent practical problem using SPME is that GC septa are easily damaged with the wide (24-gauge) SPME needles. To avoid septum coring, predrilled GC septa or septum-less injector valves may be used.

The main advantages of SPME are good analytical performance combined with simplicity and low cost [18]. It is well adapted to most compounds, which can be studied by GC. SPME produces clean and concentrated extracts and is ideal for GC–MS applications [17,19,20]. SPME is suitable for automation, which not only reduces labor costs but also often improves accuracy and precision. The main disadvantage of SPME is that it is less well adapted for quantitative analysis. Accurate measurements (in terms of quantitation) require careful control of a number of experimental variables, which is elaborate and not always feasible.

Success of SPME coupled to HS analysis using GC and GC–MS prompted studies to extend this technique to study nonvolatile, polar compounds. Compounds that are amenable for GC analysis, but have low vapor pressure, may be studied by DI SPME. In this case the fiber is immersed into the liquid sample where extraction takes place. Subsequently the fiber is inserted into the GC injector. This technique retains most advantages of SPME, notably simple operation, solvent-free extraction, and high sensitivity. However, in this case the fiber is easily damaged (mainly by irreversible absorption of large polar molecules) and can be reused only few times (which makes operation expensive). Extraction efficiency and the time necessary to reach equilibrium are influenced by several parameters (such as agitation, pH, ion strength, etc.), and reproducibility (in terms of quantitation) is usually worse than for HS analysis. To prevent the loss of polar analytes, deactivation of glassware before use is recommended [21].

A further extension of the SPME technique is coupling to HPLC (or HPLC–MS), which extends the method to (usually polar) compounds that are not amenable for GC analysis. This is also performed by DI sampling. After extraction, compounds bound to the fiber are extracted by a strong solvent. Note that the much simpler thermal desorption cannot be used, as the compounds to be studied are not volatile. This extraction takes place in a special extraction chamber, connected to a modified Rheodyne or Valco valve of an HPLC system. To facilitate HPLC analysis, a special, so-called in-tube SPME device has been developed. With this technique, organic compounds in aqueous samples are directly extracted from the sample into the internally coated stationary phase of a capillary column and then desorbed by introducing a moving stream of mobile phase.

In conclusion, SPME is an ideally suited sample preparation method to prepare samples for GC or GC–MS. Most compounds well suited for GC analysis can be extracted and concentrated using SPME, which is easy and results in excellent analytical performance. SPME is, however, less well adapted as a sample preparation for HPLC to study polar or large molecules.

4. Automation and high throughput

High-throughput (HT) analysis is becoming more and more important. It means analysis of dozens, hundreds, or even thousands of samples per day in a given laboratory or on a particular instrument. In the biomedical field, it makes large-scale experiments and testing a large number of compounds (e.g., combinatorial libraries for a particular biological effect) possible, while in the clinical field, it is essential for population-wide screening, but often also to test a particular group of patients.

The main methodologies needed for HT are automation and robotization. Most analytical methods, including sample preparation, can be adapted for this purpose. HT requires very large investment in instrumentation (and also in method development), but running costs (per sample analyzed) are much lower, mainly due to reduction of manual labor. A further advantage of automatic/robotic operation is that of finding qualified personnel, which is becoming more and more difficult.

Performing high-throughput analysis requires careful design. First, the analytical method needs to be developed in "low throughput," keeping in mind requirements for future HP experiments. In the next step this should be adapted for HT. This usually means a simplification of sample handling, parallel manipulation of a large number of samples, speeding up all steps which include long waiting time or analysis time. Bottlenecks in the sample flow should be identified and eliminated. Note that to perform HT experiments the sample preparation and analytical methods often need to be changed. Sample preparation and analytical methods need to be very robust to perform under HT conditions. In most cases analytical performance (e.g., detection limit) is not as good as in conventional analysis—this needs to be taken into account in the development phase. On the other hand, reproducibility is often improved using automatic and robotic techniques. An integral part of HT operation is proper labeling (usually bar codes are used), managing sample flow, evaluation, and reporting the results. These are controlled by special (often individually developed or adapted) software.

Sample preparation involves one step in any HT experiment—as this step is usually the most time-consuming, it is an essential aspect of designing and performing HT analysis. Luckily, most sample preparation methods can be adapted for HT. Probably most important is the use of well plates: a two-dimensional array of sample vials, usually handling 96-well (occasionally 384) samples. These are well standardized and can be used in most commercial analytical instruments. Special automatic pipettes are developed to use with these well plates, containing (8, 10, or 12) parallel pipette tips. These well plates are often used in manual operation, but they still allow high-throughput operation (96 samples can be prepared instead of 1, requiring only somewhat more time). There are commercially available and often-used versions of most sample preparation laboratory equipments (centrifuges, thermostats, automatic injectors, etc.) that can be used in combination with these well plates. The following three examples will give

information related to HT operation of techniques discussed earlier: protein precipitation, LLE, and SPE.

The protein precipitation can be easily performed by an automated liquid handler (e.g., Packard Multiprobe II, Tecan Genesis, Gilson 215, Tomtec Microtape, etc.) in a well plate or a microwell plate, by adding a water-miscible organic solvent (typically 3:1 (v/v) ratio) to the biological matrix. Proteins are then collected in the bottom of the well by centrifugation and the handler can take the aliquot of the clear liquid and transfer it to a well plate prior to LC–MS injection. Additional tasks such as adding internal standards for calibration and quality control can also be handled by automated liquid handlers.

LLE can also be automated by using a liquid handler (Tomtec Quadra96) using a microplate platform. The biological matrix is mixed with immiscible organic solvent (e.g., chloroform and ether). Depending on the particular application, LLE may require manual intervention, such as decapping tubes or vortexing (shaking). Removal of the organic layer can be done automatically; freezing the aqueous layer reduces possible errors in sampling.

SPE can be automated offline by using SPE microplates or by a multiple-tip liquid-handling workstation (e.g., Zymark XP series). The SPE extraction can be performed online as well, using a versatile automated system such as Prospekt from Spark Holland. This automated unit includes a solvent delivery unit, a cartridge transport, a sealing mechanism, and an autosampler. Samples are introduced by the autosampler and loaded to a disposable cartridge (2 mm $\times$ 10 mm); a weak solvent then elutes the unretained salts and the polar matrix components. An optimized sequence of solvents is used to wash the trapped analytes to an analytical column for HPLC separation, followed by detection. Each sample is processed by a single-use, disposable cartridge, so carryover is minimal [2].

5. Outlook

Sampling and sample preparation are cornerstones of any analytical methodology. Probably the most important advances in this field are miniaturization, simplification of methodologies, and their adaptation to HT. On the technical side, using devices analogous to very simplified chromatography have become widespread. These usually take the form of a small disposable cartridge, such as SPE, ultrafiltration tubes, etc. These are very efficient, are easy to adapt to the required problem, need small sample size, minimize problems related to contamination (carryover), and are easy to automate. Most other sample preparation equipments serve to support or complement these techniques.

Future developments in sampling are likely to follow these lines. There is more and more need for automatic operation, for both HT and reduction of manual labor.

References

1. Pawliszyn, J., *Solid Phase Microextraction: Theory and Practice*. Wiley-VCH, New York, 1998.
2. Pawliszyn, J., *Sampling and Sample Preparation for Field and Laboratory*. Elsevier, Amsterdam, 2002.
3. Richard, S.S., *Clinical and Functional Histology*. Little Brown and Co., Boston, MA, USA, 1984.
4. Watson, I.D., *Sample Preparation for Biomedical and Environmental Analysis*. Plenum Press, New York, NY, USA, 1994.
5. Sporkert, F. and Pragst, F., Use of headspace solid-phase microextraction (HS-SPME) in hair analysis for organic compounds. *Forensic. Sci. Int.*, **107**, 129–148 (2000).
6. Smith, R.M., Before the injection-modern methods of sample preparation for separation techniques. *J. Chromatogr. A*, **1000**, 3–27 (2003).
7. Buszewski, B. and Ligor, T., Single-drop extraction versus solid-phase microextraction for the analysis of vocs in water. *LC-GC Eur.*, **15**, 92–96 (2002).
8. Rasmussen, K.E., Bjergaard, S.P., Krogh, M., Ugland, H.G. and Gronhang, T., Development of a simple in-vial liquid-phase microextraction device for drug analysis compatible with capillary gas chromatography, capillary electrophoresis and high-performance liquid chromatography. *J. Chromatogr. A*, **873**, 3–11 (2000).
9. Jonsson, J.A. and Mathiasson, L., Membrane-based techniques for sample enrichment. *J. Chromatogr. A.*, **902**, 205–225 (2000).
10. Halvorsen, T.G., Bjergaard, S.P. and Rasmussen, K.E., Liquid-phase microextraction and capillary electrophoresis of citalopram, an antidepressant drug. *J. Chromatogr. A*, **909**, 87–93 (2001).
11. Soriano, T., Jurado, C., Menendez, M. and Repetto, M., Improved solid-phase extraction method for systematic toxicological analysis in biological fluids. *J. Anal. Toxicol.*, **25**, 137–143 (2001).
12. Leon-Gonzalez, M.E. and Perez-Arribas, L.V., Chemically modified polymeric sorbents for sample preconcentration. *J. Chromatogr. A*, **902**, 3–16 (2000).
13. Carson, M.C., Ion-pair solid-phase extraction. *J. Chromatogr. A*, **885**, 343–350 (2000).
14. Van Pelt, C.K., Zhang, S., Fung, E., Chu, I., Liu, T., Li, C., Korfmacher, W.A. and Henion, J., A fully automated nanoelectrospay tandem mass spectrometric method for analysis of caco-2 samples. *Rapid Comm. Mass Spectrom.*, **17**, 1573 (2003).
15. Arthur, C.L. and Pawliszyn, J., Solid phase microextraction with thermal desorption using fused silica optical fibers. *Anal. Chem.*, **62**, 2145 (1990).
16. Vas, G. and Vekey, K., Solid-phase microextraction: a powerful sample preparation tool prior to mass spectrometric analysis. *J. Mass Spectrom.*, **39**, 233 (2004).
17. Ulrich, S.J., Solid-phase microextraction in biomedical analysis. *J. Chromatogr. A*, **902**, 167–194 (2000).
18. Kataoka, H., Lord, H.L. and Pawliszyn, J., Applications of solid-phase microextraction in food analysis. *J. Chromatogr. A*, **880**, 35–62 (2000).
19. Lancas, F.M., Queiroz, M.E.C. and Eugenia, M., Tips on preparing plasma samples for drug analysis by SPME. *LC-GC*, **22**, 970 (2004).
20. Snow, N.H., Solid-phase micro-extraction of drugs from biological matrices. *J. Chromatogr A*, **885**, 445–455 (2000).
21. Wercinski, S.C.S., *Solid Phase Microextraction—A Practical Guide*. Marcel Dekker, New York, 1999.

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 5

Separation methods

KORNÉL NAGY* and KÁROLY VÉKEY

Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary

1. Introduction

Biomedical analysis nearly always relates to complex matrices. Following sampling and sample preparation, chromatography is the primary technique to separate mixtures into their chemical components. In most cases this step is necessary before structural analysis or quantitation can be performed. In general, in chromatography a fluid (containing the multicomponent sample) moves over a nonmoving (stationary) phase. When there is a strong interaction between a given compound

*Corresponding author. Present address: Nestlé Research Center, Lausanne.
Tel.: + 4121-785.8290; Fax: + 4121-785.8566. E-mail: kornel.nagy@rdls.nestle.com

and the stationary phase, the migration of the component will slow down. When the interaction is minimal, the compound will migrate with the same velocity as the mobile phase. This results in the separation of the various components of a mixture. Chromatography yields two basic pieces of information on the separated components: the degree of retention (characteristic of molecular structure) and signal intensity (related to the amount of the component). Chromatography is usually based on the distribution of the various compounds between a stationary and a mobile phase and/or on the electrophoretic mobility of the compounds. Separation can be implemented in several ways. The three major groups of chromatographic techniques are (1) gas chromatography (GC), (2) high-performance liquid chromatography (HPLC), and (3) electrophoretic techniques. These techniques differ in the applied mobile phase (gas or liquid) and in the type of retention and flow mechanism (see the following text).

As a rule of thumb, GC is used for the separation of volatile compounds. Thus, it is useful for determination of low-molecular-weight compounds (below 500 Da) but cannot be used for large, highly polar or thermally labile compounds. Implementation of GC is simple and routine. GC is mostly coupled with flame ionization detection (FID), electron capture detection (ECD), or mass spectrometry (MS).

HPLC is used for nonvolatile compounds and is well suited for the analysis of low- and high-molecular weight compounds such as peptides and proteins. HPLC is mostly coupled with ultraviolet visible (UV–VIS) wavelength spectroscopy or mass spectrometric detection.

Electrophoretic techniques are used for nonvolatile compounds, which are permanently or temporarily charged, such as proteins or organic salts. Electrophoretic techniques have an increasing importance in biomedical fields, such as proteomics.

Chromatographic techniques help in ensuring the selectivity and sensitivity necessary for clinical analysis and contribute to the success of the analytical process. Complete separation in biological samples is rarely feasible. The purpose is more often to reduce the complexity of a mixture, enrichment of a given component, or removal of interferences. In most cases chromatography is used for analytical purposes, although it may also be used in preparative chemistry. The expression "chromatographic techniques" covers a wide range of analytical methods that can separate chemical components of a sample on the basis of their molecular properties such as size or polarity. Several detailed studies can be found in the literature discussing the implementation and mechanism of separation techniques [1–11]. Commonly used chromatographic methods are listed in Table 1. In the present chapter we only provide a basic description and a brief overview: First GC, then HPLC, and finally electrophoretic techniques are discussed.

Chromatography is a collective name for methods that separate compounds based on their interaction with a mobile phase (in which the sample is dissolved or mixed) and a stationary phase. For instance, the strength of interaction between an apolar compound and an apolar stationary phase is strong; thus, the compound will

Table 1
Commonly used chromatographic methods

	Technique	Acronym	Meaning
1	2D electrophoresis	2DE	Combined application of electrophoresis and isoelectric focusing in a thin gel-based layer.
2	Affinity chromatography		A method of separating and purifying compounds using their biochemical affinity to the stationary phase.
3	Capillary electrophoresis	CE	Usually the same as CZE, but sometimes used as a collective name for several electrophoretic methods.
4	Capillary gel electrophoresis		Electrophoretic separation performed in gel-filled capillary columns.
5	Capillary isoelectric focusing		An electrophoretic technique that separates and focuses compounds into peaks according to their isoelectric points.
6	Capillary zone electrophoresis	CZE	A separation technique based on the electrophoretic mobility of analytes in electrolytes. It is performed in fused silica capillaries by applying high voltage to the ends of the column.
7	Column chromatography		Liquid chromatography performed by moving the mobile phase through a packed column using gravity.
8	Gas chromatography	GC	A type of chromatography when the mobile phase is a gas.
9	Gel filtration		Size exclusion chromatography performed with aqueous solvents for the separation of biopolymers.
10	Gel permeation chromatography	GPC	Size exclusion chromatography performed with organic solvents for the separation of synthetic polymers.
11	Gradient elution		A type of elution in liquid chromatography where the composition of mobile phase is changed during the experiment.
12	High-performance liquid chromatography	HPLC	A type of chromatography when the mobile phase is a liquid and is transferred through the column via mechanical pumps.
13	Ion-exchange HPLC	IE-HPLC	A type of liquid chromatography in which the retention is based on ion-pair formation.
14	Liquid chromatography	LC	A type of chromatography in which the mobile phase is liquid.
15	Multidimensional chromatography		A type of chromatography in which basically different separation processes are applied on the same sample in a consecutive arrangement.

(*continues*)

Table 1
Continued

	Technique	Acronym	Meaning
16	Normal-phase HPLC	NP	An expression to characterize an HPLC system. Under NP circumstances the mobile phase is less polar than the stationary phase. Typical example is elution on a silica column with hexane solvent.
17	Reverse-phase HPLC	RP	An expression to characterize a chromatographic system. Under RP circumstances the mobile phase is more polar than the stationary phase. Typical example is methanol/acetonitrile solvent on octadecyl silica phase.
18	Size exclusion chromatography	SEC	A type of liquid chromatography in which the retention is based on the hydrodynamic size of the analytes.
19	Sodium dodecyl sulfate polyacrylamide gel electrophoresis	SDS-PAGE	Electrophoretic separation performed in polyacrylamide gel in SDS-rich media. It is used to separate compounds according to their molecular weight.
20	Thin-layer chromatography	TLC	A type of liquid chromatography performed on a thin two-dimensional layer used as stationary phase.

be strongly retained on the column. On the contrary, a polar compound interacts less strongly with an apolar stationary phase; thus, it moves through the column at a faster rate. When a mixture of two different compounds is injected onto the top of a column, they will be retained to a different degree and will arrive to the end of the column at a different time. If a detector is placed at the end of the column, then the signal as a function of time depicts the elution sequence of the compounds as consecutive peaks and this is called chromatogram (illustrated in Fig. 1).

Among the various features that characterize the chromatograms the most important ones are retention time, resolution, and signal intensity. Retention time (t_R) is the time elapsed between sample introduction (beginning of the chromatogram) and the maximum signal of the given compound at the detector. The retention time is strongly correlated with the physicochemical properties of the analyte; thus, it provides qualitative information about the compound, which in simple cases may be identified using this information. Resolved compounds always have different retention times. Retention volume is a related parameter. It is the volume of the mobile phase that is required to elute a given compound from sample introduction to the detector. It can be calculated by multiplying the retention time with the flow rate of the mobile phase. The retention factor

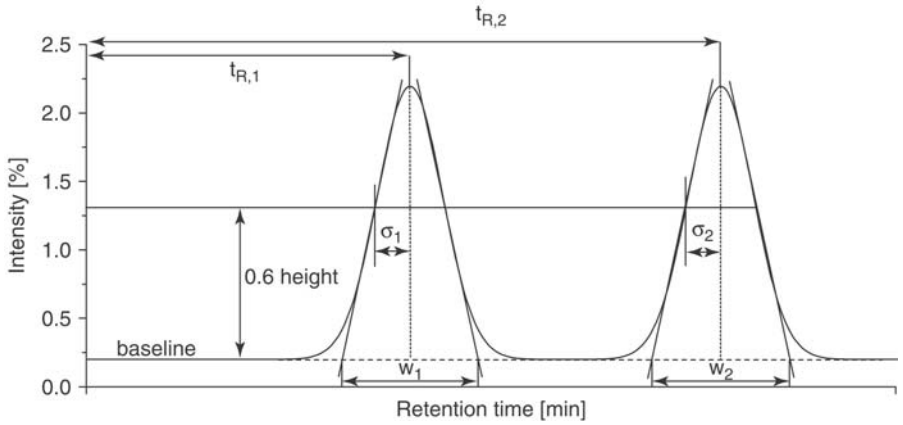


Fig. 1. An example of chromatogram showing parameters used to characterize chromatographic peaks.

(or capacity factor, often abbreviated as k) is also related, and is a measure of distribution of a given compound between the stationary and mobile phases. It expresses the strength of adsorption of the analyte on the stationary phase.

$$k = \frac{n_S}{n_M} = \frac{t_R - t_0}{t_0} \tag{1}$$

where n_S is the number of moles of the given compound in the stationary phase, n_M the number of moles of the same compound in the mobile phase, t_R the retention time of the compound, and t_0 the dead time or holdup time (retention time of a compound which does not interact with the stationary phase).

Resolution is a measure of the quality of separation between two components and is defined as:

$$R = 2\frac{t_{R,2} - t_{R,1}}{w_2 + w_1} \tag{2}$$

where R is the resolution, $t_{R,1}$ the retention time of the first component, $t_{R,2}$ the retention time of the second component, and w_1 and w_2 are the peak widths of the first and second components (projected onto the baseline), respectively. Its analytical meaning is similar to the selectivity, which is defined as:

$$\alpha = \frac{t_{R,2} - t_0}{t_{R,1} - t_0} \tag{3}$$

Both selectivity and resolution are properties of an experimental method, which reflects the extent of discriminating power between two compounds. If resolution

between the two compounds is good, then they are fully separated, arrive to the detector at different times, and do not interfere with each other. Good resolution of compounds is a prerequisite of quantitative analysis. Resolution is strongly related to the widths of chromatographic peaks. The theoretical plate number is a measure of peak width, and can be calculated by the following equation:

$$N = \frac{t_R^2}{\sigma_t^2} \tag{4}$$

where σ_t is half of the peak width measured at 0.6 peak height. The higher the theoretical number of plates the narrower the peaks are. The expression *efficiency* or *plate number/meter* (theoretical number of plates for a 1 m long column) is often used to characterize and compare performance of different stationary phases. Typical *efficiency* values fall in the 10,000–100,000 range for modern LC and in the 2500–5000 range for GC. Note, however, that HPLC columns are typically ~10 cm, while GC columns ~50 m long, so the plate number of GC columns is much higher than that of the HPLC, resulting in better resolution.

The height and area of the recorded peaks are also very important as these reflect the quantity of a given compound. Accordingly, accurate determination of peak height and peak area is a prerequisite of quantitative analysis. Note that the sensitivity of a detector is different for various compounds and different detectors also have different relative sensitivities. To perform quantitative analysis, careful calibration is always needed.

Chromatograms are usually obtained by the elution technique: The sample is injected onto the column and is carried by a fluid through the column to the detector, so various compounds arrive to the same place at different times. Chromatographic peaks obtained by the elution technique ideally possess a Gaussian-like shape. In practice, peak shapes are often different, frequently indicating problems with the separation process. The two common problematic peak shapes are shown in Fig. 2.

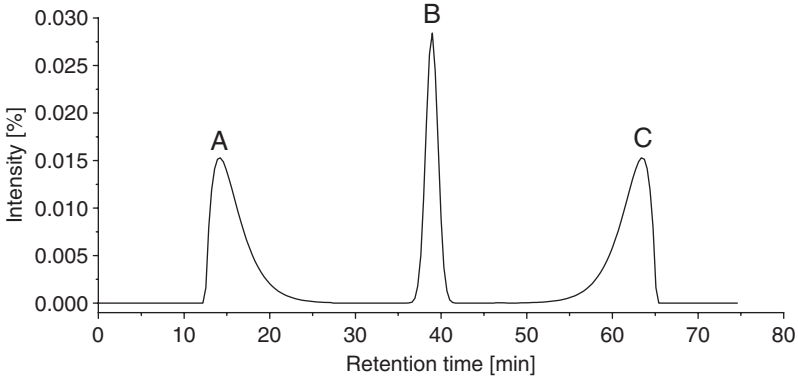


Fig. 2. Illustration of typical peak shapes.

Peak A illustrates strong tailing, which is most often the result of loose connections or the presence of large dead volumes in the system, but may also be caused by problems with the separation process. Dead volumes are often present when metal fittings are applied, so the use of flexible and easily adjustable "finger-tight" PEEK fittings and ferrules is advisable in most cases. Peak B is an ideal Gaussian-shaped peak, while peak C depicts a peak with strong fronting. This is typically the result of overloading the column and can be avoided by diluting or decreasing the amount of sample injected. Detailed description on peak shapes and their determining factors can be found elsewhere [12–14].

A common aim in chromatographic method development is to produce the highest possible resolution of the components within the shortest possible time. This is possible only if the peaks are narrow. Selective and fast methods require high theoretical plate number, high selectivity, and short retention times as shown in the model chromatogram in Fig. 3C. Other, less desirable examples are also shown in the figure, where A illustrates a nonselective, slow chromatogram, B a nonselective, but fast chromatogram, and D a selective, but slow chromatogram.

Physicochemical properties of different compounds span a very large range, so it is impossible to develop a universal method well suited for all analytical purposes. Various chromatographic techniques were developed and optimized for different analytes (a list of common techniques is provided in Table 1). As a rule of thumb, GC methods are useful for determining low-molecular-weight (below 500 Da), not-very-polar, thermally stable compounds, but cannot be used, e.g., for determination of peptides or proteins. LC is the method of choice for nonvolatile, polar or thermally labile compounds, which represent the vast majority of

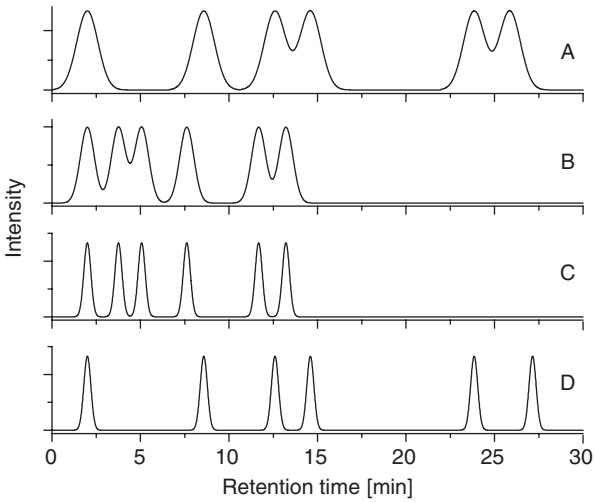


Fig. 3. Illustration of variations in selectivity and retention time in chromatograms.

compounds found in biological systems. The most widespread liquid chromatographic method is HPLC, in which high pressure (50–400 bar) pushes the mobile phase through the column. Electrophoretic techniques are increasingly used and are well suited for the analysis of ionized (or ionizable) molecules (including macromolecules).

Chromatographic separations must be followed by detection to see the result of the separation process. GC typically uses FID, ECD, or mass spectrometric detection. In HPLC, UV–VIS and MS detection are very common, while one-dimensional (1D) and two-dimensional (2D) gel electrophoretic techniques generally use staining to make component spots visible. In the last decade, detection using MS was particularly often used as it provides structural information and can be utilized to increase selectivity and specificity and to lower detection limits.

Chromatography combined with MS is also often considered as the most efficient "gold standard method." MS provides unusually high selectivity and specificity because it delivers molecular mass and structural information on a given compound/chromatographic peak. Using MS allows faster chromatograms, as problems caused by coelution may be overcome by selective MS detection. Using high resolution or tandem mass spectrometry (MS/MS) further increases selectivity. The use of MS detectors often results in lower detection limits, as its high selectivity reduces chemical noise (which is often the most serious issue in analysis, especially in the case of biological samples). MS expands the applicability of chromatographic methods, e.g., by overcoming problems of UV detection, as in the case of apolar compounds lacking suitable UV absorption band. On the contrary, mass spectrometric detection puts some restrictions on the chromatographic method used. A typical limitation is that the commonly used potassium phosphate buffer blocks the orifice of a mass spectrometer. To overcome this problem, only volatile buffers, such as ammonium formate, can be used in HPLC/MS applications [15]. In general, combination of MS with chromatography provides far more advantages than disadvantages. The most important advantages are increased selectivity, shorter analysis times, lower detection limits (especially for biological samples), and simplification of sample preparation protocols.

Quantitation is a widely used application of chromatography for a wide variety of compounds with biological importance. Just like in any other case, quantitation is based on a calibration curve that determines the relationship between the measured signal and the concentration of the compound of interest. Plotting detector response as a function of analyte concentration, one can obtain the calibration curve. Ideally there is a linear relationship between the detector signal and the sample amount, but in practice calibration curves often deviate from linearity both at very low and very high sample concentrations. For quantitation purposes, the linear middle range is desirable. Linearity of the calibration curve is often characterized

by the R^2 value. The closer this value is to 1, the better is the linearity (typically linearity better than 0.99 is required). The lower end of the linear range of the calibration curve determines the limit of quantitation. Typically this is defined by the position, where the calibration curve deviates from the trendline by 10%. This is the smallest analyte amount that can be quantified by the method. The smallest sample amount that can be detected by the given method (defined typically by a signal-to-noise ratio of 3) is the limit of detection. Note that it is always smaller than the limit of quantitation, which is frequently estimated as 10 times signal-to-noise ratio. Repeatability and reproducibility are also important parameters of an analytical process, and must be determined to check the reliability of the results. Typically "same day" and "day-to-day" repeatability values are calculated from 3 to 10 replicate measurements. For the validation of a chromatographic method other parameters are often needed as well, such as robustness, precision, accuracy, and recovery—but these topics are outside the scope of this chapter [16–18].

Chromatography—like other sciences—uses special terminology and acronyms. Navigating through these may occasionally become frustrating for the nonspecialist reader. As a guide, commonly used terms and acronyms in chromatography are summarized in Table 2.

Table 2
Commonly used terms and acronyms in chromatography

	Term	Acronym	Meaning
1	Capacity factor	k	A number characterizing the capacity and retention of chromatographic columns.
2	Carrier gas		The gas used in GC as the mobile phase to carry the analytes from the injector to detector.
3	Chromatogram		Plot of the detector response as a function of time.
4	Coating (GC capillary)		Mostly polymeric material on the inner wall of GC capillaries used as the stationary phase.
5	Dead time	t_0, t_M	Time interval for an absolutely nonbinding compound to travel through the column. Also called holdup time.
6	Efficiency		Analytical power of a column filling material expressed as number of theoretical plates per 1 m long column.
7	Effluent		The mobile phase is called effluent when leaving the column.
8	Eluent		Mobile phase that elutes the analytes.
9	Elution		The process of driving the analyte from the entry to the end of the column.

(*continues*)

Table 2
Continued

	Term	Acronym	Meaning
10	End-capping	ec	An additional treatment of HPLC columns. Residual silanol group are reacted by monofunctional chlorosilanes, improving column properties.
11	Flow rate		The speed of the mobile phase given as volume/time.
12	Height equivalent to a theoretical plate	HETP, *H*	Same as theoretical plate height. A number characterizing the quality of the column filling. It is expressed as the length of a column that would be equivalent to one theoretical plate determined by the plate theory.
13	Isotherm		A process or experiment observed or performed at a constant temperature.
14	Kovats index		A reference number characterizing the polarity and retention time of compounds. It expresses the carbon number of an alkane reference compound that exhibits the same retention as the analyte.
15	Limit of detection	LOD	The smallest analyte amount, which can be detected by a method (usually estimated as 3 times signal/noise).
16	Limit of quantitation	LOQ	The smallest analyte amount that can be quantitated by a method (usually estimated as 10 times signal/noise).
17	Make up gas		An auxiliary gas used in GC to aid the flame ionization process.
18	Matrix		The entire sample excluding the analyte. It is practically the environment of the analyte in the sample.
19	Mobile phase		It is the same as the eluent in HPLC or carrier gas in GC. It is the phase that can be moved relative to another (stationary) phase.
20	Octadecyl silica phase	C18	A widely used HPLC column filling. It is chemically modified silica gel that contains octadecyl silane chains on the surface.
21	Peak broadening		A phenomenon that deteriorates chromatographic performance (broadens the peak). It is the result of various disrupting effects that may occur during chromatography.
22	Peak fronting		A phenomenon when the symmetry of the chromatographic peak deteriorates. The left side of the peak broadens.
23	Peak tailing		A phenomenon when the symmetry of the chromatographic peak deteriorates. The right side of the peak broadens.

Table 2
Continued

	Term	Acronym	Meaning
24	Peak width	Sigma, *w*	Width of a chromatographic peak measured either on the projected baseline (*w*) or at 60% height (also called 2 times sigma).
25	Purge		A collective name for all those events where a chromatographic volume is cleaned by flushing it with mobile phase at high flow rate.
26	Repeatability		Repeatability characterizes the analytical power of the actual method. It shows how large is the deviation of the results if one person repeats the experiment using the very same conditions on the very same instrument with the same sample.
27	Reproducibility		Reproducibility characterizes the analytical power of the actual method. It shows how large is the deviation of the results if different persons repeat the same experiment using the same conditions on different instruments.
28	Resolution	*R*	Resolution is the extent of separation between two components.
29	Retention factor	*k*	The extent of retention in a given chromatographic system is characterized by the retention factor. If retention factor is high, then the analyte binds strongly to the stationary phase and the retention time will be long.
30	Retention index		A number expressing the extent of retention of a given compound compared to the retention of the reference compound.
31	Retention time	t_R	Time elapsed between injection and maximum detector response for a compound.
32	Retention volume	VR	It is the volume of eluent that passes through the column while eluting a given compound.
33	Robustness		Robustness is a characteristic of the developed method. It represents the sensitivity of the method to the change of experimental parameters.
34	Selectivity	α	Selectivity is the measure of discrimination among analytes. A method is selective if it distinguishes among the measured compounds easily.
35	Specificity		Specificity is a characteristic of the method. A method is called specific for a compound if it can distinguish it from other compounds and can identify it with full confidence.
36	Split injection		Split injection is an injection technique commonly used in GC. When performing split injection, the injected sample is splitted

(*continues*)

Table 2
Continued

	Term	Acronym	Meaning
			in the injector and only a small portion of the sample enters the chromatographic column.
37	Splitless injection		Splitless injection is an injection technique commonly used in GC. When performing splitless injection, the injected sample is not splitted, but the whole amount is directed into the column. After having the appropriate sample amount on the column, the rest of the sample is flushed from the injector.
38	Staining		Staining is a commonly used visualization technique in 2D gel separations. In a staining process the separated spots are treated with staining reagent making the spots visible.
39	Stationary phase		It is the phase that is considered as static relative to another (mobile) phase. This means practically the sorbent of the columns.
40	Theoretical plate height	HETP, H	Same as height equivalent to a theoretical plate.
41	Theoretical plate number	N	Theoretical plate number is a number that characterizes the separation efficacy of the column. The higher this number the narrower are the peaks in the chromatogram.
42	UV/VIS detection	UV, UV/VIS	A commonly used detection in HPLC that measures the ultraviolet (UV) or visible (VIS) absorbance of the sample.
43	Validation		A procedure designed to estimate the reliability of the results measured by a given method.
44	van Deemter equation		Equation explaining the separation and peak-broadening effects in liquid chromatography. Its simplified form: $H = A + B/u + Cu$.

2. Gas chromatography

GC [5,19–24]—as the name implies—is a separation technique where the applied mobile phase is a gas, while the stationary phase is a solid or a liquid. It has relatively few variants; now nearly exclusive capillary GC is used. The sample is evaporated in an injector and a gas flow carries it through an open capillary tube (column) to the detector. The schematic diagram of a gas chromatograph is shown in Fig. 4.

The heart of the GC system is the capillary column, which is operated inside a special thermostat. The inner wall of this capillary is coated with a liquid stationary phase, which binds and thus retains the components of the sample. Separation is based on the distribution of various compounds between the liquid (stationary) and the gas (mobile) phases. Accordingly, retention of a given compound on the

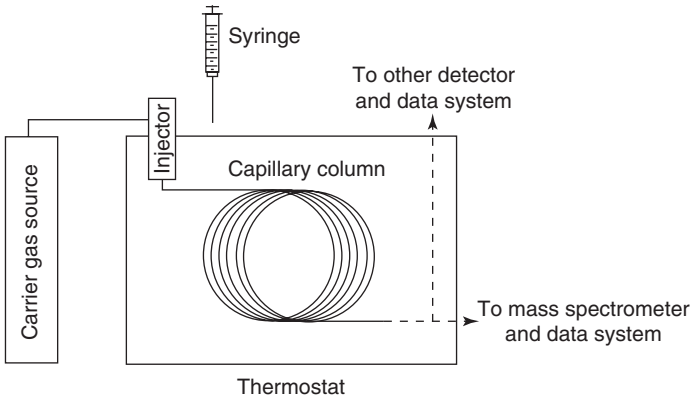


Fig. 4. Schematic representation of a GC system.

column depends strongly on not only the vapor pressure of the analyte but also the coating of the capillary column. Depending on the degree of retention, various components of a mixture will arrive to the detector at different times. The chromatogram is obtained by plotting the signal intensity vs. arrival time. Detection can be performed by a variety of techniques; the most common ones are FID, ECD, and MS. In GC applications relatively few parameters need to be optimized, so method development is usually simpler and quicker than in the case of HPLC. GC and GC–MS have been used in analytical practice for a long time, so a large number of methods are developed, validated, and widely accepted. This means that for many applications there are already well-tested GC methods available. Various GC columns are available; their most important characteristic is the type of the stationary phase. Other features are thickness of the stationary phase (film), inner diameter, and length of the capillary. The majority of today's applications work with 20–60 m long, few hundred micrometer wide fused silica capillary columns coated with a thin (0.01–5 μm) liquid (occasionally 5–50 μm solid) stationary phase. The outside of the capillaries is coated with polyimide to make them flexible. Long, narrow capillaries coated with thin liquid film are the best for high resolution, while wide capillaries coated with thick film have higher sample capacity. A typical column used in analytical practice is 60 m long, has 0.32 mm internal diameter, and is coated with 0.5 μm thick stationary phase.

The most commonly used stationary phases are silicon based, such as the apolar 100% polydimethylsiloxane phase. Addition of 5% diphenyldimethylpolysiloxane makes the phase slightly more polar, favoring analysis for moderately polar compounds. Probably this is the most commonly used stationary phase in GC. To separate compounds of high polarity, polar phases (such as polyethylene glycol which has the trade name Carbowax) give the best results. Several other phases are used in practice, such as hydrocarbon, phtalate, glycol ester, or nitrile-based phases. Note that while these stationary phases are liquid,

they must not evaporate (even at the highest analysis temperature) and must not react with the analyte. The so-called bleeding of a column is a typical problem in GC, which means that the stationary phase slowly evaporates from the surface during analysis, deteriorating column performance and sensitivity. Bleeding effects can be minimized by avoiding water for dilution of the sample and avoiding acidic or basic samples (or such additives). Bleeding might also occur if the column is operated at very high temperature (above 250–300°C). For such applications the use of special heat-resisting columns is needed.

The mobile phase (also called carrier gas) is an inert gas such as helium, argon, nitrogen, or hydrogen; its selection has little influence on the analytical performance. Performance of the GC system can be modeled by various equations, which include the peak-broadening effects. For liquid-coated capillary columns, commonly the Golay equation is used [19]; its simplified form is shown in the following equation:

$$H = \frac{B}{u} + C_M u + C_S u \tag{5}$$

Here H is the theoretical plate height (reflecting the separation power of the system, the smaller the H the better the separation is), u the linear flow rate, B, C_M, and C_S are constants representing the peak broadening effects (the longitudinal diffusion of the analytes in the mobile phase, mass transfer in the mobile phase, and mass transfer in the stationary phase, respectively). This relationship (which is very similar to the van Deemter equation used in HPLC) includes the parameters of a given GC system and describes the effect of the flow rate on the separation power. C_M and C_S are proportional to the square of column diameter, so smaller diameter columns provide smaller (better) plate heights. Film thickness also influences H, which decreases with film thickness. However, thin films have less sample capacity, so the column can get easily overloaded. Note also that GC columns can be characterized by the so-called beta value, $\beta = d_i/(4 \times d_f)$ where d_i is the inner diameter of the column and d_f the thickness of the liquid stationary phase. Columns with $\beta < 100$ are usually suited for analysis of very volatile compounds, columns with $100 < \beta < 400$ are applicable for general purposes, and columns with $\beta > 400$ are suited for the analysis of compounds of high boiling point.

The main limitation of GC is the need to evaporate the sample. This limits the type of compounds that can be studied. Polar, ionic, or thermally labile compounds (such as salts, peptides, etc.) or those with molecular mass above 500 Da can rarely be studied. To extend the range of compounds amenable for GC analysis, derivatization methods have been developed to increase volatility [25–28]. Derivatization makes it possible to use GC–MS for analysis of various small organics in body fluids. Note, however, that derivatization is time-consuming and it is a potential source of artifacts.

The most important parameter of a GC analysis is temperature. It has a profound influence on the vapor pressure of analytes, and therefore on the partitioning

between the liquid and gas phase, and changes retention of compounds to a very large degree. Initially GC analysis was often done at a constant temperature (isothermal); now temperature is commonly changed during analysis (temperature programming). If the temperature is increased, retention time will be shortened; however, if it is decreased the retention time will increase. When retention is reduced then the separation efficacy will also be decreased, so one must always find a trade-off between good resolution and acceptable separation time. Optimization of GC methods is predominantly done by temperature programming, which is an essential feature of modern GC applications. In practice, it means that the GC column is kept isotherm at a given temperature for a certain time, and then the column temperature is raised to X°C (with 5–20°C/min rate) and maintained at that temperature. If needed, the temperature can be further raised in the second (or third) step.

A typical example for a GC program for separating widely different compounds is the following: Start at 60°C with a 5 min long isotherm. Then, increase temperature by 10°C/min rate to 220°C and maintain it for 20 min. If unresolved peaks occur in the chromatogram, the initial temperature may be decreased or the heating rate can be slowed down (to 5°C/min for instance) to increase retention and enhance selectivity. If all peaks are well separated, the initial temperature and/or the final temperature may be increased to speed up the analytical process. To reduce contamination and the possibility of artifacts, columns need to be regularly "conditioned" (after a day's work). This means column temperature should be increased for 2–3 h to at least 20°C higher than the maximum temperature used during the analysis. This ensures that strongly retained contaminations leave the column.

Another critical feature of GC analysis is sample injection. A schematic diagram of a typical GC injector is shown in Fig. 5.

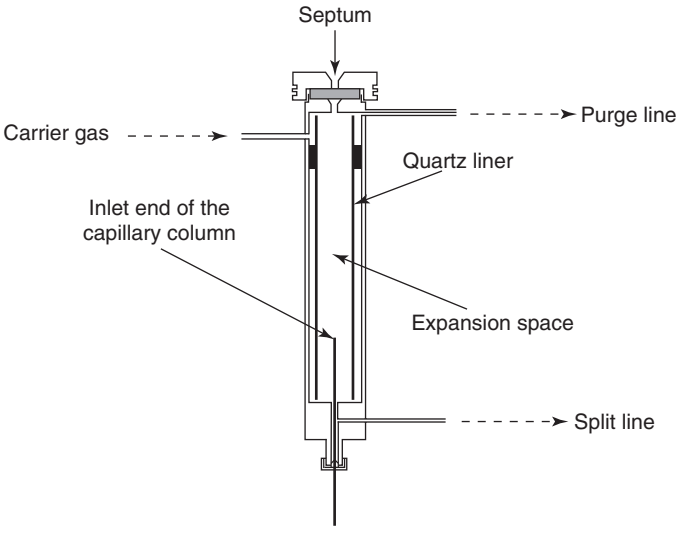


Fig. 5. Schematic diagram a typical GC injector.

The main role of the injector is to evaporate the sample completely and to get it to the front of the column. The temperature of the GC injector is crucial. It must be high enough to allow complete evaporation of the sample (200–350°C), but low enough to minimize thermal degradation. Usually it should be 20–40°C higher than the maximum column temperature used. The injector is connected to the carrier gas source and to the inlet end of the capillary column. It also has a third connection to a waste line, which serves for purging the injector. All measurements begin with injection of the liquid (occasionally gas) sample into the expansion space of the injector. First, the sample is drawn into the syringe and the septum of the injector is pierced with the sharp needle of the syringe. The sample is then injected into the expansion space of the injector and the syringe is withdrawn. The speed and accuracy of this process have key importance in GC. Note that there is pressure inside the injector, so one must always keep safe the plunger of the syringe from being expelled. The injector is kept at a high temperature in order to aid quick evaporation of the sample. After evaporation the sample is carried by the carrier gas onto the capillary column. Capillary columns have low sample capacity, while the minimal sample volume that can be reliably measured is around 1 μl. This sample amount would typically overload the column. Two basic approaches are used in GC to overcome this problem. One is the so-called split injection [29]. In this case only a small fraction of the evaporated sample is carried onto the column, most of it is carried out into the waste by the split line.

Another technique is the so-called splitless or combined split/splitless injection [29]. In this case the split line is first closed. After the desired amount of sample is loaded onto the column, the split line opens (typically 10–100 s after injection and evaporation of the sample) and the remaining (superfluous) sample is purged from the injector.

Special injectors may also be used, such as the temperature-programmed injector [29]. The main goal of this case is to use the injector to enrich analytes prior to carrying them onto the column. In this case a cold (−40 to +40°C) injector is used. An inert gas flow purges the low-boiling-point solvent from the expansion space while the high-boiling-point analytes remain condensed inside the injector. After removal of the solvent, the purge line is closed and the injector is rapidly heated. At this phase target analytes evaporate and are carried onto the column. Using this technique large sample amounts can be injected into the system and the signal-to-noise ratio can be improved.

A practical difficulty in GC is that the nonvolatile fraction of any sample (common in biological matrices) remains precipitated on the wall of the injector or in the beginning of the column. These may accumulate, decompose, and may even catalyze decomposition of analytes, deteriorating the performance of the analysis. To avoid these problems, the injector needs to be cleaned frequently, and the front end of the capillary column (0.5 m or so) needs to be cut off occasionally and discarded.

The output end of the GC column is connected to a detector such as FID, ECD, or mass spectrometer. Gas chromatography–mass spectrometry (GC–MS) is a very efficient analytical tool and also quite straightforward to use. To indicate simplicity in some GC–MS systems the mass spectrometer is also termed "mass selective detector." Coupling GC with MS usually does not deteriorate GC separation; the gas effluent of the GC can be introduced directly into the ion source of the mass spectrometer.

In summary, the most important parameters to keep in mind while planning GC experiments are the temperature program of the separation, the stationary phase of the column, the temperature of the injector, and the properly selected detection system. Typical applications of GC in the biomedical field include the determination of low-molecular-weight compounds in body fluids, such as amino acid and fatty acid profiling of blood, or organic acid profiling of urine.

3. High-performance liquid chromatography

LC is a separation technique where the applied mobile phase is a liquid, while the stationary phase may be either solid or liquid. The technique is used mainly to separate nonvolatile compounds. In its original version, a fairly large (about 1 m long, few centimeters wide) vertical column is packed with the stationary phase. The solution is introduced onto the top of the column and gravitation forces the liquid to pass through the column. This version (also called column chromatography) is often used for the separation of relatively large quantity of compounds (in the range of 100 mg). A modified version of column chromatography is flash chromatography, where the liquid flow through the column is assisted by a vacuum manifold or a vacuum pump. For analytical purposes, column and flash chromatography are not considered efficient and are superseded by HPLC.

In the case of HPLC [2,30,31] the liquid sample is driven through a packed tube (column) by liquid flow at high pressure (typically 50–400 bar) provided by mechanical pumps. Various components of the sample reach the end of the column at different times and are detected most often by UV–VIS spectrometry (which measures the absorbance of the effluent in the wavelength range ~200–600 nm) or by MS. The chromatogram is obtained by plotting the signal intensity vs. time. The sequence of the components reaching the detector strongly depends on the molecular structure of the analytes, the composition of the mobile phase, and column packing (stationary phase).

The majority of current applications use stationary phases [32] made of porous silica, aluminum oxide, or polymer particles. Solid-phase particles need to have small particle size (3–5 μm are commonly used) and a well-defined pore diameter. The most commonly used silica phases have good mechanical stability (i.e., can be used at least up to 400 bar pressure) but have low pH tolerance (so can be used only

between pH 2 and 7–8) [33]. Polymer-based phases can be used up to pH 12, but these have a lower mechanical stability and can be used only up to 50–100 bar pressure. Metal-based stationary phases (zirconium, aluminum) overcome both of these limitations, i.e., they have both good mechanical stability and are stable in a wide pH (1–14) range and up to 200°C temperature, but they exhibit undesired electrostatic interactions which may complicate development of a separation method.

Mobile phases can be selected from a wide range of solvents including water, methanol, acetonitrile, isopropanol, acetone, *n*-hexane, etc. The main parameters for selecting the mobile phase are the following: polarity (which defines the eluent strength, see the following text), miscibility, low viscosity, high boiling point, low UV light absorbance (if used with UV detection), and low toxicity.

In HPLC, the sample is dissolved in a solvent (preferably same as the HPLC mobile phase) and injected onto the column. Attention must be paid to avoid precipitation of the injected sample and blockage of the column. The HPLC column is usually a 3–25 cm long metal tube of 1–5 mm diameter. Conventionally 4.6 mm columns are used in HPLC, with a flow rate of about 1 ml/min. Nowadays narrower columns (1 and 2 mm) are becoming very popular, especially combined with MS (using much less, 50–200 μl/min solvent flow). Micro- and nano-HPLC is also gaining ground (e.g., using 75 μm diameter quartz tubes and ~200 nl/min solvent rate), especially in the field of proteomics [34]. Note that narrow columns require very small amounts of sample (approximately proportional to the internal volume of the column), and thus require very sensitive detectors.

HPLC columns are packed with the stationary phase, which retains the sample molecules. Retention of compounds depends on not only various factors predominant on molecular properties but also particle size, pore size, homogeneity of the stationary phase, viscosity and polarity of the mobile phase, etc. These effects are summarized in the van Deemter equation [14] (Equation (6), analogous to the Golay equation used in GC), which describes peak broadening in LC:

$$H = A + \frac{B}{u} + Cu \tag{6}$$

Here H is the theoretical plate height, a parameter that characterizes the effectiveness of the chromatographic separation. The smaller the H the more powerful is the separation. A is the Eddy diffusion term (or multipath term), B relates to longitudinal diffusion, C represents the resistance of sorption processes (or kinetic term), and u is the linear flow rate. For a given chromatographic system, A, B, and C are constants, so the relationship between H and u can be plotted as shown in Fig. 6.

The theoretical plate height curve has a minimum that corresponds to the optimal flow rate. The minimal theoretical plate height is influenced by the average particle size of the stationary phase. The smaller the average particle size the smaller the H and the better the resolution is [35,36]. Current technologies can provide columns

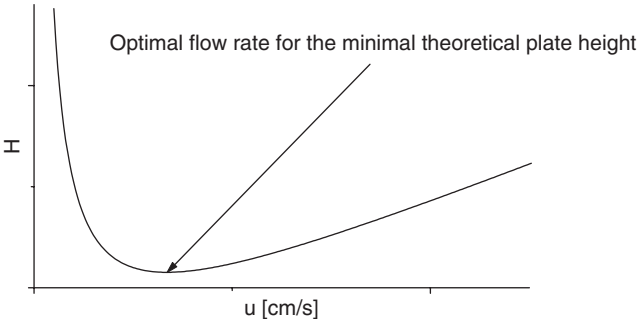


Fig. 6. Schematic representation of a van Deemter curve.

with a particle size down to 1.5 μm in diameter. High porosity (which correlates with high surface area and small pore diameter) of the particles is also important to achieve maximum selectivity. However, if the pore size is too small then target molecules cannot penetrate into the pores, deteriorating column performance. This is especially important if macromolecules are studied. Optimal pore size is therefore a compromise; usually the 10–100 nm range is considered best.

The third term of the van Deemter equation includes diffusion of analyte molecules into and out of the pores of the particles. To obtain minimal peak broadening, diffusion needs to be fast. The diffusion rate correlates with the viscosity; thus in HPLC low-viscosity solvents are preferred. Diffusion also depends on temperature; thus, maintaining the column at high temperature by using a column thermostat is often advantageous. However, it must be kept in mind that the temperature range of stationary phases is limited and that at high temperature longitudinal diffusion may become dominant, also leading to peak broadening. At high temperature, the vapor pressure of solvents increases further limiting the usable temperature range. In practice, column temperature up to 50–60°C is used.

Solvent strength is defined as the capability of the solvent to elute a given compound from the stationary phase. The stronger the solvent the quicker it can elute the analyte from the column. Elution of analytes in HPLC can be performed by two basic approaches, namely by isocratic or gradient elution. In the case of isocratic elution the same solvent mixture is used during elution, while in gradient elution the composition of the mobile phase is systematically changed, so that the solvent strength is increased. Isocratic and gradient elution techniques in HPLC are analogous to isotherm and temperature programming methods in GC. Isocratic elution has the advantage of simplicity and it is better suited for high-throughput applications. Its main disadvantage is that it cannot cope with widely different analytes. If weakly and strongly binding analytes are studied together then either the resolution will be unacceptably low at the beginning of the chromatogram, or strongly binding components will not be eluted. Gradient elution is the method of choice for separating widely different compounds. It requires better quality and more

expensive instrumentation than isocratic elution and in this case the column must be equilibrated with the initial solvent composition after each analysis (before a new sample can be injected), which increases analysis time. The main advantage of gradient elution is that at the beginning of the experiment a low solvent strength mobile phase is used; thus, analytes that bind weakly to the column can be resolved. Subsequently the solvent strength is increased gradually and compounds that bind strongly to the column are resolved and eluted too. Gradient elution therefore can deal with complex mixtures; the shape of the gradient can be optimized to achieve good resolution of all compounds while maintaining acceptable analysis time [37]. Today, gradient elution techniques are more and more widespread and they are becoming indispensable to deal with complex mixtures such as biological extracts.

Several detector types can be used for HPLC, such as UV–VIS absorbance detection (either operating at a given wavelength or using a diode array to detect the whole spectrum), fluorimetric detection, refractive index detection, evaporative light scattering detection, or mass spectrometric detection (HPLC–MS). Among these methods mass spectrometric detection is probably most selective and is rapidly gaining ground. There are several MS techniques compatible with HPLC, such as ESI, nanospray, and APCI.

In addition to the solvent, additives are often used in HPLC in low amounts (0.01–1%) to optimize performance and minimize undesired side effects, such as peak broadening. One of the prime factors determining retention is the charge state of the analyte that strongly depends on the pH. For this reason buffers (traditionally potassium phosphate buffers) are typically used to adjust the pH accurately. Note that in the case of HPLC–MS, nonvolatiles cannot be used, so typically ammonium acetate or formate buffer is preferred. Various other additives may also be used, such as trimethyl amine or trifluoroacetic acid, to suppress the interaction of analytes with the residual silanol groups of the stationary phase, thereby improving the resolution.

In summary, the most important parameters in designing and optimizing HPLC are the solvent system and the gradient program. The type of the applied column (packing type and particle size of the stationary phase, length, and diameter of the column), the flow rate, and column temperature are also significant. Overall performance of HPLC depends on several factors and cannot be optimized by considering one parameter only. Below the most common versions, normal and reverse-phase HPLC, ion exchange, and size exclusion chromatography are discussed.

3.1. Normal-phase liquid chromatography

Normal-phase liquid chromatography (NP-HPLC), as the name implies, is the original version of HPLC. Nowadays it is not often used, only when results obtained with reverse-phase LC prove unsatisfactory. It is discussed first for didactic reasons.

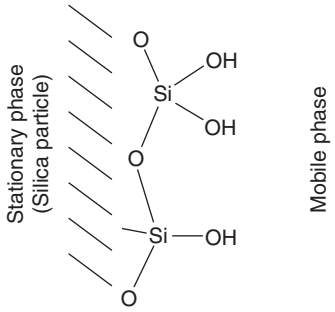


Fig. 7. Molecular properties of a typical NP stationary phase.

In NP-HPLC the stationary phase is more polar than the mobile phase and the interaction between analyte and column has predominantly polar character (hydrogen bonding, π–π or dipole–dipole interactions, etc.). The most commonly used NP stationary phase is silica gel ($[SiO_2]_x \cdot [H_2O]_y$). After column preparation the surface of silica gel consists mainly of hydroxyl groups bound to silica atoms as shown in Fig. 7.

These hydroxyl groups are often called silanol groups. These predominantly bind analytes by polar interactions. Other stationary phases are also used, such as aluminum oxide or chemically modified silica gel. In the latter case usually amino, diol, nitro, or cyano group containing chemicals are reacted with the free silanol groups to modify their binding properties.

Mobile phases in NP-HPLC are mostly apolar solvents (or solvent mixtures) such as *n*-hexane, *n*-heptane, dichloromethane, dichloroethane, diethyl ether, methyl acetate, ethyl acetate, acetone, isopropanol, ethanol, or methanol. In NP-HLPC more polar solvents represent higher solvent strength and these elute compounds faster from the column. The typical order of solvent strength is hydrocarbons $<$ ethers $<$ esters $<$ alcohols $<$ acids $<$ amines (going from weak to strong).

The biggest problem in using NP-HPLC is its dramatic sensitivity to water. Even water traces (in the mobile phase or from the sample) may bind to the column, deteriorate its performance, and cause irreproducibility. In addition, particular care must be taken to ensure accurate pH, as in NP-HPLC, retention is very sensitive to the charge state of the analyte. Owing to these practical problems NP-HPLC is relatively rarely used. Its main application fields are separation of polyaromatic hydrocarbons, sterols, vitamins, chlorophylls, ceramides, and other lipid extracts.

3.2. Reverse-phase liquid chromatography

Reverse-phase liquid chromatography (RP-HPLC) is the most important and most widely applied version of LC. It is well suited to separate both apolar and polar compounds, but less well suited for studying permanently ionized molecules. It is easy to couple with MS.

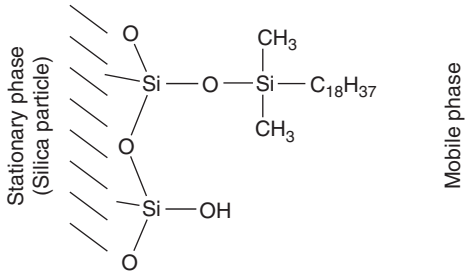


Fig. 8. Molecular properties of a typical RP stationary phase.

In RP-HPLC the stationary phase is less polar than the mobile phase and the interaction between analyte and the stationary phase has a predominantly hydrophobic (apolar) character. The most commonly used stationary phase in RP-HPLC is silica gel in which octadecyl silica chains are covalently bound to the free hydroxyl groups, indicated as a C18 phase. The typical surface of such a phase is shown in Fig. 8.

Other commonly used stationary phases are silica gels modified using octyl (indicated, e.g., as a C8 phase), hexyl, butyl, or ethyl groups. Occasionally organic polymer-based phases are also used. Modified silica gels may be used up to several hundred bars pressure and across a pH range of 2–8.5. Care must be taken to select the right pH, as the chemically bound groups begin to hydrolyze at pH below 2 and the silica gel begins to dissolve at pH higher than 8–9 [38]. Retention of compounds occurs by apolar interaction between the analyte and the immobilized octadecyl silica chain. Most compounds exhibit hydrophobic character to some extent and thus they can be analyzed by RP-HPLC. Even strongly polar or ionic substances can be analyzed by RP-HPLC if the pH is adjusted so that the analyte will be in neutral form. Such an example is RP-HPLC separation of basic amphetamines at pH 8.5 [39].

The surface of C18 phases always contains unreacted silanol groups, which may form secondary polar interactions with the analyte. This is generally disadvantageous in RP-HPLC as it often causes peak broadening [33,40]. An important improvement is the introduction of the so-called end-capping procedure: The residual silanol groups in the C18 phase are reacted with monofunctional chlorosilane, which decreases surface polarity. This very popular stationary phase is called C18ec, where the notation “ec” stands for end-capped.

Mobile phases in RP-HPLC are mostly polar solvents such as water, acetonitrile, methanol, and isopropanol. In RP-HPLC apolar solvents have high solvent strength. Accordingly, the order of solvent strength is water $<$ acetonitrile $<$ ethanol $<$ acetone (from weak to strong). The most commonly used solvent mixture is a water–acetonitrile gradient, in which the amount of acetonitrile is increased during a chromatographic run to elute first the polar components and then the more strongly bound apolar compounds. Mixtures containing a wide

range of compounds may be studied by a fast gradient starting from high water content (e.g., 90%) and ending at high (usually 100%) acetonitrile content.

RP-HPLC is widely applicable, although pH control must often be applied. Most important application areas include peptide and protein analysis (proteomics), drugs and their metabolites, fatty acids, and also volatile compounds such as aldehydes and ketones, although these require derivatization.

3.3. Ion-exchange liquid chromatography

Ion-exchange liquid chromatography (IE-LC) is not very common, but it is gaining importance [41–43]. It separates ionized compounds, which excellently complements RP-HPLC. In ion-exchange chromatography separation of different compounds is achieved by using ion–ion interactions between the analyte and the stationary phase. To ensure that this interaction is dominant, the surface of the stationary phase must contain either permanently or temporarily ionized groups and of course the sample must be in ionized form. Most commonly used stationary phases in IE-LC are chemically modified silica gels containing immobilized anionic or cationic groups. These groups are most commonly primary, secondary, quaternary amine, and carboxyl or sulfonyl groups. The retention of acidic compounds occurs with anion-exchange phases (immobilized amines), while the retention of basic compounds occurs with cation-exchange phases (immobilized acids). In performing IE-LC particular care must be paid to ensure the adequate pH, as retention is very sensitive to the charge state of the analyte. The surface of a typical anion-exchange stationary phase is shown in Fig. 9.

Here the surface of silica gel is modified by the introduction of quaternary amine groups. These groups are permanently positively charged; thus, they attract negatively charged analytes (anions).

The applied mobile phases in IE-LC are mostly solvents with acid, base, or buffer content. The strength of the mobile phase can be influenced either by changing the pH to shift the ionization state of the analyte or by displacing the analyte with solvent additives (e.g., displace a fatty acid from the cationic stationary phase by adding 1–2% of acetic acid to the mobile phase).

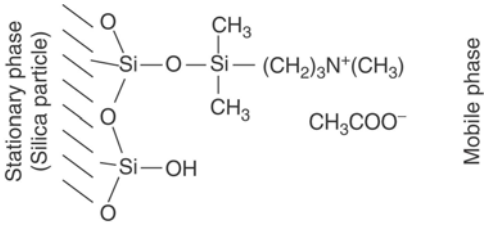


Fig. 9. Molecular properties of a typical anion-exchange stationary phase.

IE-LC is suitable for the separation of either permanently or temporarily ionized compounds. Typical application areas are separation of amino acids (e.g., amino acid analyzers) or separation of enzymatically digested protein fragments prior to reverse-phase separation. The complementary features of IE and RP chromatography can be excellently utilized in 2D chromatography (see the following text), which is gaining importance for protein analysis.

3.4. Size exclusion chromatography (gel filtration)

Size exclusion chromatography is a method where separation of different compounds occurs according to their size (hydrodynamic volume) measured by how efficiently they penetrate the pores of the stationary phase [44,45]. Size exclusion chromatography has two basic versions. When performed using organic solvents, it is called gel permeation chromatography (GPC). The main application field of GPC is polymer analysis. When size exclusion chromatography is performed using aqueous solvents, it is called gel filtration. A typical example of gel filtration is desalting of proteins. In this case the protein–salt mixture is applied onto the column. The inorganic salt ions have small size; they penetrate the small pores present in the stationary phase and therefore will be retained on the column. In contrast, large protein molecules cannot enter the very small pores and so will be eluted by the solvent flow with minimal retention. As a consequence, the proteins will first elute from the column while the salt will be retained.

3.5. 2D liquid chromatography

In 2D chromatography [46], two different chromatographic columns are connected in sequence, and the effluent from the first system is transferred onto the second column. Application of 2D LC is suggested when very complex mixtures have to be separated. In a typical HPLC experiment, the average peak width is 30 s while the chromatogram is about 1 h long, so at most 120 compounds can be separated. This peak capacity can be substantially improved when the effluent of the first column is collected in fractions and is further analyzed by a separate chromatographic run, usually based on a different separation mechanism. This can be implemented in both offline and online modes. A typical online experiment for 2D HPLC is used for proteomics applications [46], where a complex mixture of digested proteins has to be analyzed (often thousands of peptides are present in the sample). The digested sample is first injected onto a cation-exchange column, as the commonly used trypsin yields basic peptides. First, the neutral peptides elute from the column, and these are washed onto the next, very short octadecyl silica column. This column binds (and therefore concentrates) the first fraction of peptides. After changing the solvent composition (switching to a different solvent mixture) the peptide fraction is washed onto a longer, analytical octadecyl silica column, where the peptides are separated on the basis of their polarity (a typical RP-HPLC application). In the next step the

cation-exchange column is washed with an eluent containing low salt concentration, which elutes the weakly retained peptides. These are trapped, washed, and analyzed on the octadecyl silica column similarly to the first fraction. In repeated steps the cation-exchange column is washed with eluents of higher and higher salt content and thus peptides with higher and higher basicity are eluted from the column. These fractions are trapped and analyzed on the C18 column as described earlier. In summary, the peptides are fractioned according to their basicity on the first column (first dimension) and the obtained fractions are further separated on the basis of their apolar character on the second column (second dimension). This protocol reduces coelution and thus enhances the confidence of identification for unknown proteins.

4. Electrophoretic techniques

Electrophoretic techniques are well suited to separate charged compounds. Separation is due to migration induced by high voltage and takes place either in a buffer solution or in the pores of a gel filled with buffer solution. Several electrophoretic techniques are used; here only the most important ones will be discussed. Most of these methods are used for analysis, but some (such as 2D gels) also for isolating macromolecules for further studies. Electrophoretic techniques are particularly important for studying macromolecules, especially proteins.

4.1. Capillary zone electrophoresis

Capillary zone electrophoresis [9,10,47–54] (CZE) is a separation technique where components of the sample are separated using 10–30 kV potential difference between the two ends of a 50–100 μm diameter capillary filled with a buffer solution. The basic instrumental setup is demonstrated in Fig. 10.

The capillary column is immersed into two buffer-filled reservoirs. High voltage is applied to these reservoirs via platinum electrodes. The sample is stored in

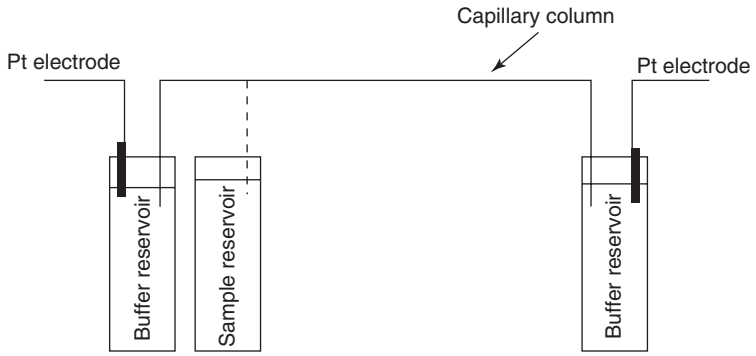


Fig. 10. Schematic representation of a capillary electrophoresis system.

a separate reservoir and can be injected into the capillary by various techniques such as a hydrodynamic or electrokinetic impulse. The injected sample volume is in the low nanoliter range.

Separation of components occurs by the simultaneous effect of the electrophoretic and electro-osmotic forces that develop inside the capillary. Electrophoretic force (and flow) is a result of the applied potential difference (high voltage) between two ends of the capillary. It attracts the positively charged ions towards the cathode (negative end) and negatively charged ions towards the anode (positive end). Electro-osmotic force is a result of the electrical double layer, which develops on the wall of the capillary and induces a flow by its motion towards the cathode. CZE provides unusually high resolution since several peak-broadening effects present in traditional HPLC are absent. The only significant peak-broadening effect in CZE is longitudinal diffusion along the column. Resolution is determined by the applied high voltage and the electrophoretic mobility of the ions. The applied flow rates in CZE are in the nanoliter range; thus, this separation technique can be coupled with nanospray MS.

4.2. Capillary gel electrophoresis

Capillary gel electrophoresis [55,56] (CGE) is very similar to CZE. The main difference is that in CGE the column is packed with a gel, which affects the motion of the analytes. Accordingly, separation will be determined not only by the electrophoretic force acting on the ions but also by the size of analyte molecules. The effect of the gel present inside the column has a similar effect to size exclusion chromatography (see earlier). A typical application is the separation of proteins in a capillary which is filled with polyacrylamide gel and sodium dodecyl sulfate (SDS). The presence of SDS aids the electrophoretic mobility of proteins, as it coats their surface proportional to their size. Consequently, the molecular structure will have little influence on mobility, so macromolecules will migrate according to their molecular mass. This technique is very similar to SDS-PAGE.

4.3. Capillary isoelectric focusing

Capillary isoelectric [54,57,58] focusing is closely related to the techniques discussed above, but separates compounds based on their isoelectric point. Separation occurs in a capillary, which is internally polymer-coated to eliminate the electro-osmotic flow. The cathode end of the capillary column is immersed into a base and the anode end into an acid. This results in the formation of a pH gradient along the column. Similarly to capillary electrophoresis, positively charged ions migrate towards the cathode and negatively charged ions migrate towards the anode. The predominant effect in this case is the pH dependence of the charge state of the

analyte. A positively charged compound will migrate in the capillary column towards the cathode, but during this migration it arrives into an increasingly basic environment. At the position where the pH is equal to the isoelectric point of the compound, the net charge on the analyte becomes zero (this is, in fact, the definition of the isoelectric point) and the compound stops migrating. This way each compound is concentrated at the pH value that is the same as its isoelectric point. The separated zones can be displaced from the capillary by either a hydrodynamic or an electrokinetic impulse and measured as a chromatogram.

4.4. Sodium dodecyl sulfate polyacrylamide gel electrophoresis

An especially important technique in proteomics is sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) [59,60]. This method can separate high-molecular-mass proteins or glycoproteins (up to several hundred kilodaltons). Just like in other electrophoretic methods, the separation of the analytes occurs by migration induced by a high potential difference. In this case an anionic detergent (SDS) is used to aid solubility, denaturation, and charging of proteins. SDS wraps around the peptide backbone of proteins and confers multiple negative charges to the protein. The amount of bound SDS is proportional to the size of the protein; thus, the net charge and therefore the migration of the protein will be proportional to its size and molecular weight. Separation is often implemented on a vertically positioned gel strip, so the separated proteins form horizontal bands on the gel. To visually observe the protein bands the gel is stained [61] typically using Coomassie blue or copper chloride chemicals. After the separation the individual bands in the gel are often cut out for further analysis (usually by MS), i.e., the SDS-PAGE can be used for small-scale preparative purposes as well.

4.5. 2D gel electrophoresis

2D gel electrophoresis [59,60,62–65] is performed on a plate by the combination of isoelectric focusing in 1D, and SDS gel electrophoresis in the other direction. The separation is based on two different, unrelated (orthogonal) phenomena, and provides exceptionally high resolution. The separated compounds (mostly proteins) form spots on the plate, which may be cut out for further studies. Although it requires very careful work and large experience, 2D gels are very powerful and are capable of resolving over a 1000 spots in a plate. 2D gel electrophoresis has become one of the most important and widely used techniques in the field of proteomics.

In practice, 2D gel electrophoresis is implemented mostly on porous agarose or polyacrylamide gel in the form of a homogenous, flat, square-shaped layer. In a 2D gel experiment, the sample is first separated by applying isoelectric focusing on a strip (first dimension). Then, this strip is attached to a gel plate and further

separation is performed by SDS gel electrophoresis (second dimension). Although separation by electrophoretic mobility in a gel is very similar to capillary gel electrophoresis, isoelectric focusing in a gel is somewhat more complex. Isoelectric focusing in a flat gel is achieved by applying pH gradient along one edge of the surface. The pH gradient can be formed either by applying the so-called ampholyte chemicals or by using a gel that consists of prefabricated gel strips with immobilized buffer on the surface. To reduce secondary effects (hydrogen bonds and hydrophobic interactions between the analytes and the gel) additional chemicals (such as urea or thiourea) are often used. The separated compounds are stained for visualization using various methods such as Coomassie blue staining or copper chloride chemicals. Developed 2D plates are typically scanned and then analyzed by advanced computerized techniques, identifying those spots that change (are overexpressed or underexpressed) between two different samples. These spots are typically cut out (manually or robotically) and the respective proteins are identified by MS.

5. Future trends

The most important role of chromatographic techniques is that they ensure the necessary selectivity and chemical purity prior to detection. Although there have been dramatic improvements in detection systems, it is still very important to boost the performance of separation methods. Better separation often means lower detection limits, better quantitation, and more confident identification of unknowns. As most compounds in biological systems are nonvolatile, LC-based techniques dominate over GC in the biomedical field.

We see three major aspects for future developments in chromatography. One relates to improving analytical performance: to lower detection limits, to increase selectivity, to be able to analyze less sample, etc. To achieve this, manufacturers are constantly modifying their instruments, new chromatographic columns become available, and especially pure solvents are used. For example, ultra performance liquid chromatography (UPLC) [35,36] uses very small particle size (approximately 1.5 μm). This results in narrower peaks but needs unusually high pressure (up to $\sim$1000 bar). Significant improvements are emerging in the field of GC as well. Application of time-of-flight MS for detection provides acquisition rates at hundreds of spectra per second, which opens up new possibilities for ultra-fast GC. Using this technique chromatograms take only a few minutes, and peak widths are less than a second [66,67]. Such improvements will appear in the future, but a major breakthrough is not expected. A different option to improve performance is the online combination of techniques—this is capable of achieving stunning results. To enhance selectivity different chromatographic

techniques are often combined in order to separate those compounds that would coelute by applying merely one technique. These multidimensional techniques are already often used [46], but they will spread even more as they become available as standard components of HPLC systems. Simple HPLC–MS combinations are already considered routine. More complex combinations, such as 2D HPLC combined with high-resolution tandem MS, are also likely to become more common.

The second trend to watch is miniaturization. This is advantageous not only because sample amount is often limited but also because performance may be improved, and running costs can be reduced (e.g., by using less chemicals). Small size also means that more equipment can be put into the (often-limited) laboratory space. The main limitation of miniaturization is sensitivity. In this respect MS, due to its sensitivity, is also invaluable. Luckily an important mass spectrometric technique, nanospray ionization, is ideally suited for coupling to nano-HPLC [34] (requires nl/min flow rates). This reduces sample requirement, and also facilitates coupling MS with electrophoretic techniques.

The third and possibly the most important trend is high throughput and automatization/robotization. The prerequisite is very robust methodology, which is becoming available. Most high-quality instruments are capable of automatic operation; this will become increasingly widespread. This can reduce labor costs and may make individual-based medication and population-wide medical screening possible.

References

1. Armstrong, D.W. and Boehm, R.E., Gradient Lc Separation of Macromolecules—Theory and Mechanism. *J. Chromatogr. Sci.*, **22**, 378–385 (1984).
2. Bird, I.M., High-Performance Liquid-Chromatography—Principles and Clinical-Applications. *Br. Med. J.*, **299**, 783–787 (1989).
3. Celis, J.E. and Gromov, P., 2D protein electrophoresis: can it be perfected? *Curr. Opin. Biotechnol.*, **10**, 16–21 (1999).
4. Dolan, J.W., Temperature selectivity in reversed-phase high performance liquid chromatography. *J. Chromatogr. A*, **965**, 195–205 (2002).
5. Eiceman, G.A., Hill, H.H. Jr. and Gardea-Torresdey, J., Gas chromatography. *Anal. Chem.*, **70**, 321R–339R (1998).
6. Ettre, L.S., Headspace-gas chromatography—An ideal technique for sampling volatiles present in non-volatile matrices, *in* Headspace Analysis of Foods and Flavors. *Adv. Exp. Med. Biol.*, **488**, 9–32 (2001).
7. Fey, S.J. and Larsen, P.M., 2D or not 2D. *Curr. Opin. Chem. Biol.*, **5**, 26–33 (2001).
8. Jiskra, J., Claessens, H.A., Cramers, C.A. and Kaliszan, R., Quantitative structure-retention relationships in comparative studies of behavior of stationary phases under high-performance liquid chromatography and capillary electrochromatography conditions. *J. Chromatogr. A*, **977**, 193–206 (2002).
9. Robert, F., Bouilloux, J.P. and Denoroy, L., Capillary Electrophoresis—Theory and Applications. *Ann. Biol. Clin. (Paris)*, **49**, 137–148 (1991).

10. Tagliaro, F., Turrina, S. and Smith, F.P., Capillary electrophoresis: Principles and applications in illicit drug analysis. *Forensic Sci. Int.*, **77**, 211–229 (1996).
11. Barth, H.G., Boyes, B.E. and Jackson, C., Size-Exclusion Chromatography. *Anal. Chem.*, **66**, 595R–620R (1994).
12. Gotmar, G., Fornstedt, T. and Guiochon, G., Peak tailing and mass transfer kinetics in linear chromatography—Dependence on the column length and the linear velocity of the mobile phase. *J. Chromatogr. A*, **831**, 17–35 (1999).
13. Papai, Z. and Pap, T.L., Analysis of peak asymmetry in chromatography. *J. Chromatogr. A*, **953**, 31–38 (2002).
14. Piatkowski, W., Gritti, F., Kaczmarski, K. and Guiochon, G., Influence of the particle porosity on chromatographic band profiles. *J. Chromatogr. A*, **989**, 207–219 (2003).
15. McCalley, D.V., Comparison of peak shapes obtained with volatile (mass spectrometry-compatible) buffers and conventional buffers in reversed-phase high-performance liquid chromatography of bases on particulate and monolithic columns. *J. Chromatogr. A*, **987**, 17–28 (2003).
16. Krouse, S.O., Analytical method validation for biopharmaceuticals—A practical guide. *Biopharm. Int.*, **I**, 26 (2005).
17. Swartz, M. and Krull, I., Analytical method validation: Accuracy in quantitation. *LC–GC*, **23**, 46 (2005).
18. Taverniers, I., De Loose, M. and Van Bockstaele, E., Trends in quality in the analytical laboratory. II. Analytical method validation and quality assurance. *Trends Anal. Chem.*, **23**, 535–552 (2004).
19. Golay, M.J.E., In: Desty, D.H. (Ed.), *Gas Chromatography*. Butterworths, London, 1958.
20. Ettre, L.S. and Hinshaw, J.V., *Basic Relationships of Gas Chromatography*. Advanstar, Cleveland, 1993.
21. Karasek, F.W. and Clement, R.E., *Basic Gas Chromatography—Mass Spectrometry Principles and Techniques*. Elsevier, Amsterdam, 1988.
22. Bartle, K.D. and Myers, P., History of gas chromatography. *Trends Anal. Chem.*, **21**, 547–557 (2002).
23. Grant, D.W., *Capillary Gas Chromatography*. Wiley, Chichester, 1996.
24. Eiceman, G.A., Gardea-Torresdey, J., Overton, E., Carney, K. and Dorman, F., Gas chromatography. *Anal. Chem.*,**76**, 3387–3394 (2004).
25. Zaikin, V.G. and Halket, J.M., Derivatization in mass spectrometry—2. Acylation. *Eur. J. Mass Spectrom.*, **9**, 421–434 (2003).
26. Halket, J.M. and Zaikin, V.G., Derivatization in mass spectrometry—1. Silylation. *Eur. J. Mass Spectrom.*, **9**, 1–21 (2003).
27. Halket, J.M. and Zaikin, V.G., Derivatization in mass spectrometry—3. Alkylation (arylation). *Eur. J. Mass Spectrom.*, **10**, 1–19 (2004).
28. Halket, J.M. and Zaikin, V.G., Derivatization in mass spectrometry—5. Specific derivatization of monofunctional compounds. *Eur. J. Mass Spectrom.*, **11**, 127–160 (2005).
29. Bailey, R., Injectors for capillary gas chromatography and their application to environmental analysis. *J. Environ. Monit.*, **7**, 1054–1058 (2005).
30. Lindsay, S., *High Performance Liquid Chromatography*. Wiley, Chichester, 1992.
31. Katz, E.D., *High Performance Liquid Chromatography Principles and Methods in Biotechnology*. Wiley, Chichester, 1996.
32. Majors, R.E., A review of HPLC column packing technology. *Am. Lab.*, **35**, 46 (2003).
33. Espada, A. and Rivera-Sagredo, A., Ammonium hydrogencarbonate, an excellent buffer for the analysis of basic drugs by liquid chromatography-mass spectrometry at high pH. *J. Chromatogr. A*, **987**, 211–220 (2003).

34. Ishihama, Y., Proteomic LC-MS systems using nanoscale liquid chromatography with tandem mass spectrometry. *J. Chromatogr. A*, **1067**, 73–83 (2005).
35. Swartz, M.E., Ultra performance liquid chromatography (UPLC): An introduction. *LC–GC*, **23**, 8–14 (2005).
36. Swartz, M.E., UPLC (TM): An introduction and review. *J. Liq. Chromatogr. Rel. Technol.*, **28**, 1253–1263 (2005).
37. Snyder, L.R., *Chromatogr. Rev.*, **7**, 1 (1965).
38. Claessens, H.A. and van Straten, M.A., Review on the chemical and thermal stability of stationary phases for reversed-phase liquid chromatography. *J Chromatogr. A*, **1060**, 23–41 (2004).
39. Pichini, S., Pacifici, R., Pellegrini, M., Marchei, E., Lozano, J., Murillo, J., Vall, O. and Garcia-Algar, O., Development and validation of a high-performance liquid chromatography—Mass spectrometry assay for determination of amphetamine, methamphetamine, and methylenedioxy derivatives in meconium. *Anal. Chem.*, **76**, 2124–2132 (2004).
40. Nawrocki, J., The silanol group and its role in liquid chromatography. *J. Chromatogr. A*, **779**, 29–71 (1997).
41. Crimmins, D.L., Applications of strong cation-exchange (SCX)-HPLC in synthetic peptide analysis. *Methods Mol. Biol.*, **36**, 53–64 (1994).
42. Bahr-Lindstrom, H., Moberg, U., Sjodahl, J. and Jornvall, H., Ion-Exchange High-Performance Liquid-Chromatography Steps in Peptide Purifications. *Biosci. Rep.*, **2**, 803–811 (1982).
43. Edelson, E.H., Lawless, J.G., Wehr, C.T. and Abbott, S.R., Ion-Exchange Separation of Nucleic-Acid Constituents by High-Performance Liquid-Chromatography. *J. Chromatogr.*, **174**, 409–419 (1979).
44. Kostanski, L.K., Keller, D.M. and Hamielec, A.E., Size-exclusion chromatography—a review of calibration methodologies. *J. Biochem. Biophys. Meth.*, **58**, 159–186 (2004).
45. Irvine, G.B., High-performance size-exclusion chromatography of peptides. *J. Biochem. Biophys. Meth.*, **56**, 233–242 (2003).
46. Liu, H.B., Lin, D.Y. and Yates, J.R., Multidimensional separations for protein/peptide analysis in the post-genomic era. *Biotechniques*, **32**, 898 (2002).
47. Watzig, H. and Gunter, S., Capillary electrophoresis—A high performance analytical separation technique. *Clin. Chem. Lab. Med.*, **41**, 724–738 (2003).
48. Kasicka, V., Recent advances in capillary electrophoresis of peptides. *Electrophoresis*, **22**, 4139–4162 (2001).
49. Li, S.F., *Capillary Electrophoresis*. Elsevier, Amsterdam, 1992.
50. Grossman, P.D. and Colburn, J.C., *Capillary Electrophoresis Theory and Practice*. Academic Press, San Diego, 1992.
51. Petersen, J.R., Okorodudu, A.O., Mohammad, A. and Payne, D.A., Capillary electrophoresis and its application in the clinical laboratory. *Clin. Chim. Acta*, **330**, 1–30 (2003).
52. Hernandez-Borges, J., Neususs, C., Cifuentes, A. and Pelzing, M., On-line capillary electrophoresis-mass spectrometry for the analysis of biomolecules. *Electrophoresis*, **25**, 2257–2281 (2004).
53. Cooper, J.W., Wang, Y.J. and Lee, C.S., Recent advances in capillary separations for proteomics. *Electrophoresis*, **25**, 3913–3926 (2004).
54. Hutterer, K. and Dolnik, V., Capillary electrophoresis of proteins 2001–2003. *Electrophoresis*, **24**, 3998–4012 (2003).
55. Pantazaki, A., Taverna, M. and Vidal-Madjar, C., Recent advances in the capillary electrophoresis of recombinant glycoproteins. *Anal. Chim. Acta*, **383**, 137–156 (1999).
56. Engelhardt, H. and Grosche, O., Capillary electrophoresis in polymer analysis, *in New Developments in Polymer Analytics I*, **150**, 189–217 (2000).

57. Shimura, K., Recent advances in capillary isoelectric focusing: 1997–2001 *Electrophoresis*, **23**, 3847–3857 (2002).
58. Kilar, F., Recent applications of capillary isoelectric focusing. *Electrophoresis*, **24**, 3908–3916 (2003).
59. Hu, Y., Huang, X., Chen, G.Y.J. and Yao, S.Q., Recent advances in gel-based proteome profiling techniques. *Mol. Biotechnol.*, **28**, 63–76 (2004).
60. Oh-Ishi, M. and Maeda, T., Separation techniques for high-molecular-mass proteins. *J. Chromatogr. B*, **771**, 49–66 (2002).
61. Patton, W.F., A thousand points of light: The application of fluorescence detection technologies to two-dimensional gel electrophoresis and proteomics. *Electrophoresis*, **21**, 1123–1144 (2000).
62. Van den Bergh, G. and Arckens, L., Recent advances in 2D electrophoresis: an array of possibilities. *Expert Rev. Proteomics*, **2**, 243–252 (2005).
63. Lilley, K. S., Razzaq, A. and Dupree, P., Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Curr. Opin. Chem. Biol.*, **6**, 46–50 (2002).
64. Gorg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R. and Weiss, W., The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **21**, 1037–1053 (2000).
65. Molloy, M.P., Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients. *Anal. Biochem.*, **280**, 1–10 (2000).
66. Cramers, C.A., Janssen, H.G., van Deursen, M.M. and Leclercq, P.A., High-speed gas chromatography: an overview of various concepts. *J. Chromatogr. A*, **856**, 315–329 (1999).
67. van Deurse, M.M., Beens, J., Janssen, H.G., Leclercq, P.A. and Cramers, C.A., Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography. *J. Chromatogr A*, **878**, 205–213 (2000).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 6

Mass spectrometry instrumentation and techniques

ÁRPÁD SOMOGYI*

Department of Chemistry, University of Arizona, Tucson, AZ, USA

*Tel.: +1-520-626-7272; Fax: +1-520-621-8407. E-mail: asomogyi@email.arizona.edu.

1. Introduction

In this chapter we provide a general overview of mass spectrometry instrumentation and techniques. After the discussion of some general questions and main features of mass spectrometers, the ionization methods, separation techniques, mass analyzers, and tandem mass spectrometry will be discussed. At the end of the chapter, a few arbitrarily chosen mass spectrometry terms will be mentioned for clarification. Somewhat surprisingly, these terms are not well understood and misleadingly used in everyday jargon.

One can justifiably argue that such a detailed discussion of instrumentation and technical features may not be interesting for general users with no mass spectrometry background. The author has a different opinion based on his many years of experience in teaching mass spectrometry for a quite general audience, including not only chemists but also biochemists, biologists, geneticists, medical doctors, and other colleagues working in clinical laboratories. These "inexperienced general users" find the discussion of instrumentation interesting and useful for better understanding of the mass spectra, and most importantly, their needs, and at times the limitations of mass spectrometry in certain areas of their research.

Of course, a compromise should be made and it is not the purpose of this chapter to bury the reader with a lot of technical details. The discussion presented here mimics a "lecture style" presentation, i.e., when simple but important questions are asked and analogies are given for better understanding. For technically inclined readers, we provide some references for guidance. We further encourage the readers to find more relevant and detailed works in relation to their research. These works (by the hundreds) are easily available on the Internet, for example.

Maybe the first questions we should ask at this point are: *"Why do we need this book at all? Is mass spectrometry so much better than any other analytical method?"* The golden rule in analytical chemistry is not to rely exclusively on one analytical method but rather use as many as you can and put the pieces of information together to get the best answers possible to your questions. What one should consider is the structural information content provided by a given analytical technique per unit time. For example, to study chirality in carbohydrate derivatives, nuclear magnetic resonance (NMR) spectroscopy is a much more reasonable choice than mass spectrometry. This does not mean that mass spectrometry cannot be used to study chirality (in fact, there are several papers in the literature in this field), but at present NMR undoubtedly provides stereochemical information in a much shorter time than mass spectrometry. Another example is the determination of protein structures. Obviously, X-ray crystallography can provide the greatest information content, such as bond lengths, bond angles, and torsion angles, but this technique requires the preparation of a pure (and in most cases, crystalline) protein that may take a lot of time. Even

though mass spectrometry cannot provide "X-ray quality" structural information, it can be used to check, for example, protein sample purity and to sequence proteins in a reasonably short time. (Note here that although it is widely used, the term *"structural information"* is not well defined and can mean different parameters (such as bond lengths and angles in X-ray crystallography) or structural units/chemical groups (such as the order of the amino acids in peptides and/or proteins).) Protein purity measurements can be performed within minutes with matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry, and sequencing of protein mixtures can be done in a couple of hours with current nanospray/high-performance liquid chromatography tandem mass spectrometry (nanospray–HPLC–MS/MS) measurements by using very little sample. The generally short analysis times make mass spectrometry suitable for high-throughput analyses, which is a significant advantage in clinical laboratories. The great sensitivity of mass spectrometry is definitely one of its strengths over other analytical techniques. Fluorescent-tag spectroscopy can, in principle, compete with mass spectrometry, but the application of this technique requires a more intensive pretreatment of the sample. For correctness, it should be noted that sample preparation for mass spectral analyses is also necessary, but the general trend in research and application is to reduce this time.

In summary, we can objectively state that mass spectrometry is among the most powerful analytical tools in clinical and medicinal chemistry. The samples from these laboratories are very often *complex mixtures* that may contain small amount of physiologically important analytes (e.g., drugs and metabolites) buried in the dirty environment of "biological matrices." Clinical diagnostic laboratories produce large number of samples, the timely analysis of which is crucial to make correct diagnosis. Pharmacokinetics (drug metabolism) studies also require the *quantitative* analysis of large number of samples taken at different times from different biological fluids, e.g., urine or blood. Multiple reaction monitoring (MRM) on a triple quadrupole instrument coupled with HPLC separation is a perfect technique for these quantitative studies and provides much more relevant information than an HPLC analysis with a ultraviolet (UV) detector only. Another important advantage of using a mass spectrometer over a UV detector is that structural information on coeluting components can be routinely obtained by HPLC–MS/MS measurements, but coelution may be overlooked by using solely a UV detector. Thus, mass spectrometry overlaps with many other analytical techniques providing not only an alternative way of analysis but also more coherent and reliable information on components of complex mixtures. Together with many other areas of applications (such as environmental, forensic, and material sciences), mass spectrometry is an important tool in medicinal chemistry with an expanded role and availability in more and more laboratories. The main aim of this chapter is to shed some light on the physical phenomena that make mass spectrometry such a powerful analytical technique.

2. General questions about mass measurement and mass spectrometry

How do we measure masses of big and small objects? It is relatively easy to determine the mass of heavy things, such as a book, a car, or a human being. In most cases, we use one of nature's four forces, the gravitational force (more specifically the Earth's gravity), to help us out and, in fact, we measure the *weight* of an object and use this information to determine the *mass*. For example, if you have a patient who complains about weight loss, you simply ask him or her to stand on a conventional scale and measure his/her weight. If the person weighs 70 "kilos" (used in everyday language), the mass of the person is approximately 70 kg. As a doctor, you can easily monitor the change in mass by measuring the weight in an easy, conventional way.

With lighter and lighter objects (or, equivalently, smaller and smaller masses) the use of a conventional scale would not be adequate—just think about measuring the weight (mass) of a light feather (for example, the one that flows with the wind in the beginning and the end of the movie *Forrest Gump*). With smaller and smaller masses, we would need more and more sensitive scales but, eventually, there is a lowest mass limit (e.g., a microgram, 10^{-6} g) that we could measure in the conventional way of measuring weights (i.e., by using the gravitational force).

We must have a different approach if we want to measure the mass of much lighter species, such as atoms and molecules. Fortunately, nature offers us a relatively simple way. This is because besides the gravitational force there are three other forces in nature. These are (i) the strong force (that holds the atomic nuclei together), (ii) the electroweak force (which is responsible for radioactivity of certain isotopes some of them are even used in clinical diagnostics), and (iii) the electromagnetic force (which is related to moving (accelerating) electronically charged particles). For our present goal of measuring the mass of atoms and molecules, the latter one, the electromagnetic force, is crucial. What we need to do is relatively simple: We have to make the atoms and molecules charged by a process called *ionization* and allow them to interact with electrostatic, magnetostatic, or electromagnetic fields by which the ions are separated (*ion separation*).

What are mass spectrometers? The instruments in which originally neutral atoms and/or molecules become charged (*ionized*) and are subjected to electrostatic, magnetostatic, or electromagnetic fields (*ion separation*) are called *mass spectrometers*. Large number of possible combinations of ionization (*ionization methods*) and ion separation (*mass analyzers*) are available in a great variety of both homemade and commercially available mass spectrometers. The common feature of the majority of mass spectrometers is that ionization and ion separation occur in the gas phase. The analyzed compounds need to be vaporized or transferred into vacuum either before or during the ionization. Most mass spectrometers operate in the vacuum range of 10^{-4} to 10^{-11} Torr.

How can mass spectrometry be used for chemical analysis? It is quite difficult to give a brief definition of mass spectrometry that fully covers all of its crucial features. A short definition recommended in the book by Sparkman (2000) is that "mass spectrometry is the study of matter based on the mass of molecules and on the mass of the pieces of the molecules" [1]. In broader terms, we can also say that mass spectrometry is a powerful tool in analytical and bioanalytical chemistry that provides detailed structural information on a wide variety of compounds with molecular weight (MW) of 1–1,000,000 Da by using a small amount of sample (nanogram, picomole, or femtomole of material). Another important feature is that mass spectrometers are easily coupled with separation technology, such as gas chromatography (GC) or HPLC. Mass spectrometry is an "ideal" tool to analyze complex mixtures, e.g., peptides resulting from the enzymatic digestion of proteins. With automated analyses, mass spectrometry is also a high-throughput technique with the capability of analyzing several hundreds of samples a day per instrument.

What is a mass spectrum? Fig. 1 shows a 70 eV electron impact (EI) ionization spectrum of acetone. This spectrum is a plot of relative abundance *versus* mass-to-charge ratio (*m/z*). The term "relative abundance" is used because the vertical axis is calculated by assigning the most intense ion signal to 100 (base peak) and the other ion signals (peak intensities) are normalized to this value. We measure the *mass-to-charge ratios* (*m/z*) from which the mass of a given ion can be determined based on the knowledge of the charge state. Obviously, if the charge state is one (such as in singly charged ions formed by loosing an electron, e^-), the *m/z* value directly gives the ion mass. The charge states of multiply charged ions can easily be determined, as will be discussed in the following text.

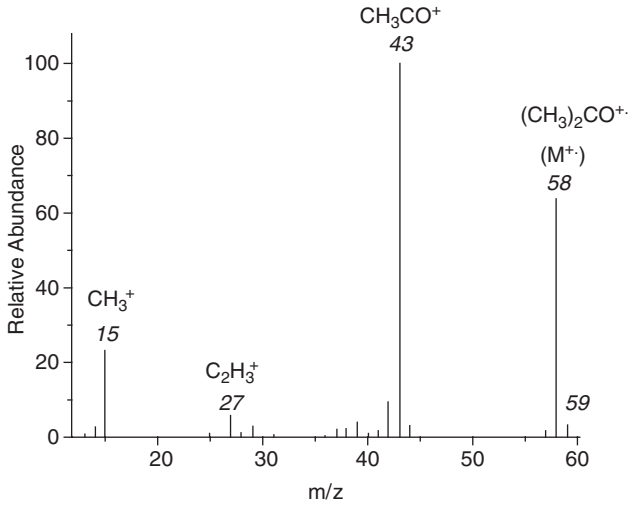


Fig. 1. Electron impact (EI) ionization spectrum of acetone (70 eV).

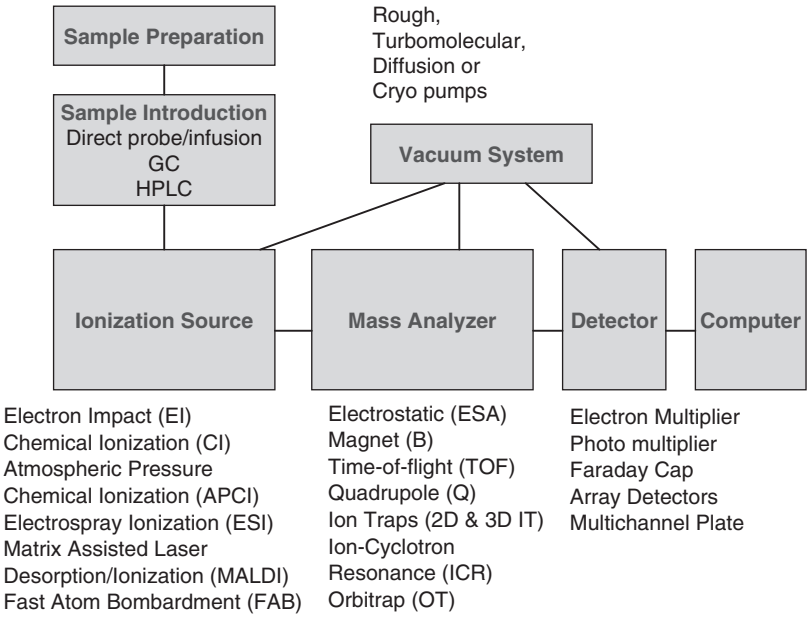


Fig. 2. Basic components of a mass spectrometer.

What are the general components of a mass spectrometer? A simplified block diagram of a mass spectrometer is shown in Fig. 2.

As was discussed in detail in Chapter 4, sample preparation is crucial, especially for samples of biological/biochemical origin. Samples can be introduced via a direct inlet, a GC, or an HPLC. Direct introduction may include a heated reservoir (for volatile compounds that are liquids at room temperature), a direct insertion probe (for relatively pure, synthesized solid organic compounds (EI) or fast-atom bombardment (FAB) and biomolecules (MALDI), and a direct infusion or flow injection for electrospray ionization (ESI) or atmospheric pressure chemical ionization (APCI, see the following text). GC and HPLC are strongly recommended and routinely used for the analysis of complex mixtures. (These separation techniques will be discussed briefly in Section 3, and has already been discussed in somewhat more detail in Chapter 5.)

Ionization is a crucial process occurring in the ionization source of mass spectrometers: There are several requirements about the ionization process: (i) The ionization process and ion extraction from the ionization source should be reasonably efficient to maintain low detection limits (high sensitivity); and (ii) the ionization efficiency, desirably, should not be sample dependent and the generated ion current should stay steady for reliable quantitation. The current state-of-the-art mass spectrometers are equipped with efficient ionization sources; however, for quantitation the use of internal standards is strongly recommended.

There are two main ways of generating positively charged ions: either by the removal of an electron, e.g., EI and field desorption (FD) ionizations or by addition of a proton or other "cationizing" agents, such as Na^+, K^+, Ag^+, etc. In the latter case, the proton/cation transfers are established by the so-called "soft" ionization techniques, including chemical ionization (CI), APCI, FAB, liquid secondary ion mass spectrometry (LSIMS), laser desorption (LD), MALDI, surface-enhanced laser desorption ionization (SELDI), ESI, desorption electrospray ionization (DESI), and direct analysis in real time (DART). Note that proton detachments can easily be achieved by most of the soft ionization techniques leading to the formation of negatively charged ions that are widely investigated as well.

Another important part of a mass spectrometer is the mass analyzer that is used to separate the ions. The simplest way of ion separation is just to let them fly and measure their time of flight. This type of analyzer is called time of flight (TOF). Here, electrostatic potential gradients are used to accelerate/decelerate the ions. Ion separation is achieved by the interaction of ions with an electrostatic (electric sector analyzer, ESA or orbitrap (OT)) or a magnetostatic (magnet, *B*) field. A resonant electromagnetic field is applied in quadrupoles (Q), and three-dimensional or linear ion traps (3D-IT and LTQ, respectively). A combination of electric (*E*) and magnetic (*B*) fields is used in Fourier transform ion cyclotron resonance (FT-ICR) instruments. Spatial coupling of mass analyzers is also used to perform tandem mass spectrometry (MS/MS) experiments. These types of experiments will be discussed later in this chapter (Section 6).

The final step of a mass spectral analysis is recording of the mass spectrum by detecting the ions after their separation. The detection of ions can be obtained consecutively in time ("sweeping" techniques) where a characteristic parameter of the analyzer, e.g., the magnetic field strength or radio frequency (RF) field amplitude, is being varied in time so that only ions with a particular *m*/*z* can hit the detector at a given time. In contrast, ions or ion packets can be detected simultaneously by recording the signal associated with all the ions at the detector plates. This complex ion signal (transient) is then deconvoluted by Fourier transformation (FT) that provides us the mass spectra. Modern mass spectrometers are equipped with detectors of great sensitivity. The detectors most commonly used include the electron multiplier, the photomultiplier, the conversion dynode, the Faraday cap, the array detector, and the charge or inductive detector. Detailed descriptions of these detectors are beyond the scope of the present chapter.

From the operational point of view, reliable vacuum systems are a prerequisite for mass spectral measurements. In most cases, manufacturers apply differential stage pumping to achieve the required pressure range(s). Rotary pumps are used to provide an initial vacuum of approximately 10^{-2} to 10^{-3} Torr. High-vacuum pumps such as diffusion pumps (10^{-6} to 10^{-8} Torr), turbomolecular pumps (10^{-7} to 10^{-8} Torr), and cryopumps (10^{-9} to 10^{-11} Torr) are used to reduce pressure further. Adequate knowledge in vacuum technology is essential in instrument design; however, this is also beyond the scope of this chapter.

As mentioned, mass spectrometry is a high-throughput analytical method. One can easily generate several megabytes or even gigabytes of data in an hour of operational time. Successful data processing in a timely manner requires state-of-the-art computers with intelligent data processing and search programs. In some cases, such as for proteomics research, clusters of computers are used to improve the speed and the reliability of database searching.

Finally, we note that there are numerous books and articles available in the literature for those who are interested in rigorous details of mass spectrometry. To adequately cite all these works would create a long list. Therefore, we provide a few to guide the technically inclined readers. For example, for general descriptions about mass spectrometry instrumentation, terminology, and mass spectral interpretation, we recommend the books by Sparkman [1], Watson [2], Chapman [3], Gross [4], Busch et al. [5], and McLafferty and Tureček [6]. For those who are interested in biological mass spectrometry and proteomics, the books by Siuzdak [7], Liebler [8], and Baer et al. [9] are recommended. Additional references will be recommended in the following sections.

3. Separation techniques: gas chromatography (GC), and high-performance liquid chromatography (HPLC)

Why do we need separation techniques? As will be discussed in Sections 5 and 6, state-of-the-art mass analyzers and tandem mass spectrometry allow mass spectrometry to be a powerful tool for the analysis of complex mixtures. The coupling of classical separation techniques with mass spectrometry further improves the utility of these combined techniques for mixture analysis. Mass spectrometers are the most sensitive and structure-specific *detectors* for separation techniques that, in general, provide more detailed and reliable structural information on components of complex mixtures than other conventional detectors (such as flame ionization, UV, reflective index detectors, etc.).

A simplified schematics for three main separation techniques, namely GC, high-performance or high-pressure liquid chromatography (HPLC), and supercritical fluid chromatography (SFC), are shown in Fig. 3. In all cases, the analyte molecules in a mixture (such as M1 and M2) are partitioned between a liquid-phase film on a solid substrate and a carrier flow (mobile phase). In GC, the carrier is a gas, most commonly helium (He), in HPLC, the carrier flow is a combination of common solvents, such as water, methanol, acetonitrile, etc., and in SFC the mobile phase is a supercritical fluid, usually CO_2. The partition of analyte molecules between the carrier phase and the liquid (stationary) phase depends on many factors, such as volatility, polarity, and hydrophobicity/hydrophilicity of the analyte, the chemical composition of the liquid phase, the flow rate, and the temperature applied. This partition can be visualized as a flooding river carrying debris of different sizes and shapes (analyte

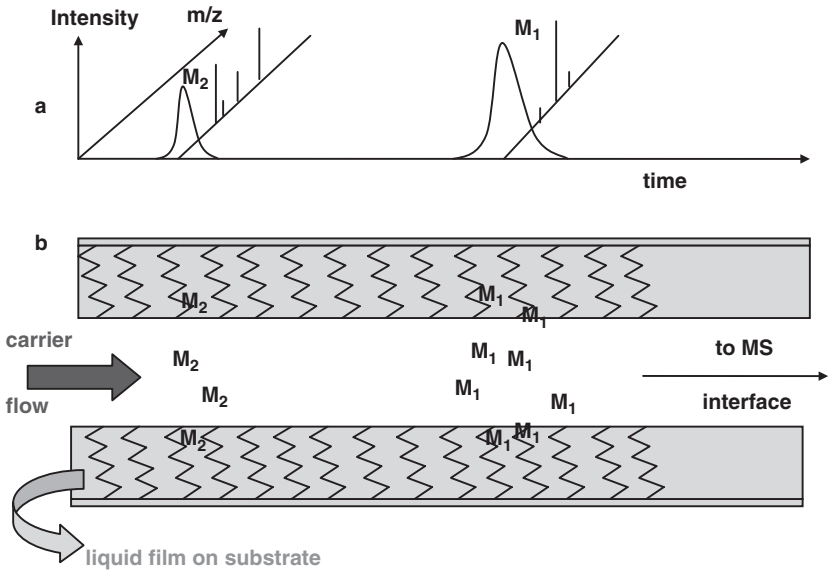


Fig. 3. (a) GC chromatogram and mass spectra of a two-component system. The plot of total ion intensity vs. increasing time is the total ion chromatogram (TIC). Individual mass spectra are obtained in a real-time mode for each component. (b) A simplified mixture flow of a two-component system in a GC column.

molecules) that are bouncing back and forth between the water (mobile phase) and the branches of the trees and bushes (stationary liquid phase) on the riverbank.

In GC, there is not much variance in the "chemical composition" of the mobile carrier phase, which is usually He gas. Therefore, chemical composition of liquid-phase molecules on the column and the temperature change during separation determine the effectiveness of separation. Combinations of chemical structures in the liquid-phase chains are common, such as the use of different ratios of methyl/phenyl silicons. The retention time, i.e., time necessary for a compound to pass through the column increases with the volatility of the compounds. To achieve better separation, a temperature gradient is applied that reduces peak broadening due to diffusion.

GC–MS is still widely used technique in environmental, forensic, and planetary (space) sciences. It is, however, limited to volatile and thermally stable compounds as they are injected to the GC via a high-temperature (250–300°C) injection port. Nonvolatile compounds can be analyzed after specific derivatization such as methylation, silylation, etc.; however, that requires additional sample preparation time. This is not always feasible as HPLC–MS is a better technique for a large variety of nonvolatile compounds, including those of biological importance. These include drugs and their metabolites, peptides, proteins, oligosaccharides, and oligonucleotides. For more details about GC/MS operation

and for a practical user guide, we recommend the book by McMaster and McMaster [10].

In HPLC, a greater variety of both the mobile (carrier) phase and the liquid phase is available to optimize separation for a wide variety of compounds. By varying the ratio of solvents by applying a solvent gradient, one can change the polarity of the mobile phase, which is a unique feature of HPLC compared to GC. Two main categories of liquid phases are applied: the "normal" phase and the "reverse" phase liquid layers. In reverse phase, nonpolar alkyl chains are exposed to the mobile phase. This provides stronger interactions with nonpolar (hydrophobic) analytes that will appear at longer retention time compared to more polar (hydrophilic) analytes. For further details of separation mechanisms and ionization techniques used in HPLC–MS, a good introductory book by Ardrey [11] is recommended.

The advantage of combining GC, HPLC, and SFC with mass spectroscopy is that fast-scanning mass analyzers allow us to record several mass spectra every second so that numerous mass spectra are produced during a chromatographic run. These mass spectra can either be direct "electron or chemical ionization" spectra (GC–MS) that are rich in fragments (see Section 4) or tandem mass spectra (see Section 6), in which an ion of interest at its retention time is selected and then further fragmented by an ion-activation method, usually collision-induced dissociation (CID). Fig. 4 shows such a combined HPLC–MS/MS run for a peptide mixture obtained by digesting a protein. Fig. 4a shows the base peak ion current as a function of time. The mass spectrum (MS) at a particular retention time (26.47 min) is shown in Fig. 4b. It is clear from this MS spectrum that there are two coeluting components (see doubly charged ions at m/z 571.4 and 643.2). The doubly charged ion at m/z 571.4 is then selected and fragmented to produce the MS/MS (fragmentation) spectrum (Fig. 4c). The fragment ions provide important structural information such as peptide sequence. Although not shown here, the structural information on the other component (m/z 643.2) was also obtained in about a second. Thus, MS/MS spectra are automatically generated for coeluting components allowing us to derive structural information, e.g., peptide sequence.

4. Ionization methods

Why do we need different ionization methods? Depending on the chemical properties of a molecule studied, different ionization methods should be applied. All ionization methods are not applicable to all molecules. For example, nonvolatile, heat-sensitive molecules, such as most of the biomolecules, cannot be ionized by EI ionization because the prerequisite for this ionization is the (thermal) evaporation of the sample. Also, the ion yield (ionization efficiency) depends on the

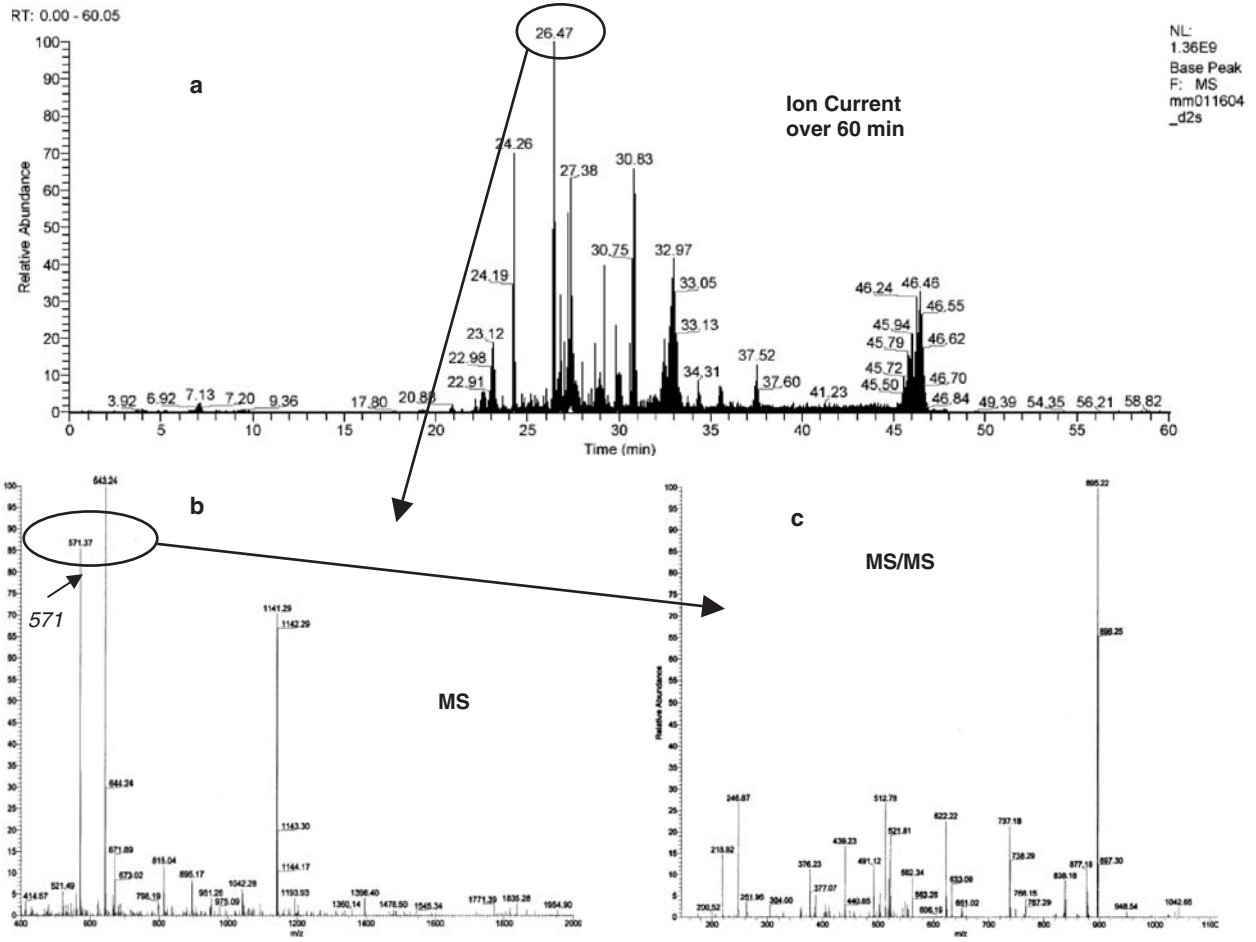


Fig. 4. HPLC is used to separate the components of a protein digest mixture. (a) Base peak ion current as a function of time. MS and MS/MS mass spectra are recorded in real time. (b) Full MS spectrum obtained at retention time = 26.47 min. Two main coeluting components are detected (see, e.g., doubly charged ions at m/z 571.4 and 643.2). (c) The tandem MS/MS (fragmentation) spectrum of the doubly charged peptide ion at m/z 571. The m/z values of the fragments are used to sequence the peptide.

chemical structure of a molecule so that the application of more than one ionization method for a given compound is desirable. In this chapter a brief summary of the most commonly used ionization methods is provided.

4.1. Electron impact (EI) ionization

Volatile molecules can be ionized in the gas phase by colliding them with a beam of high-energy (70 eV) electrons. Typically, lower molecular mass compounds are more volatile. EI is suitable for the analysis of these compounds in the molecular mass range of 1–800 Da. (Note, however, that there are compounds, such as fluorinated hydrocarbons or some transition metal complexes, that have MWs higher than 1000 Da and, yet, they are still volatile enough for EI analysis.)

A scheme of an EI ionization source is shown in Fig. 5. The electrons are emitted from a heated filament (made of tungsten or rhenium) and accelerated toward the source chamber. In the ionization chamber, some of these accelerated electrons collide with the evaporated neutral molecules so that the emission of two electrons occurs leaving behind a positively charged molecular ion. To form positively charged ions, the average energy of the electrons must exceed the ionization potential of the (originally) neutral molecule. Although the ionization potentials of most

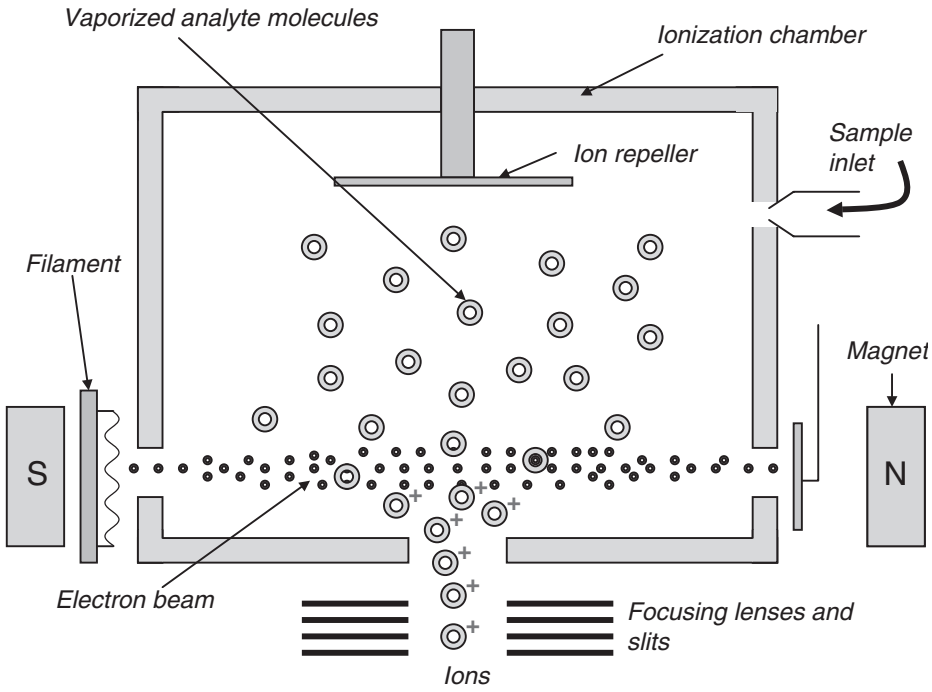


Fig. 5. Schematics of an electron impact (EI) ionization source.

common organic compounds are in the range of 8–12 eV, for routine and comparative studies a 70 eV electron beam is applied. At around this value of energy, the ionization efficiency (ion yield) is constant. Desirably, the electron beam should be well focused and a narrow energy spread of the beam should be maintained. This can be achieved by appropriate focusing and by a magnet (indicated by S (south) and N (north) poles in Fig. 5).

In most cases, the average background and/or analyte pressure in the EI source is low enough (less then 10^{-4} to 10^{-5} Torr) to avoid ion–molecule collisions (i.e., the mean free path is much longer than the dimension of the ionization chamber). As a consequence, EI mass spectra, in general, are free from ions originating from ion–molecule reactions. The applied repeller voltage of 1–5 V is high enough to force the ions to leave the source within a few microseconds. Thus, parent and fragment ions detected in EI mass spectra must be formed within the time frame of a few microseconds. This means that the unimolecular rate constants (k) are in the range of about 10^5 to 10^6 s^{-1}. To drive fragmentation reactions with this rate, a significant amount of internal energy is required, i.e., the kinetic shift (the difference between the actual average internal energy and the activation energy) is relatively large. The two main theory (Rice, Raisberger, Kespel, and Marcus (RRKM) and quasi equilibrium theory (QET)) that describe the main features of ion activation and fragmentation are beyond the scope of this chapter, but we recommend some fundamental works by Beynon and Gilbert [12], Cooks et al. [13], Forst [14], McLafferty and Tureček [6], and Vékey [15] and Drahos and Vékey [16].

The excess internal energy can easily be provided by collisions with 70 eV electrons since the electron energy is significantly larger than the ionization energies of common organic molecules (8–10 eV). Thus, during the ionization not only the elimination of an electron from a molecule (M) occurs but also an excited molecular ion (M^*) is obtained (Equation (1)).

$$M + e^- \rightarrow (M^*)^{+\cdot} + 2e^- \tag{1}$$

Owing to the excitation of the molecular ion, the extra internal energy deposited via EI allows the ion to fragment in the microsecond timescale, and if the internal energy is high enough, the fragments (F_{ij}) can even fragment further. The ions appearing in the mass spectra are, therefore, a result of competitive and consecutive reactions as illustrated by the "fragmentation" matrix of Equation (2).

$$(M^{+\cdot})^* \begin{cases} \rightarrow F_{11} \rightarrow F_{12} \rightarrow F_{13} \cdots \\ \rightarrow F_{21} \rightarrow F_{22} \rightarrow F_{23} \cdots \\ \rightarrow F_{31} \rightarrow F_{32} \rightarrow F_{33} \cdots \end{cases} \tag{2}$$

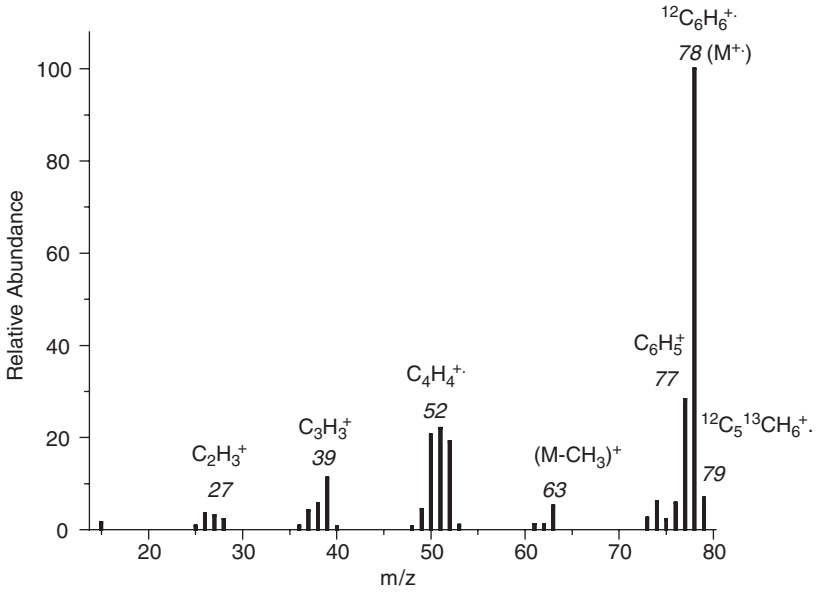


Fig. 6. Electron impact (EI) ionization spectrum of benzene (70 eV).

In Equation (2), F_{ij} denotes the fragment ion formed in the *i*th competitive and *j*th consecutive fragmentation step.

Three characteristic 70 eV EI ionization spectra are shown in Figs. 1, 6, and 7a (acetone, benzene, and tributyl amine, respectively). In the EI spectrum of acetone (Fig. 1) the molecular ion is at *m*/*z* 58 and this is the *nominal* MW of the neutral

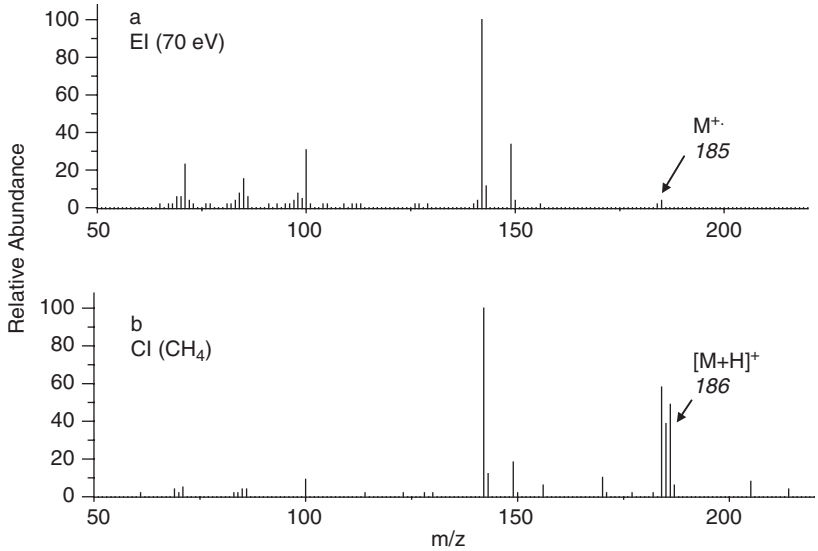


Fig. 7. (a) 70 eV EI spectrum and (b) CI spectrum (with methane reagent gas) of tributyl amine.

acetone. More precisely, this is the nominal mass of the acetone molecular ion. (For the definition of nominal and accurate masses, and isotope patterns, see the following text.) The ion at *m/z* 59 corresponds to a molecule in which one carbon is a ^{13}C isotope. (A reminder: Historically, mass spectrometry was developed to separate and determine the masses of different isotopes of elements.) Other peaks in the spectrum in Fig. 1 correspond to fragment ions of the molecular ion of acetone. The most abundant ion is a fragment ion at *m/z* 43 corresponding to the acetyl cation (CH_3CO^+). The most intense peak in the mass spectrum is called the *base peak* and, conventionally, all the other peak intensities are normalized to the intensity of the base peak (which is taken as 100%). Other ions include the ions at *m/z* 15, 14, and 13 corresponding to the methyl cation (CH_3^+) and subsequent (consecutive) hydrogen losses from the methyl cation. Formation of both the acetyl and methyl cations can be associated with a *direct* (C−C) *bond cleavage*. On the contrary, ions at *m/z* 27 and 29 originate from *rearrangement processes* in which some bonds are being broken while others are being formed. The activation energies of rearrangement reactions, in general, are lower than those of direct bond cleavages. However, rearrangement reactions require specific orientation (conformational and/or other rearrangement) of the atoms in the fragmenting ions, which is manifested in lower *frequency factors*. (For more details of ion-fragmentation mechanisms, see refs. [6,12,13,15].)

As demonstrated in Fig. 1, the molecular ion peak ($M^{+\cdot}$) is not necessarily the most intense (base) peak in the spectrum (in case of acetone, its relative intensity is approximately 64% and represents only about 28% of the *total ion intensity*). This is the measure of the "fragility" of the molecular ion; the relative intensity of the molecular ion increases with its stability. In comparison to the acetone molecular ion, for example, the peak corresponding to the molecular ion of benzene is the base peak in the 70 eV EI spectrum, which can easily be rationalized by the more conjugated (more stable) character of this ion (compare Figs. 1 and 6). The presence of the fragment at *m/z* 63 is particularly interesting because it corresponds to the loss of a methyl *radical* from the benzene molecular ion. Such a loss is difficult to rationalize from a closed-ring-type molecular ion. Instead, isomerization of benzene molecular ion to a $[CH_3-C{=}C{=}C{=}C-CH_3]^{+\cdot}$ conjugated structure is assumed. Isomerization reactions are quite common in mass spectrometry so that the structure of the molecular ion and that of the corresponding neutral are not necessarily the same.

Even though the 70 eV spectra of acetone (Fig. 1) and benzene (Fig. 6) show characteristic differences, there are important similarities as well. In both cases, the *odd-electron* molecular ions tend to lose neutral *radicals* to form *even-electron* cations. These reactions are driven by the fact that, generally speaking, even-electron ions are more stable than odd-electron ions. (As will be seen in the following text, this general rule has an important consequence for fragmentation of even-electron-protonated (or deprotonated) molecules that prefer to lose even-electron fragments, e.g., small neutral molecules.)

Although with different intensities, the molecular ions of both acetone and benzene can easily be observed in Figs. 1 and 6. Thus, the molecular mass information can be deduced for both compounds based on the position of the molecular ion peaks. There are cases, however, when the molecular ion is so fragile that it fragments completely, so that the molecular ion peak is either of very low intensity (see, e.g., Fig. 7a for tributyl amine) or not detected at all. Obviously, in these cases, the molecular mass determination based on 70 eV EI spectra becomes ambiguous, if not impossible. To reduce the fragmentation efficiency of the molecular ion, i.e., to obtain MW information, more gentle ("soft") ionization methods are required. Five of these methods are discussed in the following text.

4.2. Chemical ionization (CI)

CI is a gas-phase *ion–molecule reaction* in which the analyte (molecule) is ionized via a proton transfer process. (For more detailed description of the CI processes and analytical applications, a "classic" book by Harrison [17] is recommended.) The formation of the reactive ions in this ion–molecule reaction process is triggered by EI ionization of a *reagent gas* that is, most commonly, methane, isobutene, or ammonia. The partial pressure of the reagent gas (1–0.1 Torr) is much higher than that of the analyte (ca. 10^{-4} to 10^{-5} Torr), so the gas molecules can be considered as a protective shield for the analyte molecules to avoid direct EI ionization. EI ionization of methane results in the fragmentation of methane molecular ion and some of these ions react with neutral methane. The ionization of the analyte molecule occurs by proton transfer between reagent gas ions and the analyte, or to a less extent, by adduct formation. Some characteristic mechanistic steps for methane CI can be summarized as follows:

$$CH_4 + e^- \rightarrow CH_4^{+\cdot} \rightarrow CH_3^+,\ CH_2^{+\cdot},\ CH^+ \tag{3}$$

$$CH_4^{+\cdot} + CH_4 \rightarrow CH_5^+ + CH_3 \tag{4}$$

$$CH_3^+ + CH_4 \rightarrow C_2H_5^+ + H_2 \tag{5}$$

$$CH_5^+ + M \rightarrow [M + H]^+ + CH_4 \tag{6}$$

$$C_2H_5^+ + M \rightarrow [M + H]^+ + C_2H_4 \tag{7}$$

$$C_2H_5^+ + M \rightarrow [M + C_2H_5]^+ \tag{8}$$

The CH_5^+ ion formed in Reaction (4) is a strong acid. Reaction (6) is, therefore, likely exothermic and the protonated molecule can gain enough internal energy to fragment. Thus, methane is considered as a relatively "hot" CI gas.

Reaction (6) can be rewritten to illustrate the proton-transfer process in a more general way:

$$RH^+ + M \rightarrow [M + H]^+ + R \tag{9}$$

Here, RH^+ is the "protonated" reagent gas, M the analyte, $[M + H]^+$ the protonated analyte molecule, and R the reagent gas. Energetically, this reaction is preferred if the enthalpy change is negative ($\triangle H_r < 0$). This is true if the *proton affinity (PA)* of the analyte is greater than that of the gas. (As we will see later, this statement will be generalized for all of the soft ionization methods discussed below by substituting the "specific" reagent partner (gas) with a more general "matrix.") *PA* of a molecule (M) is defined as the negative value of the heat of the following reaction:

$$M + H^+ \rightarrow [M + H]^+ \tag{10}$$

$$PA = -\Delta H_r \tag{11}$$

Notice that the $\triangle H_r$ for Reaction (10) is less than zero due to the large value of heat of formation of proton (1530 kJ/mol). PAs of several compounds have been reported by Meot-Ner in the literature with special attention to the calibration of the PA scale [18]. The order of PAs of the most commonly used CI gases is: methane (PA = 5.7 eV) < isobutene (PA = 8.5 eV) < ammonia (PA = 9.0 eV). This is in agreement with the "strong acid" character of CH_5^+, which also implies that practically all organic compounds can be protonated by methane CI.

To illustrate characteristic differences between EI and CI spectra, the 70 eV EI and methane CI spectra of tributyl amine (MW: 185 Da) are shown in Fig. 7a and b, respectively. In the EI spectrum, the molecular ion at *m/z* 185 ($M^{+\cdot}$) is very low in intensity, making the MW determination somewhat ambiguous. In the case of CI ionization using methane as a reagent gas, the peak corresponding to the protonated molecule $[M + H]^+$ can easily be recognized at *m/z* 186. Owing to the low PA of methane, fragmentation of the $[M + H]^+$ still occurs providing structural information. For example, a fragment ion at *m/z* 142 can be assigned as $(C_4H_9)_2N{=}CH_2^+$. Although the ion at *m/z* 142 is the base peak in both the EI and CI spectra, there are important differences between the mechanisms leading to this ion: In the EI mode, this ion is formed by the loss of propyl radical from the odd-electron molecular ion, while in CI this ion is generated by the loss of a neutral molecule, propane, from the even-electron-protonated molecule. This is again consistent with the relative stability of even-electron ions (as discussed earlier).

The spectra in Fig. 7 can also be used to illustrate another important rule in mass spectrometry, the *nitrogen rule*. The nitrogen rule states that any common organic molecule or odd-electron ion that contains odd number of nitrogen atoms has an odd (nominal) molecular mass. For example, tributyl amine contains one nitrogen atom; thus, the nominal MW must be odd, and so it is (185 Da). However,

acetone (MW = 58) and benzene (MW = 78) that contain no nitrogen atoms have even molecular mass. The nitrogen rule has many important implications and applications. For example, the fragment ion at *m*/*z* 142 cannot be an odd electron ion containing one nitrogen. Indeed, it is an even-electron ion containing one nitrogen atom (see above). Naturally, the nitrogen rule can be applied for protonated (or deprotonated) molecules as well. Of course, in this case the numbers "flip" around: The nominal mass of an ion corresponding to a *protonated molecule* that contains odd number of nitrogen atoms must be an even number (see, e.g., the $[M + H]^+$ of tributyl amine at *m*/*z* 186). Detailed interpretation of EI and CI spectra are beyond the scope of this book, but for interested readers the book by McLafferty and Tureček [6] is, again, strongly recommended.

4.3. Fast-atom bombardment (FAB) and liquid secondary ion mass spectrometry (LSIMS)

FAB and LSIMS are closely related soft ionization techniques. A simplified scheme for both techniques is shown in Fig. 8.

The main difference between the two techniques is that a neutral atomic beam (Ar or Xe) is used in FAB, while a Cs^+ cation beam is used in LSIMS as a primary (ionizing) beam. In both cases, the analyte is mixed with a high-viscosity liquid matrix (proton-transfer agent), such as glycerol, thioglycerol, *m*-nitrobenzene alcohol, triethanol amine (TEA), etc. Combinations of these matrix components are also used to enhance ionization efficiency. For example, a glycerol:thioglycerol:*m*-NBA 2:1:1 mixture containing 0.1% of trifluoroacetic acid (TFA) effectively generates protonated molecules for many organic compounds. Although FAB and LSIMS

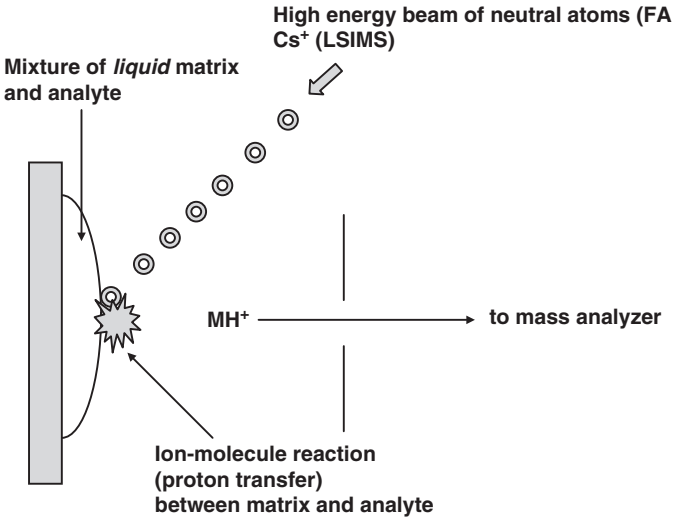


Fig. 8. A simplified scheme for FAB and LSIMS ionization.

ionizations are sometimes considered to be "outdated," they are useful and alternative ionization techniques to electrospray/nanospray ionization (ESI), APCI, and MALDI. The application of FAB and LSIMS is especially justified in an academic environment for elemental composition determination of relatively small synthetic organic compounds by accurate mass (high resolution) experiments. For more details about the chemical aspects of FAB, we recommend the review by Fenselau and Cotter [19].

4.4. Electrospray ionization (ESI)

The significance of ESI in the analysis of biomolecules by mass spectrometry is well acknowledged by awarding a shared Nobel Prize in 2002 to its inventor John Fenn (currently at the Virginia Commonwealth University, Richmond, VA, USA). As he pointed it out demonstratively, *"we taught elephants to fly."* Elephants, of course, stand for a wide variety of large biomolecules including peptides, proteins, oligonucleotides, oligosaccharides, glycolipids, etc. (For early papers on ESI ionization, see, e.g., the ones by Doyle et al. [20] and Fenn et al. [21, 22], and for an overview book, see the one edited by Cole [23].)

Electrospray ionization is an ionization process by which analyte molecules or ions present originally in solution are transferred to the gas phase through either solvent or ion evaporation. Although the experimental setup is relatively simple, the ion-formation mechanisms are still under systematic studies [24–26]. A scheme for an electrospray source is shown in Fig. 9, while a simplified ion-formation mechanism is indicated in Fig. 10.

In ESI the analyte previously dissolved in a solution is introduced into the ESI source via a needle either by direct infusion or as an eluent flow from an HPLC chromatograph. The most commonly used solvents include water, methanol, and acetonitrile. Their combinations and specific use depend on the solubility of the analyte. For direct infusion, a typical flow rate is in the range of 1–5 μl/min. More

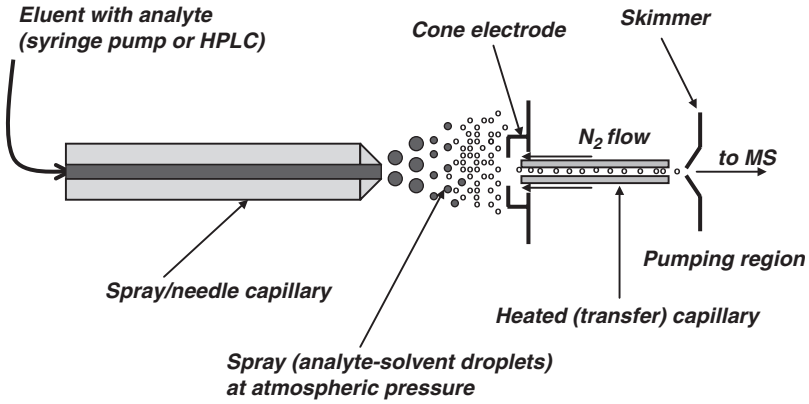


Fig. 9. Characteristic components of an electrospray ionization (ESI) source.

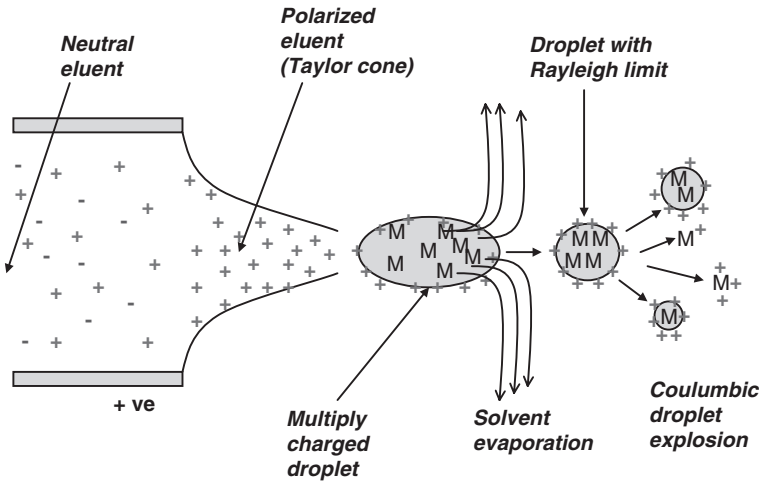


Fig. 10. A simplified mechanism of ion formation in the electrospray ionization process.

recently, samples were analyzed with much lower flow rates of a few tens of nanoliters per minute. This technique is called *nanospray* ionization that requires much less sample amount providing greater sensitivity than "conventional" ESI. The ESI flow rate is characteristically higher when HPLC is used for sample introduction (approximately 50 μl/min). In many cases, microbore analytical columns are used in the HPLC analysis. When higher flow rates are necessary for the HPLC analysis, the eluent leaving the HPLC column can be split so that only a small percentage of it is transferred to the ESI needle.

The electrospray itself is formed as a result of a large electrostatic potential difference between the syringe needle and a counter (cone) electrode. In the absence of this electrostatic field, the droplet formed at the end of the syringe needle would simply drop to the ground whenever the adhesive surface tension cannot compensate for the weight of the droplet. However, in the presence of a large electrostatic field, the solution at the end of the needle is polarized (see Taylor cone in Fig. 10) and torn away from the needle. This way, depending on the applied potentials, positively or negatively charged droplets are formed (Fig. 10). (To form positively charged droplets, the needle potential can be kept, for example, at +4 kV, and cone voltage at, e.g., 200 V.)

The more or less uniformly sized droplets enter a heated transfer capillary in which the solvent molecules are being further evaporated. As a consequence, the surface charge density increases until the droplet size reaches the "Rayleigh" limit at which the surface tension cannot compensate for the "Coulombic" repulsion associated with the surface charge. At this point, the droplet explodes ("Coulombic" explosion, Fig. 10) and smaller size droplets are formed. This process can continue until virtually no solvent molecules are present, but only

protonated (or deprotonated) analyte molecules. Alternatively, ions can be evaporated directly from the charged droplets. It is easy to envision that in these process multiply charged ions can easily be generated. In fact, formation of *multiply charged ions* is an important characteristic feature of ESI. Note that depending on the transfer capillary temperature and the solvent used, small droplets can survive the journey through the transfer capillary. To retain these droplets and prevent them from entering the ion guide and analyzer region of the mass spectrometer, a skimmer is used at the "entrance" of the lower pressure ion guide/mass analyzer section. Another disadvantage of the "linear" arrangement sketched in Fig. 9 is that small salt particles can also enter the mass analyzer region causing contamination. Perpendicular (or Z-type) sprays are currently developed and successfully used to overcome this problem.

A typical ESI spectrum for a protein (lysosyme) is shown in Fig. 11. The multiply charged molecular ion pattern is clearly recognizable. Note that although this ESI spectrum corresponds to only one protein, there is a *mixture* of ions in the spectrum each of which has a different mass-to-charge ratio (reminder: In mass spectrometry, the *m*/*z* ratio is measured). To calculate the molecular mass (or MW) of the protein, the charge states of the individual ions should also be determined. Thus, we have two unknowns, the MW and the charge state (*n*). To determine

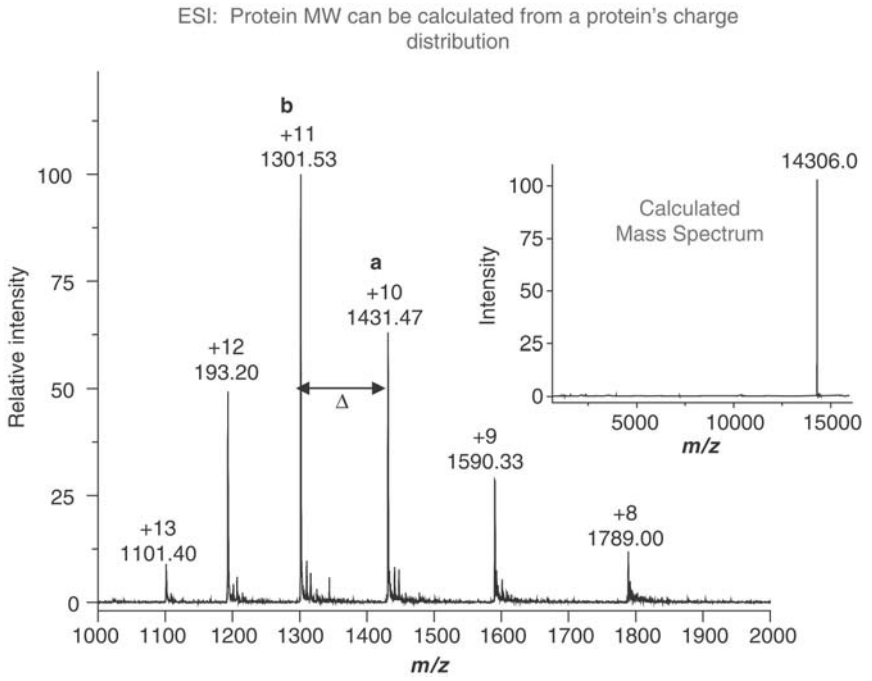


Fig. 11. ESI spectrum of lysosyme.

the values of two unknowns, we need at least two independent equations. For example, for two neighboring ions, the *m*/*z* values of which are denoted by ***a*** and ***b*** in Fig. 11, the relationship between ***a***, ***b***, the charge state, and MW can be written as:

$$\left(\frac{m}{z}\right)_a = \boldsymbol{a} = \frac{[\mathrm{MW} + n]}{n} \tag{12}$$

$$\left(\frac{m}{z}\right)_b = \boldsymbol{b} = \frac{[\mathrm{MW} + n + 1]}{(n+1)} \tag{13}$$

With simple arithmetic rearrangements, the charge state at the ion ***a*** can be determined as:

$$n = \frac{(\boldsymbol{b} - 1)}{(\boldsymbol{a} - \boldsymbol{b})} \approx \frac{\boldsymbol{b}}{\Delta} \tag{14}$$

Equation (14) provides a "geometric" solution that is easy to remember. The procedure is simple: (i) Select any two neighboring peaks in the spectrum, (ii) calculate the difference between these two peaks (Δ), and (iii) divide the *m*/*z* of one of the peaks by this difference, which will give the charge state of the other one. This simplified method works because, in most cases, $b >> 1$, so the correction is negligible. In the spectrum, shown in Fig. 11, $\Delta \approx 130$ so that $1301/130 \approx 10$ and this is the charge state of ion ***a***. The MW of the protein can then be determined as $[1431.47 \times 10 - 10] \approx 14{,}305$ Da. The correction by 10 is necessary because the measured ion mass is larger than the MW by the mass of 10 ionizing protons. For more precise determination, all peaks should be considered in a similar way and the calculated MW values should be averaged. This simple mathematical process is called "deconvolution" and provided by several manufacturers as a part of their data processing program. It will be demonstrated in the FT-ICR mass analyzer section (Section 5.4) that, if the resolution of a mass spectrometer is good enough to separate the isotopes of a given ion, the charge state can directly be determined by using the observed *m*/*z* differences (1/*n*) between the (carbon) isotope peaks of an ion as well.

ESI has a great advantage over other "matrix assisted" ionization methods (such as FAB (LSIMS) and MALDI) that peaks associated with matrix ions do not appear in the ESI spectra. This is especially useful for the analysis of smaller molecules, such as pharmaceutical products and their metabolites. Another advantage is related to sample introduction: Because samples are introduced in solution, HPLC is, naturally, a good "coupling" component of ESI making it suitable for mixture analysis. Another consequence of the sample introduction in solution phase is that ESI provides a way to study biomolecules in their native-like (solution phase) environment. For example, noncovalent (e.g., enzyme/substrate) interactions or protein denaturing kinetics can be followed by ESI measurements at least in a qualitative way. Other advantages of ESI/nanospray include a wide mass range of the

compounds to be measured with good sensitivity (from the low picomole to a few tens of femtomole level). As an impressive feature of these ionization methods, we refer here to the analyses of large protein complexes with molecular masses greater than 200,000 Da, e.g., by the groups of Robinson [27] and Wysocki [28]. Disadvantages include easy contamination of the transfer line and low salt tolerance so that sample pretreatment may be necessary before analysis.

These drawbacks can be overcome by using two recently developed desorption ionization techniques. One of them is *desorption electrospray ionization (DESI)* [29]. This technique is related to both ESI and desorption ionization methods, such as secondary ion mass spectrometry (SIMS) and LD. In DESI, electrosprayed charged droplets generated from solvents are directed at a surface of interest in air. No matrix is necessary and the surface investigated can easily be moved during the analysis allowing "mapping" of the surface for certain analytes. The usefulness of DESI has been demonstrated for small molecules (drugs), peptides, and proteins, as well as for *in vivo* analysis [29]. Another technique is the so-called *direct analysis in real time (DART)* developed by Cody et al. [30]. DART refers to an atmospheric-pressure ion source that allows analysis of gases, liquids, or solids on surfaces in open air. The DART source operates by exposing the sample to a dry gas stream (typically He or N_2) that contains long-lived electronically and/or vibrationally excited atoms or molecules. The excited-state species can directly interact with the sample to desorb and ionize the sample (Penning ionization). Similarly to DESI, DART has been successfully used for the direct analysis of samples such as clothing, human skin, pills, plant materials, etc.

4.5. Atmospheric pressure chemical ionization (APCI)

As the term implies, in APCI, analyte molecules are ionized by ion–molecule reactions that take place at *atmospheric pressure*. A scheme for an APCI source is shown in Fig. 12. This ionization technique shows similarities with ESI in that the samples are sprayed into the source; thus, this technique is also very commonly used with HPLC (in fact, APCI allows higher flow rates than ESI). On the contrary, there are significant differences between ESI and APCI. First, in APCI, the samples are sprayed into a *heated* ionization source ($t > 400$°C) so that the analyte molecules are vaporized. (This implies that APCI is not suitable for the analysis of thermally labile compounds.) An essential part of the APCI source is a corona discharge in which O_2 and N_2 molecules are ionized and further react with solvent molecules in the gas phase at *atmospheric pressure* to form ions that will protonate (or deprotonate) the analyte molecules. (Reminder: In "classical" CI, ion–molecule reactions also take place in the gas phase, but at lower pressure (1−0.1 Torr).) APCI is widely used for ionization of smaller molecules, such as drugs and their metabolites, pesticides, steroid derivatives, lipids, etc. ESI and APCI are often compared to each other in several applications. For example, for the determination of cyclosporin A in rat plasma, see, e.g., the work by Wang et al. [31].

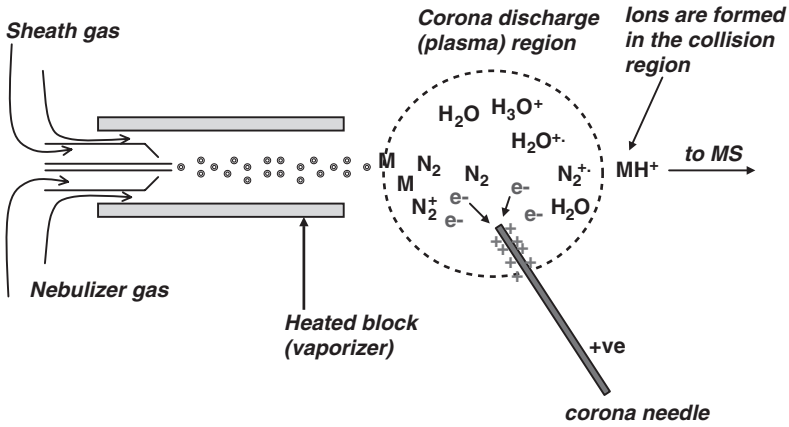


Fig. 12. Main components of an APCI ionization source. Ion formation in the corona discharge and collision region is also indicated.

4.6. Matrix-assisted laser desorption/ionization (MALDI)

"Elephants," i.e., large biomolecules, can be taught to fly not only by ESI but also by using laser desorption ionization (LDI). This technique was also greatly acknowledged by awarding the shared 2002 Nobel Prize to its developer Koichi Tanaka [32]. At about the same time, Karas and Hillenkamp realized that laser ionization efficiency could be significantly improved by applying a matrix so that a modified method termed "matrix-assisted laser desorption/ionization (MALDI)" was developed [33].

As the name indicates, MALDI is a desorption/ionization method, so it has some similarities with FAB and LSIMS ionization (see Section 4.3). There are, however, important differences as indicated in Fig. 13. First, the analyte is crystallized together with the matrix, i.e., no liquid matrix is involved. Second, the primary beam is a laser (photon) beam and not a particle beam. Many of the commercially available instruments are equipped with a N_2 laser, the frequency of which falls in the UV region (337 nm). Infrared (IR) lasers are also used, but they are not as common as the N_2 laser. The most common matrices used in UV MALDI experiments include nicotinic acid, benzoic acid and cinnamic acid derivatives, dithranol, azobenzoic acid derivatives, etc. (A section of a MALDI plate in Fig. 14. shows spots with different colors that are associated with different matrices.) These matrices have two common structural features: They contain a group that is a source of an acidic proton and they have absorption at or around 337 nm (the N_2 laser wavelength). Similarly to ESI ionization, the mechanisms of MALDI processes are still under investigation (see, e.g., Dashtiev et al. [34] and Vértes et al. [35]). The most important steps of MALDI ionization mechanism can be briefly summarized as follows. The matrix molecules that are in great excess to the analyte molecules are electronically excited by the UV laser, and this energy is transferred to vibrational energy

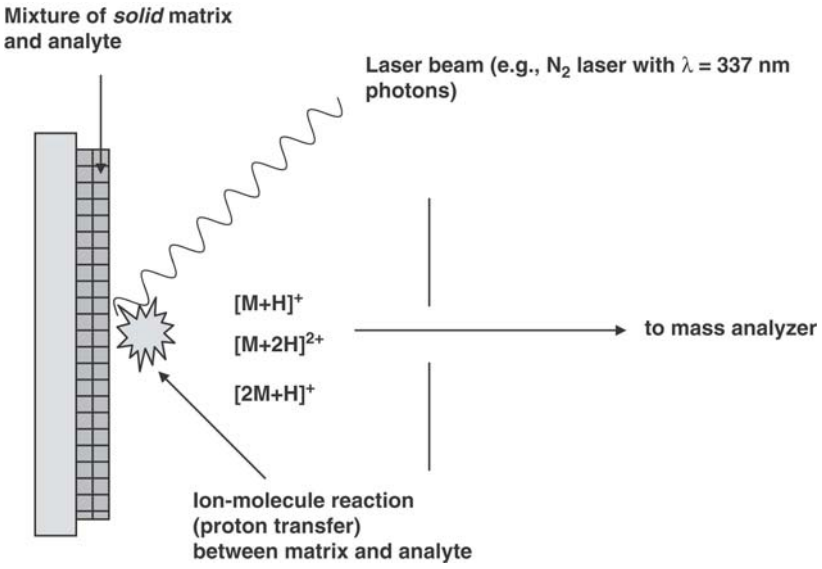


Fig. 13. Matrix-assisted laser desorption/ionization (MALDI).

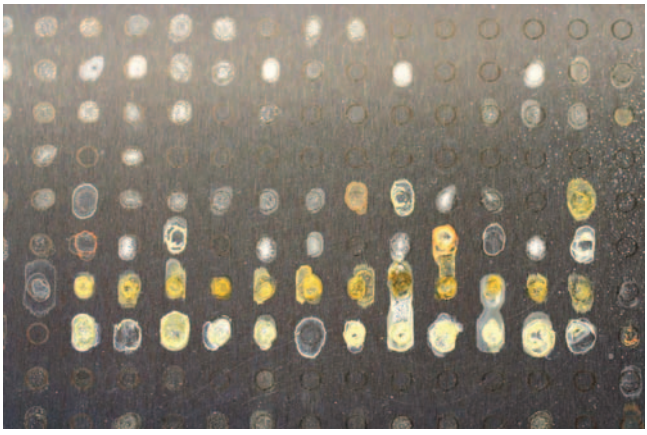


Fig. 14. Section of a 384-well MALDI plate. Different colors indicate different matrices used for different types of compounds such as proteins (sinapinic acid, bright white spots), peptides (α-cyano cinnamic acid white spots), synthetic polymers (dithranol, yellow spots), etc.

that is also manifested in local heating (melting) of the crystal. The locally melted crystal is then transferred to the vacuum carrying the analyte molecules. In this plume, proton transfer between the matrix and analyte molecules takes place.

In the MALDI process, mostly singly charged ions are formed, although these ions can be accompanied by some doubly and, occasionally, triply charged ions. In addition, noncovalent adducts, such as dimers, trimers, etc., of proteins and/or matrix adducts of certain analytes can also be observed. This is well illustrated in

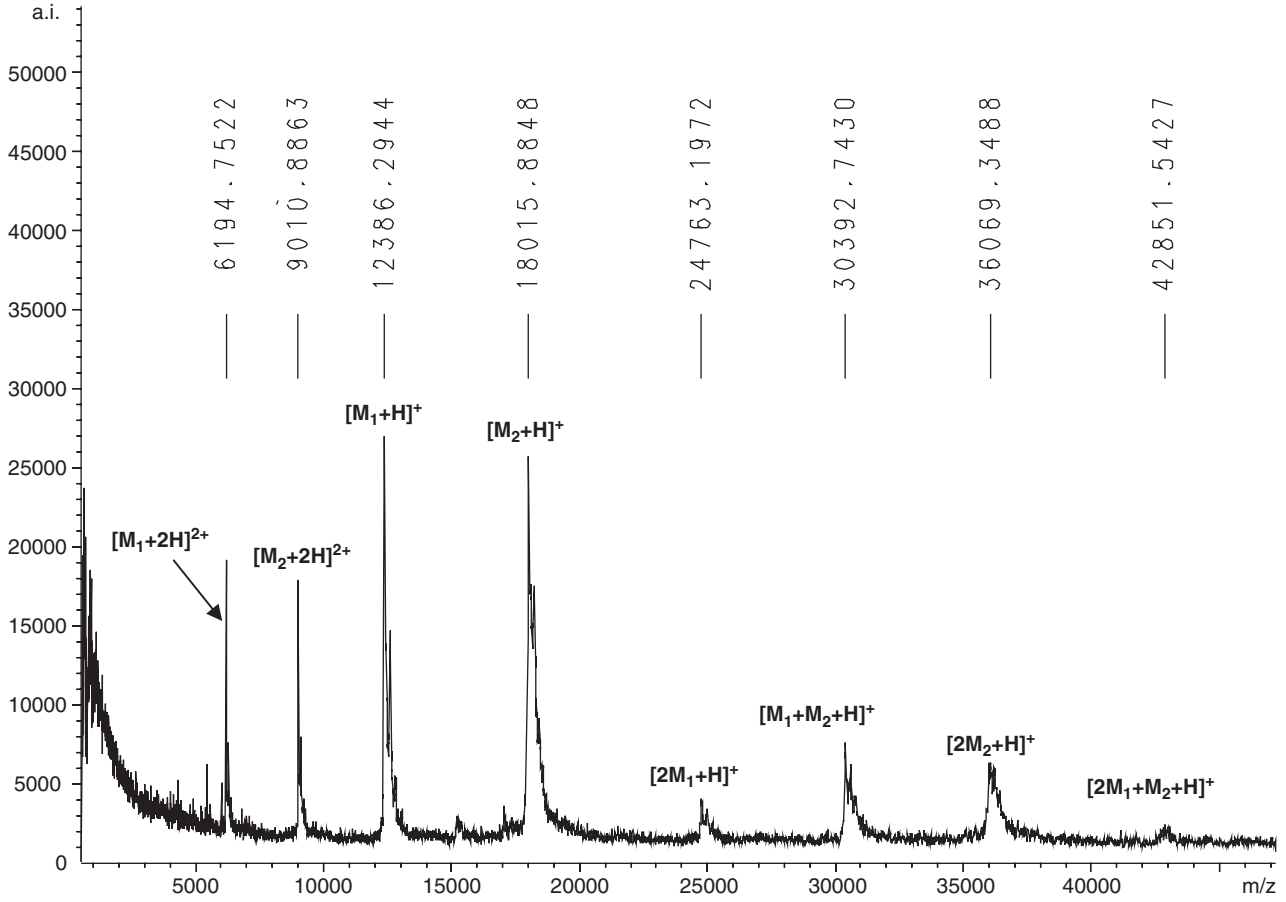


Fig. 15. MALDI-TOF spectrum of two proteins obtained by using a linear acquisition mode with sinapinic acid as matrix. The two proteins have average molecular weights $M_1 = 12{,}385$ Da and $M_2 = 18{,}014$ Da, respectively.

Fig. 15, which shows a characteristic MALDI-TOF spectrum of a two-component protein mixture using sinapinic acid (SA) as a matrix. A broad, tailing peak is observed in the lower mass region of the spectrum (<2000 Da) that is associated with ions from the matrix. This matrix interference is not disturbing in this particular case of protein analysis but could be a disadvantage when smaller molecules are studied. One way to eliminate the matrix interference is the application of the so-called "desorption/ionization on silicon" (DIOS) technique [36].

DIOS is a matrix-free laser desorption ionization technique that uses a pulsed laser shined on porous silicon. Because the porous silicon is a UV-absorbing semiconductor, a conventional N_2 laser (337 nm) can also be used in DIOS. No other significant modifications are necessary, so a conventional MALDI-TOF instrument can be used for DIOS experiments. In comparison with direct laser desorption/ionization (LDI), DIOS does not result in fragmentation of the generated ion. This is because the UV photon energy is mostly absorbed by the silicon surface and only part of this energy is transferred to the analyte. The large surface area of porous silicon allows obtaining low detection limits. High-throughput DIOS analysis of several small molecules has already been demonstrated and mechanistic studies have been carried out by the groups of Siuzdak and Vértes [36, 37].

MALDI ionization is, in general, more sensitive than ESI; it is routinely used for the analysis of peptides at the low femtomole or high attomole level. Another advantage of MALDI is that, coupled with a TOF analyzer, biomolecules up to about 500,000 Da can be investigated. The mass range can even be stretched further to 1–2 MDa, but detection of these "ultrahigh" masses requires sensitive and specific detectors, such as the low-temperature detector used by, e.g., the Zenobi group [38]. Disadvantages of MALDI include low salt tolerance, even though it is generally less critical than for ESI. There are several cases when sample pretreatment (e.g., desalting by the solid phase extraction) is desirable, if not required.

5. Mass analyzers

Why do we need mass analyzers? Obviously, it is not enough to generate ions by different ionization techniques (see Section 4) but it is also necessary to separate them from each other. Mass analyzers are used for ion separation, and several mass analyzer types are commercially available. Overview of these mass analyzers can be simplified by considering two basic physical phenomena:

i) charged species can be easily accelerated by applying an electrostatic potential difference, and
ii) an accelerated electric charge generates an electromagnetic field.

As a consequence, we can separate charged species based on their

i) time of flight (TOF analyzers), and

ii) their interaction with an electrostatic (electrostatic analyzer (ESA) and orbitrap (OT)), a magnetostatic (magnetic sector (B)) and a magnetostatic and electrostatic fields (ion cyclotron resonance (ICR) analyzers), or an electromagnetic field (quadrupole (Q), three-dimensional quadrupole ion trap (3D QIT), or "two dimensional" linear ion-trap (LTQ) mass analyzers).

One can argue that this classification is quite arbitrary and does not follow the conventional or historical description of mass analyzers. We believe, however, that emphasizing the similarity of physical phenomena that are essential for ion separation helps greatly in better understanding their basic operational principle. Combinations of these analyzers are also very common, especially in tandem mass spectrometers that will be discussed briefly in Section 6.

A few general and desirable properties of mass analyzers should be mentioned here. (i) To achieve good selectivity, mass analyzers should separate the ions with reasonable resolution. (ii) Sensitivity (i.e., the number of ions detected) depends not only on the ionization efficiency but also on the transmittance of mass analyzers. (Note that state-of-the-art detectors are good enough to detect only 10–100 ions.) (iii) It is desirable to have a mass analyzer that is compatible with the ionization source that can provide continuous or pulsed ion beams with either low or high initial kinetic energy. (iv) Finally, the mass analyzers should have the appropriate mass-to-charge (m/z) limit to be able to detect compounds with a wide molecular mass range.

In the forthcoming sections a brief summary of the most important mass analyzers is presented. Detailed description of their operation is beyond the purpose of this book, but readers with special interest in the physics and mathematics of operations can easily find hundreds of articles and books in the literature. We provide a few relevant references for guidance throughout the text.

5.1. Time-of-flight (TOF) analyzers

An excellent and detailed overview of TOF analyzers can be found in Cotter's book [39]. The principle of operation is relatively simple: By applying an electrostatic acceleration field (V), ions with a charge of zq (where q is the unit charge and z indicates the charge state) will gain a well-defined kinetic energy (E_{kin}) from which the velocity (v) of the ion can be determined.

$$E_{kin} = z\,q\,V = \frac{1}{2}mv^2 \tag{15}$$

If we just simply let the ion with a velocity of v to fly for a distance of d, the time of flight (t) can be determined as:

$$t = \frac{d}{v} \tag{16}$$

Note that in "real" instruments this equation is more complicated, but this does not undermine the importance of the simple fact that the TOF can directly be correlated to the mass-to-charge ratio (m/z) [39]. By combining Equations (15) and (16) this relationship can be written as:

$$t = \left(d/\sqrt{(2V)}\right)\left(\sqrt{(m/zq)}\right) \tag{17}$$

This means that the TOF is proportional to the square root of the mass-to-charge ratio (m/z), i.e., lighter ions have shorter arrival time than the heavier ions. This is illustrated in Fig. 16.

The analogy for a TOF analyzer could be a track-and-field race with a noticeable difference that heavier (overweight) persons do not necessarily run slower than lighter (underweight) people. A sharp start signal is obviously necessary to start a "fair" run. This start signal can conveniently be a laser pulse, so TOF analyzers are naturally coupled with MALDI ionization sources (MALDI-TOF instruments). An alternative way of generating ion packets is the application of a perpendicular pulse to an originally continuous ion beam. This pulsing technique is used in Q-TOF instruments, for example. Even though the laser pulse or the ion beam pulsing is relatively short in time, ions are formed (or pulsed) with a

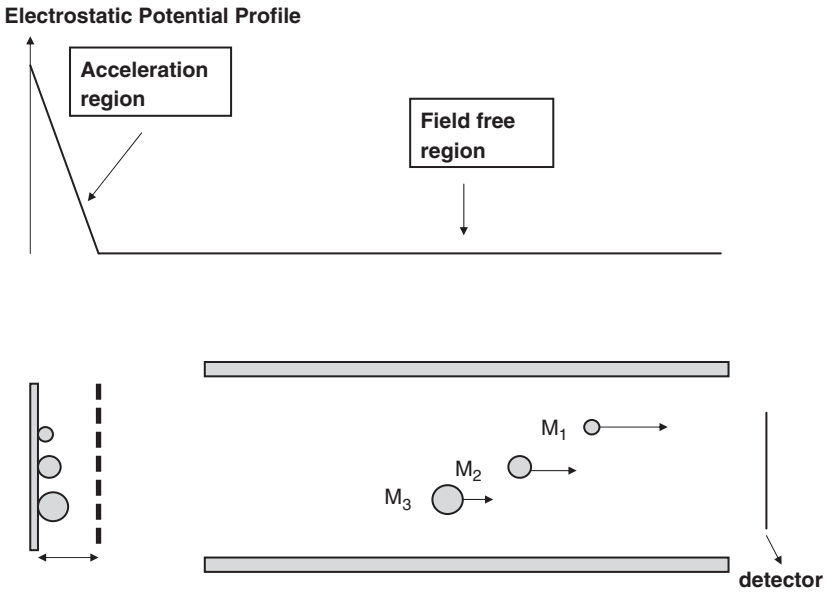


Fig. 16. Schematic diagram of a TOF analyzer: Lighter ions fly faster than heavier ones ($v(M_1) > v(M_2) > v(M_3)$) (masses and velocities are not in scale). The upper part of the figure shows a simplified electrostatic potential indicating the acceleration region and the field-free region.

noticeable spatial and velocity spread. These definitely decrease separation efficiency of the ions that may be manifested in not adequate resolution and also not satisfactory mass accuracy. Nevertheless, even the low-resolution, "linear mode" ion detection technique provides important information about protein purity (see, e.g., Fig. 15). This linear MALDI-TOF acquisition technique is also widely used in cell imaging (or protein profiling) studies that have great potential in clinical diagnosis.

The resolution and mass accuracy of TOF analyzers has been significantly improved by introducing the delayed extraction technique and the reflectron (ion mirror). The delayed extraction is used to compensate for the spatial spread, while the reflectron is used to reduce the velocity (kinetic energy) spread. The latter is illustrated in Fig. 17. Ions with higher velocity (but with the same *m*/*z* ratio) penetrate deeper in the electrostatic field of the reflectron so that they are forced to travel a longer distance. Hence, ions with greater velocities have a longer flight path than those with lower velocities. The detector should be positioned at a place where the faster moving ions "catch up" with the slower ions.

TOF analyzers have several advantages: (i) A reasonably good resolution (up to approximately 20,000) can be achieved, (ii) large mass range (up to approximately *m*/*z* 2,000,000) is accessible with special detectors, (iii) fast duty cycles (10–5000 scans/spectra/s) can be used, and (iv) its high transmission provides excellent sensitivity (e.g., at the 1–10 fmol level). It is not surprising, therefore,

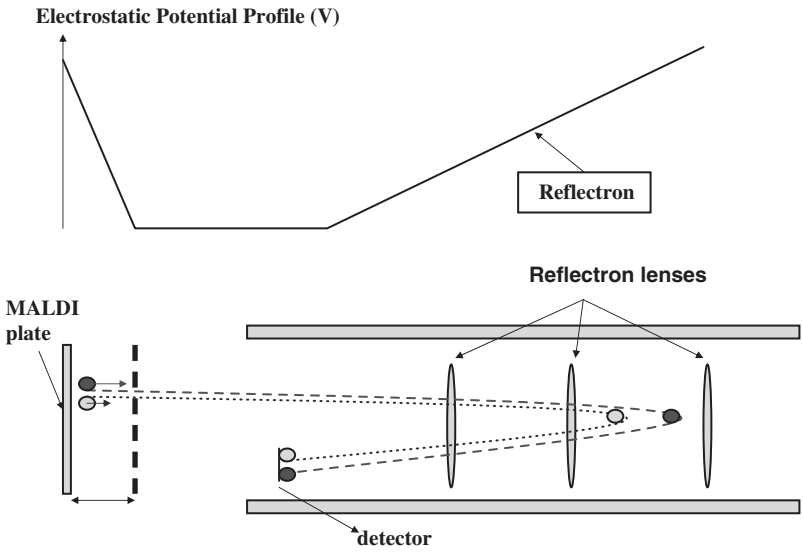


Fig. 17. Schematic representation of a reflectron (ion mirror) in a TOF analyzer: Ions with the same *m/z* formed with higher initial velocity (- - -) penetrate deeper in the electrostatic field of the reflectron than those with lower initial velocity (· · ·).

that MALDI-TOF and Q-TOF instruments are widely used for high-throughput analysis of samples of biological origin.

5.2. Interaction with electrostatic fields: electrostatic (ESA) and orbitrap (OT) analyzers

The simplest analyzer based on interaction with an electrostatic field is the ESA that separates the ions according to their *kinetic-energy-to-charge ratio* (E_{kin}/zq). This analyzer is historically used in sector instruments either (i) in front of the magnet to focus the ion beam leaving the ionization source (EB instruments) or (ii) after the magnet to detect, e.g., the kinetic energy of fragments of a selected ion (mass-selected ion kinetic energy spectra (MIKES)).

A recent development is the so-called "orbitrap" mass analyzer developed by Makarov and colleagues [40,41]. The potential distribution of the electrostatic field is a combination of a quadrupole and a logarithmic potential. The ion motion in such a field is quite complex, yet it is a well-defined oscillating motion along the axial electrode (Fig. 18). The frequency of this motion, which is related to the square root of the m/z ratio, can be measured with high accuracy so that the OT analyzer is one of the high-resolution analyzers. If ions with different m/z ratios are present, the measured signal can be deconvoluted with the Fourier transformation (FT) technique (see also the ICR analyzer in this section). At the present stage of development, OT analyzers are mostly used for exact mass measurements and as a second-stage mass analyzer in tandem MS/MS experiments (such as in a linear ion-trap/orbitrap combination).

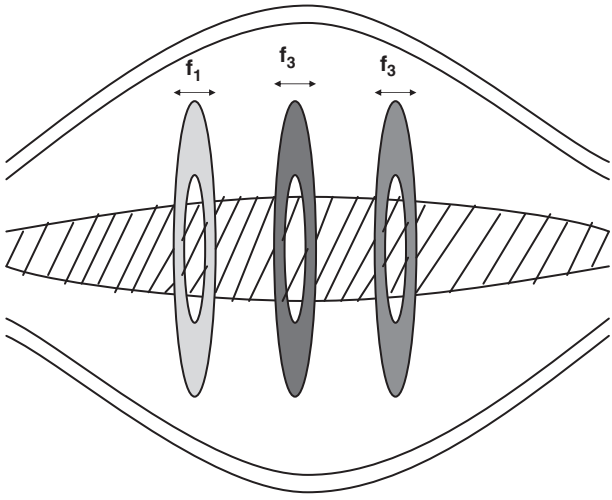


Fig. 18. Orbitrap mass analyzer: The rings are associated with ions with different m/z ratios, and they oscillate with m/z-related frequencies (f_i) along the axial electrode.

5.3. Interactions with a magnetic field: magnetic (B) analyzer

Historically, the magnetic analyzer was the first mass analyzer used, e.g., by Thomson and Aston to separate isotopes of elements in the beginning of the twentieth century. The principle of separation is based on the Lorenzian force that acts on a charged particle in a magnetic field (B) as a centripetal force so that a circular motion with a radius of r is generated (Fig. 19):

$$zq[\mathbf{v} \times \mathbf{B}] = \frac{mv^2}{r} \tag{18}$$

Equivalently, Equation (18) can be written as:

$$Br = \frac{mv}{zq} \tag{19}$$

Equation (19) indicates that the magnet separates ions according to their *momentum-to-charge* ratio.

The instrument geometry is fixed (i.e., r is constant) so that an ion with a given m/z can be detected at a given and well-defined B. By changing the magnetic field in time, a mass spectrum with a defined m/z range can be obtained. By incorporating the acceleration voltage (V) and so the kinetic energy of an ion into Equation (18), the mass-to-charge ratio can be written as:

$$\frac{m}{z} = \frac{B^2 r^2}{2V} \tag{20}$$

Even though magnetic analyzers are now considered outdated and are not widely used, it should be acknowledged that they played an important role in establishing mass spectral fragmentation rules and also the theory of mass spectra (see, e.g., ref. [13]). The fading glory of magnetic sector mass analyzers is related to the fact that they are just too bulky and cannot compete with other high-resolution analyzers such as the OT and ICR mass analyzers.

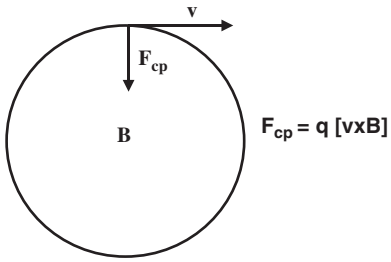


Fig. 19. Circular motion of a charged species generated by a centripetal force ($\boldsymbol{F}_{\mathbf{cp}}$) in a magnetic field. **B** is perpendicular to the plane.

5.4. Interactions with a magnetic and electrostatic field: ion cyclotron resonance (ICR) analyzers

Similarly to the magnetic analyzer, ion separation is based on the circular motion of a charged species in a magnetic field in ICR instruments as well. The difference is that in the ICR, ions undergo several full cycles in the ICR cell. Other differences, such as the application of electrostatic trapping fields, are also crucial and the ion motion is, in fact, much more complex than implied by the simplified discussion below. For the technically inclined reader, we recommend the book by Marshall and Verdun [42]. If only ions with the same m/z ratios are present, such as indicated in Fig. 20, ion motion can be related to a regular sine function, the frequency of which (ω) is inversely proportional to the m/z ratio:

$$\omega = \left(\frac{q}{m}\right)B \tag{21}$$

Note that Equation (21) is just a different representation of Equation (18).

It is easy to picture that if ions with different m/z ratios are present, circular motions with different frequencies (ω_i) are detected. Therefore, the detected signal will be a combination of sine functions with different ω_i frequencies (i.e., m/z ratios) and amplitudes (A_i) that are related to ion intensities. Such a complex signal is shown in Fig. 21a. This signal is then deconvoluted by using the well-known Fourier transformation and the corresponding mass spectrum is obtained (Fig. 21b). Owing to this "relationship" between an ICR signal and Fourier transformation, the term FT-ICR instruments is often applied.

Note that in contrast to a "sweeping" detection of ions in magnetic analyzers (i.e., when the magnetic field is changed in time), the FT-ICR technique detects all

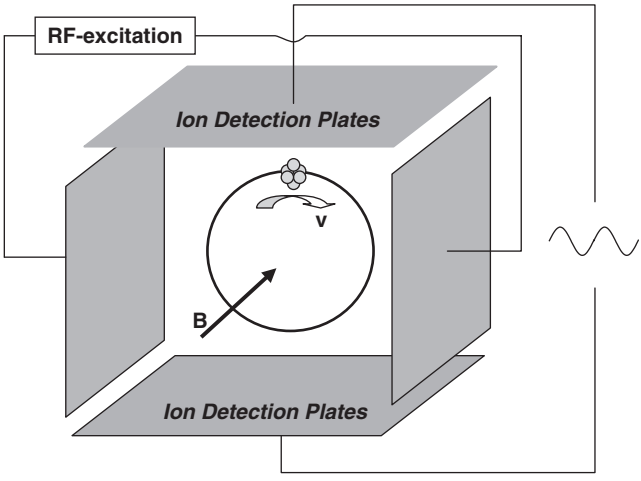


Fig. 20. Coherent circular motion of ions with the same m/z ratio in a cubic ICR cell results in a "pure" sinusoidal signal. For a better view, the planes perpendicular to the magnetic field are omitted.

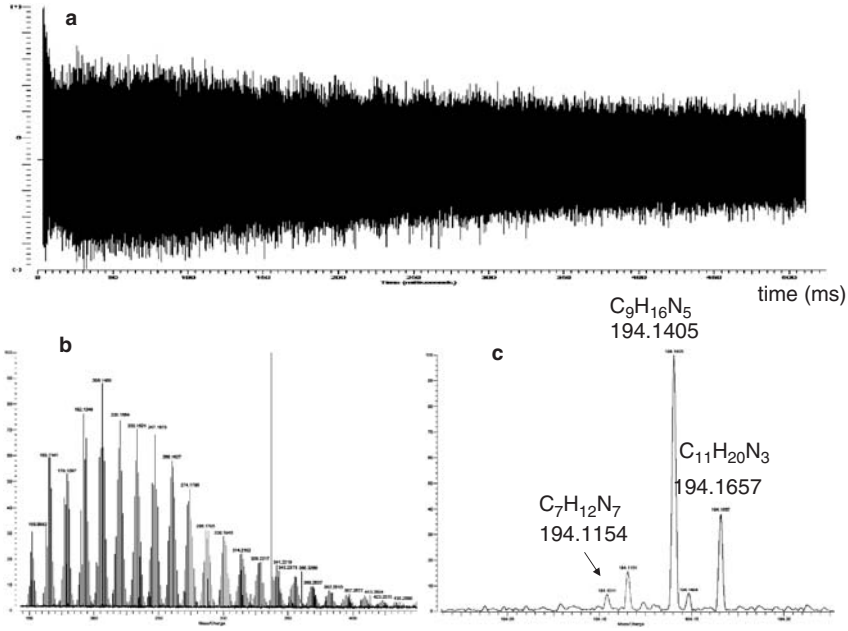


Fig. 21. (a) A detected 500 ms transient signal in an ICR cell, and (b) the corresponding mass spectrum obtained by Fourier transformation, and (c) a small part of the spectrum around the nominal mass *m*/*z* 194, showing five different ions separated easily by the ultrahigh resolving power of the FT-ICR.

ions at the same time with high resolution and mass accuracy. A good analogy of an FT-ICR analysis is the identification of individuals outside a lecture room by recording the noise behind a closed door with a microphone (or simply just by listening). Assuming that everybody talks at the same time, deconvolution of the noise from the room could lead to the unambiguous identification of every individual from the outside (knowing, of course, their characteristic "voice" frequencies). In this respect, our ears are our best Fourier transformers. Obviously, the longer we listen, the greater the reliability of identification. This is valid for ion-signal detection as well: Longer transients in time provide us better mass resolution. To achieve longer detection time, one should maintain the ion trajectories close to the detection plates, i.e., we should maintain ion velocities (v) unchanged for a reasonable amount of time. Ions can lose their velocities by colliding with residual gas molecules in the cell. Ultrahigh vacuum is, therefore, a requirement to achieve ultrahigh resolution. Another requirement is a strong, stabile, and homogenous magnetic field that is maintained by superconductive magnets. These requirements make FT-ICR instruments more expensive than most other instruments.

The main power of FT-ICR instruments lies in the ultrahigh resolution and mass accuracy. Very complex mixtures, such as crude oil samples and protein digests (peptide mixtures), can be easily analyzed without separating the individual components

prior to mass analysis. Different ion-activation methods are also easily applicable so that FT-ICR is widely used in tandem mass spectrometry (see Section 6).

5.5. Interaction with electromagnetic fields: quadrupole (Q) analyzers

As the name indicates, quadrupole mass analyzers consist of four parallel rods just as indicated in Fig. 22.

In a quadrupole mass analyzer, direct current (DC) and alternate current (AC) voltages are applied to the rods in such a way that two opposite rods have the same voltage, while the perpendicular ones have a voltage with opposite sign (+ and −, respectively). To be able to interact with this vibrating electromagnetic field in between the rods, the ions should enter the quadrupolar field with low velocity (e.g., with a few eV kinetic energy). Consequently, no high voltage (HV) is necessary to accelerate the ions before the mass analysis. This is particularly useful when ESI or APCI is used as there is relatively high pressure in the source region (which may result in HV discharge).

Two opposite rods have positive potential except for a short period of time when the negative RF voltage exceeds the positive DC voltage (Fig. 22). Because of this short period of negative potential, only the lighter positively charged ions will be defocused by these rods. This also means that these rods will focus only relatively large positively charged ions so that they can be considered as high mass filters. With similar considerations, the other two rods will defocus the positively charged ions most of the time and will focus only the lighter (positively charged) ions during a short

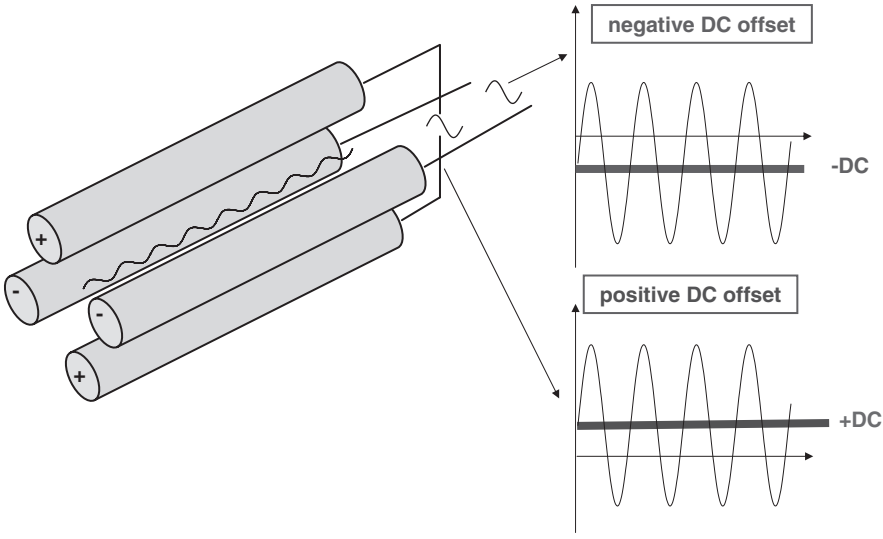


Fig. 22. Schematic representation of a quadrupole mass analyzer and a voltage profile on the rods. At a particular AC and DC voltages, only the ion with a given *m*/*z* passes through the quadrupole field.

period while the voltage is positive on these rods (Fig. 22). Thus, the other pair of rods works as low-mass filter. By applying the AC and DC voltages to all of the rods a band filter is created, i.e., a filter that allows an ion with a given m/z to pass through the rods and, subsequently, to reach the detector. (Notice that in this case the ions are ejected axially from the quadrupolar field.) By changing the absolute values of the AC and DC voltages, but keeping their ratio constant, the mass spectrum can be acquired. Note that in the so-called "RF only" operation, no DC voltage is applied to the rods. In this case, all ions with higher m/z than the low mass cutoff will pass the quadrupole analyzer. This RF-only operation is quite often used to focus an ion beam consisting of ions with different m/z ratios. For more details of the operating concept of the quadrupole analyzer, see, e.g., the well-written paper by Miller and Denton [43].

Quadrupole mass analyzers have several advantages such as no requirement for very high vacuum ($>10^{-7}$ Torr), and their relatively fast and simple operation for high-throughput analysis. Disadvantages include low transmittance, a low m/z cutoff, and low (generally unit) resolution.

5.6. Interaction with electromagnetic fields: linear ion-trap quadrupole (LTQ) analyzers

As discussed earlier, in the "conventional" way of quadrupole operation the ions are not trapped in between the rods but fly alongside them. However, it is also possible to trap ions in between the quadrupole rods for a certain amount of time and detect them by radial ejection (Fig. 23) [44]. The relatively large volume of ion

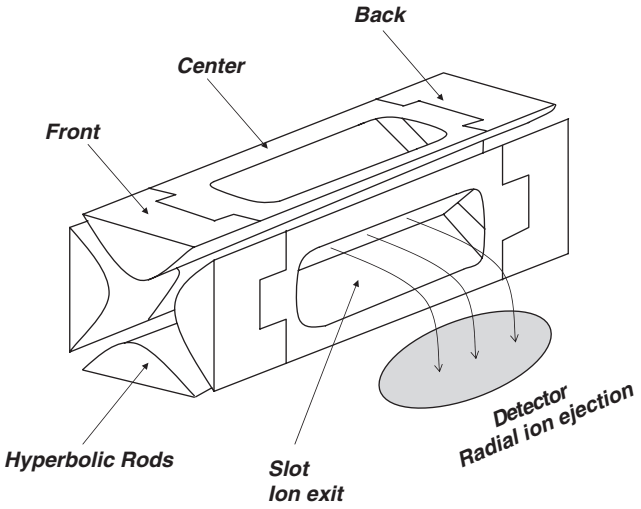


Fig. 23. A schematic representation of a linear ion-trap mass analyzer (LTQ). Ions are stored in between the quadrupole rods by applying trapping potentials at front and back sections. After some time, the ions are then ejected radially to detect them by two parallel detectors (Only one of them is shown for clarity).

storage allows more ions to be trapped than in a conventional 3D-IT instrument (see the following text). This, together with the axial ion detection, significantly increases the sensitivity of ion detection. Therefore, LTQ mass analyzers are more and more often used in areas where sensitivity is a crucial issue, such as pharmacokinetics and proteomics (including, e.g., posttranslational modification studies).

5.7. Interaction with electromagnetic fields: three-dimensional quadrupole ion trap (3D QIT) analyzers

Historically, the 3D QIT analyzers have been developed before the linear ion traps. Their invention paved the road to small, bench-top mass spectrometers. The operational principle of a QIT is similar to that of the quadrupole even though their physical appearances are quite different. 3D QIT analyzers consist of three main parts: the end cap, the entrance cap, and the inner ring (doughnut) (Fig. 24). By applying a RF field the ions can be oscillated in the trap (see a simplified illustration in Fig. 24). The ion trajectories are stabilized by a buffer gas (most often He). By ramping the voltage, the ions are ejected out of the trap through the exit hole. Intuitively, it is easy to predict that this analyzer can store less ions and more ions are lost during the ejection than in the LTQ instruments. Indeed, the 3D QIT is less sensitive than the LTQ.

The 3D QIT instruments have played and still play a revolutionary role in high-throughput mass spectral analyses. They are literally "work horses" that can operate in a "24/7" mode. Instrument maintenance is easy and not time-consuming. One disadvantage is that usually only unit resolution is achievable, but this drawback is overshadowed by the easy use for tandem MS/MS experiments (i.e., structural investigation, including, e.g., peptide sequencing that is fundamental for proteomics studies).

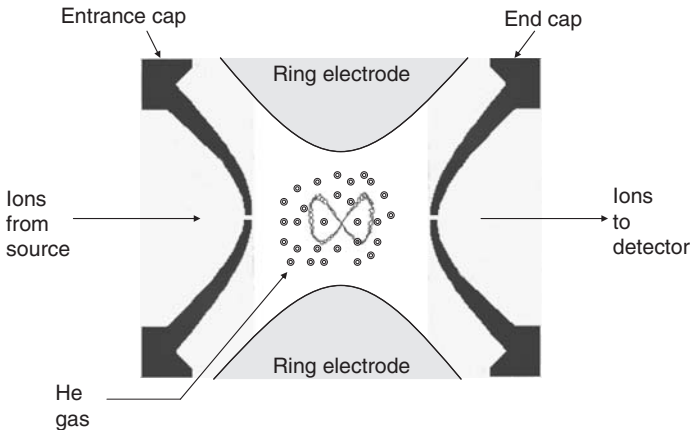


Fig. 24. Main components of a 3D QIT and a characteristic ion trajectory in the trap.

6. Tandem mass spectrometry (MS/MS)

Why do we need tandem mass spectrometry (MS/MS)? Is single stage MS not satisfactory? The rapid development of tandem MS/MS techniques have been triggered by the introduction of soft ionization techniques (most importantly ESI, nano-ESI, and MALDI). Soft ionization techniques usually provide intact, nonfragmenting ions that are crucial for molecular mass determination. However, a molecular mass (even an accurate one) does not provide enough information about the structure of the compound, simply because of the possible existence of structural isomers. For example, peptides of YAGFL and AFGLY have exactly the same MW, yet they differ significantly in their sequence.

Tandem mass spectrometry is an invaluable analytical technique to obtain structural information on originally stable, nonfragmenting ions. The main difference between a single-stage MS and tandem MS/MS is illustrated in Fig. 25. In the regular MS mode, ions formed in the ionization source are separated by a single-stage mass analyzer. The problem is that either the ions originating from the source may represent molecular ions of certain components of a mixture or some of the lower *m*/*z* ions can be fragments of ions of larger *m*/*z* ratio (i.e., they can be in a precursor–fragment relation). Even if a separation technique (GC or HPLC) is used prior to ionization, *coelution* may occur so that ions formed in the source at the same (retention) time may represent different components of a mixture. With tandem mass spectrometry, however, any individual ion can be selected and then activated to generate fragments of the selected ion. These fragments are characteristic for the precursor ion structure. The fragments originating exclusively from the precursor ion can then be analyzed separately with another mass analyzer. In short, there are three main steps in tandem mass spectrometry: (i) ion selection, (ii) ion activation (fragmentation), and (iii) analysis of the fragments of the selected ion.

There are several tandem MS/MS instrument types available commercially. A detailed overview of these instrumentations and techniques is beyond

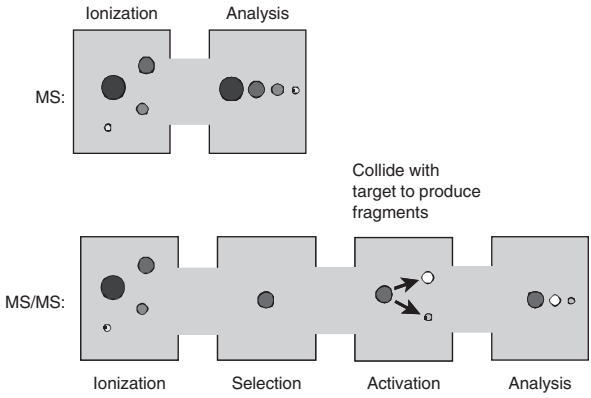


Fig. 25. Single-stage MS and tandem MS/MS.

the scope of the book, but a brief classification and summary are provided here.

Tandem MS/MS experiments can be performed either (i) *in space* (just as illustrated in Fig. 25) or (ii) consecutively *in time* in trapping analyzers. In the latter case, every step of a tandem mass spectrometry measurement (ion selection, ion activation, and fragment analysis) occurs in the same trap (i.e., same space but at different time). The great advantage of trapping instruments is that these steps can be consecutively repeated many times so that we can get structural information on the second, third, and, in general, the *n*th generations of fragments (MS^n techniques).

Depending on the resolving power of the first mass analyzer, ions can be selected either monoisotopically or with multiple isotopes. For practical reasons, and to improve sensitivity, multiple isotope selections (i.e., 2–3 *m*/*z* units) are preferred, especially for automated runs.

The main purpose of the ion-activation step is to provide additional internal energy to the originally "cold" ions that have not enough internal (vibrational) energy to fragment within the timescale of the instrument. As summarized in Table 1, this goal can be achieved in different ways, but all of them are related to

Table 1
A brief summary of ion-activation methods commonly used in tandem mass spectrometry experiments

Activation partner	Modes and instruments	Number of collisions	Amount of internal energy deposited and dominant fragmentation processes
Gas (He, Ar, Xe)	Low energy (eV)		
	QQQ	Several	Medium energy
	QIT, LTQ	Multiple	Low energy
	Sustained off-resonance irradiation (SORI-CID) in FT-ICR	Multiple	Low to medium energy
	High energy (keV)		
	TOF–TOF	Single	High energy
	Sector–TOF	Single	High energy
Photon	IRMPD (QIT, FT-ICR)	Multiple	Low energy
	BIRD (FT-ICR)	Multiple	Low energy
Low-energy elecron	ECD (FT-ICR)	Single	Low energy
Low-energy anion	ETD (LTQ)	Single	Low energy
Surface	eV SID	Single	High energy, with relatively narrow distribution
	Q-SID-Q		
	Sector–TOF		
	FT-ICR		
	keV SID = SIMS		

collisions of the selected ions. The colliding partner most often is a gas (gas-phase collision-induced dissociation, CID [5,45], including sustained off-resonance irradiation-CID, SORI-CID in FT-ICR instruments [46]), but it may be a photon (photodissociation [47], infrared multiphoton dissociation (IRMPD) [48–51], or blackbody irradiative dissociation (BIRD) [52–53]), a low-energy electron (electron capture dissociation (ECD) [54–56]), a negatively charged anion (electron transfer dissociation (ETD) [57]), or a surface (surface-induced dissociation, SID [58–60]). The extent of fragmentation depends on the amount and distribution of internal energy deposited to the selected ions: As intuitively expected, more internal energy triggers more extensive fragmentation. For further details of the internal energy distribution and its influence on fragmentation, see, e.g., the books by Cooks et al. [13], Forst [14], and Beynon and Gilbert [12], as well as a detailed tutorial by Vékey [15].

Ion activation can also be classified as low- (eV) or high-energy (keV) collisions. In this case, the laboratory collision energy is used for guidance, but it should be noted that only the so-called "center of mass" energy is available for the kinetic-to-internal-energy ($T \rightarrow V$) transfer. Another way of grouping ion-activation methods is to consider the number of collisions so that we can talk about single or multiple collision conditions. These are also indicated in Table 1.

The final appearance of a tandem MS/MS spectrum depends not only on the mode of ion activation (high vs. low internal energy deposition) but also on the time lag between ion activation and the recording of the fragmentation spectrum. This can easily be understood if one accepts that ion fragmentation is assumed to be unimolecular after ion activation: As expected, a longer time gap leads to more fragments. The main practical conclusion is that tandem MS/MS spectra of the same precursor ion can be quite different if they are acquired in different instrument configurations. This is well illustrated in Fig. 26 for protonated *N*-acetyl OMe proline [61]. In an ion-trap instrument, where the internal energy is deposited in small increments by multiple collisions, only low-energy processes, such as the loss of methanol (CH_3OH) and a subsequent loss of CO, are observed. On the contrary, when the internal energy is deposited in one step (CID in QQQ and SID in Q-TOF), the high-energy process of the ketene loss becomes a competitive channel and the corresponding fragment ion at *m*/*z* 130 is clearly detected. Despite this dependence on experimental conditions, tandem MS/MS spectra can be reasonably compared if obtained under similar instrumental conditions.

The tandem MS/MS spectra shown in Fig. 26 are typical "product ion" spectra. In most tandem MS/MS applications this scan mode is used to obtain structural information of a selected (precursor) ion. A variation of product ion scans are used also in multiple reaction monitoring (MRM), which is a useful technique for *quantitation* and kinetic studies (see the following text). Other MS/MS scan types are also applied even though not all of them are easily available in all tandem MS/MS

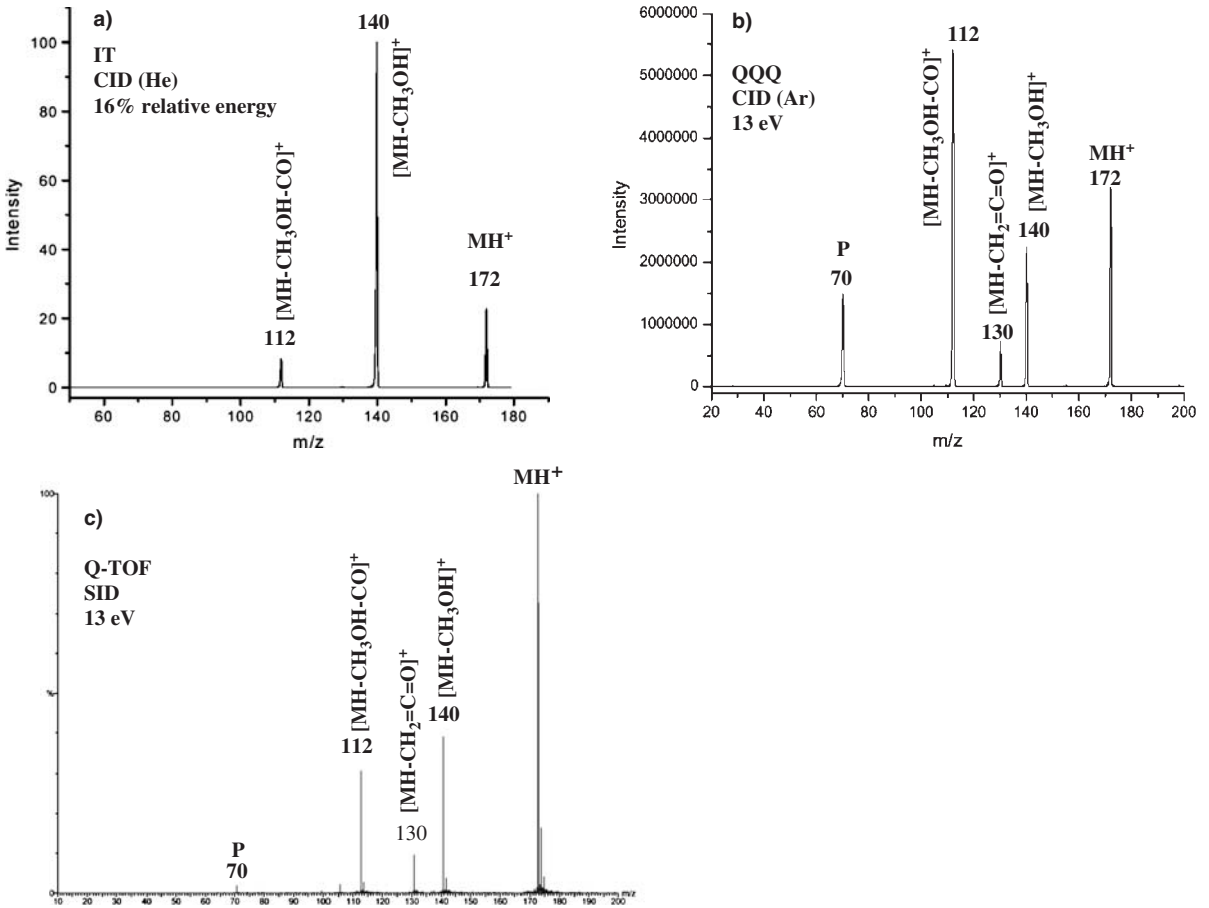


Fig. 26. Tandem MS/MS spectra of protonated *N*-acetyl OMe proline obtained using different ion-activation methods and instruments, such as gas-phase collisional activation in (a) a Thermoelectron (Finnigan) LCQ classic 3D QIT instrument, (b) a Thermoelectron (Finnigan) triple quadrupole (QQQ) instrument, and (c) surface-induced dissociation (SID) in a Micromass Q-TOF instrument.

Table 2
MS/MS scan modes applicable in a triple quadrupole instrument (QQQ)

Scan modes	Quadrupole 1 (Q1)	Quadrupole 2 (Q2)	Quadrupole 3 (Q3)
Product ion	Select a desired *m*/*z*	Ion activation/ dissociation	Scan for fragments of *m*/*z*
Precursor ion	Scan for parents of a given fragment, F	Ion activation/ dissociation	Select and monitor fragment, F
Neutral loss	Scan	Ion activation/ dissociation	Scan with shift of mass of the neutral
Selected/multiple reaction monitoring (SRM or MRM)	Select a desired *m*/*z*	Ion activation/ dissociation	Select and monitor desired fragment(s)
Ion–molecule	Select a desired *m*/*z*	Ion–molecule reactions	Scan for reaction products

instruments. For easier understanding, these scan modes are summarized in Table 2 for a triple quadrupole (QQQ) instrument.

In precursor ion scan mode, all precursors that form a given fragment ion (F) are detected. In this mode, the second quadrupole is set to the given *m*/*z* value of the fragment ion and the first quadrupole is scanned so that the software can reconstruct the precursor ion spectrum. In other words, the software "finds" those precursor ions from which a given fragment is formed. In the neutral loss scan both the first and the second quadrupoles are scanned in a synchronized way, i.e., their scans are "shifted" by the desired mass of the neutrals (e.g., by 18 for water or by 80 (HPO_3) for protein phosphorylation studies). Neutral loss measurements are, therefore, especially useful for detection of laboratory (or natural) modifications of an analyte. If higher sensitivity is desired, the second quadrupole is not scanned over a wide mass range of the fragments but, instead, it is set up to monitor only a selected fragment or fragments (SRM or MRM scans). Finally, we note that tandem mass spectrometry can also be used for studying ion–molecule reactions, i.e., when a selected ion reacts in the "collision cell" (second quadrupole, or in general, in any trapping analyzer). Specific examples for ion–molecule reactions include hydrogen/deuterium (H/D) exchange studies or reactions of multiply charged (positive) ions with anions in a linear ion trap (charge transfer dissociation (CTD)). These ion–molecule reactions are particularly useful to distinguish between structural isomers of ions with the same chemical formula.

Throughout this book the readers can find beautiful examples for the application of tandem MS/MS. Although many good instruments are available commercially, there is no "ideal" or "perfect" instrument for *general* use. Depending on the application, different instruments may have ideal performance. We encourage the reader to carefully choose an instrument that is the best for his/her needs. Hopefully, the information and discussion presented in this chapter will help the reader in making the best choice.

7. Selected terms for clarification

The author of this chapter takes the freedom to explain a few selected terms that—based on his teaching experience—are often not well understood and may even be misused. Although this is not an exhaustive list, clarification of these terms may help in reading the whole book more comprehensively by readers whose primary expertise is not in mass spectrometry. For a detailed guidance of terminology, we refer again to the book by Sparkman [1].

Nominal mass is the integer mass of the most abundant naturally occurring stable isotope of an element. As a consequence, the nominal mass of an ion is the sum of the nominal masses of the elements in the empirical formula (for the acetone molecular ion the formula is $C_3H_6O^{+\cdot}$; thus, the nominal mass is 58). This, sometimes, is mistaken with the *average* molecular mass that is also commonly used, e.g., by synthetic chemists. In the average molecular mass, average atomic masses are used: the accurate atomic masses of various isotopes, weighted by their natural abundance. For example, the average atomic weights of chlorine and bromine are 35.5 and 80 Da, respectively. Nonetheless, in mass spectral analysis of common organic molecules these average masses are never measured. Instead, each isotope is observed, in the case of the above-mentioned case of chlorine an ion pair corresponding to ^{35}Cl and ^{37}Cl, while in the case of bromine those related to ^{79}Br and ^{81}Br.

Accurate mass is the experimentally measured mass of an ion that is precise enough to determine its elemental composition, e.g., has at least ± 5 ppm accuracy. This accuracy can easily be achieved in the *m*/*z* range of 1–3000 Da by using various commercially available mass spectrometers, such as sectors, TOF, QTOF, OT, or FT-ICR instruments. For accurate mass measurements, appropriately chosen internal standards are required. In a general procedure, the analyte ion is bracketed by two internal standard ions, the *m*/*z* values of which are known with very high accuracy. The spectrum obtained with the internal standard is then recalibrated by using the standard *m*/*z* values of the internal standard ions ("peak matching" technique). A wide variety of internal standards can be used. The most commonly applied ones include perfluoro kerosene (PFK) for EI ionization and polyethylene (PEG) or polypropylene (PPG) glycol in FAB (LSIMS), ESI, and MALDI measurements. The use of peptide internal standards is also common in accurate mass measurements by ESI and MALDI.

Although they are related, the *measured accurate mass* should be distinguished from the *calculated exact mass,* which is the mass determined by summing the exact isotope masses of the elements present in a particular ion. For example, the calculated exact mass of $^{12}C_3{}^{1}H_6{}^{16}O^{+\cdot}$ (acetone molecular ion) is 58.0419 Da, and a measured accurate mass could, for example, be 58.0416 Da, which corresponds to a -5.2 ppm error. In this calculation, the following atomic (isotopic) exact masses are used: ^{12}C: 12.0000 Da, ^{1}H: 1.007825 Da, and ^{16}O: 15.9949 Da. (For

accurate isotope masses and an exact mass calculator program, see, e.g., the web sites www.sisweb.com/referenc/source/exactmaa.htm and www.sisweb.com/referenc/tools/exactmass.htm, respectively.)

Fig. 27 illustrates the differences between nominal, exact, and average masses for singly protonated alanine oligomers: $[(Ala)_5 + H]^+$ and $[(Ala)_{50} + H]^+$. When the nominal ion mass is relatively small (e.g., around 400 Da), the nominal, exact, and average masses do not differ significantly (Fig. 27a and c). However, with increasing masses, differences between the nominal, exact, and average masses become more and more significant that should be accounted for (Fig. 27b and c). If, for example, the resolution of a mass spectrometer is not good enough to separate individual isotopes, the measured peak will provide only an "envelope" of the isotope distribution from which the *average* mass of the ion can be approximately determined (see solid curve line in Fig. 27b).

Resolution of a mass spectrometer is related to the ability of a mass analyzer to separate two ions with different *m/z* ratios. The resolution is defined as $R = M/\triangle M$, where M is a given mass and $\triangle M$ is the difference between the given mass and the neighboring mass peak with, for example, 10% peak height (see Fig. 28).

The terms "low resolution" and "high resolution" are often misused meaning "not accurate mass/survey" and "accurate mass" measurements, respectively. High

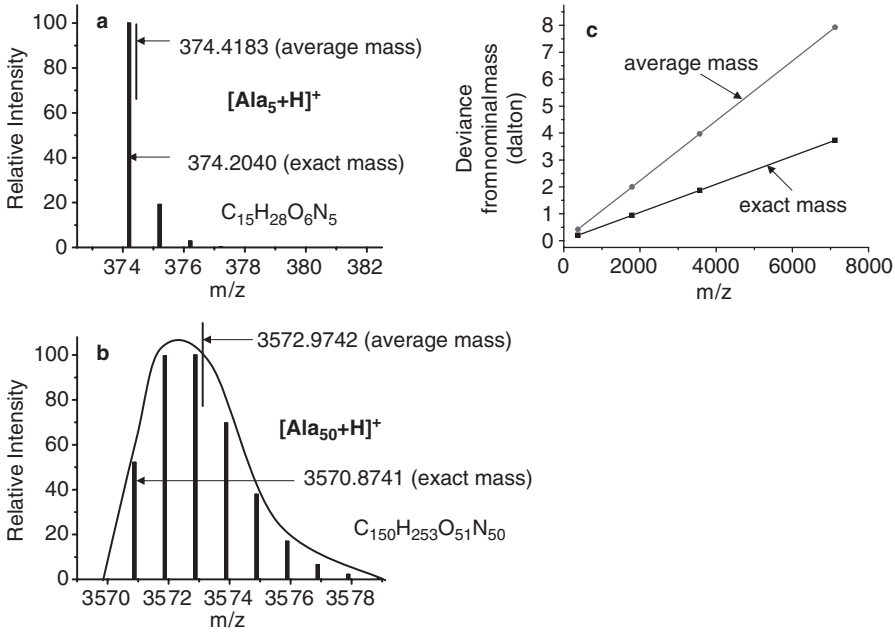


Fig. 27. Calculated isotope pattern distribution of singly protonated (a) Ala_5, and (b) Ala_{50}, and (c) the calculated deviance between the nominal mass, the exact mass of the first isotope peak, and the average mass for singly protonated polyalanines as a function of the nominal mass.

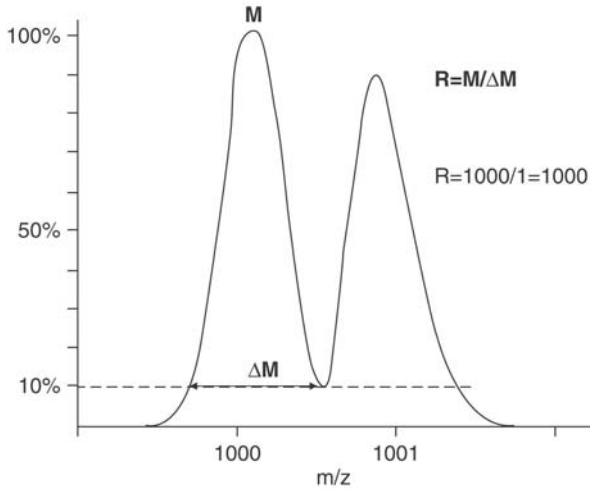


Fig. 28. Resolution of a mass spectrometer: The presented ion separation is related to a resolution of 1000 (at 10% valley).

resolution is only a prerequisite for accurate mass measurements in which the use of internal standards with precisely known ion masses and a stable mass scale or calibration of the instrument are required.

Acknowledgment

The author greatly appreciates the careful reading and comments by Drs. Ronald Wysocki (Director of Chemical Synthesis Laboratory, Department of Chemistry, University of Arizona) and Linda Breci (Associate Director of the Proteomics Consortium, Department of Chemistry, University of Arizona).

References

1. Sparkman, O.D., *Mass Spectrometry Desk Reference.* Global View Publishing, Pittsburg, PA, 2000.
2. Watson, J.T., *Introduction to Mass Spectrometry* (3rd edition). Lippincott-Raven Publishers, Philadelphia, New York, 1997.
3. Chapman, J.R., *Practical Organic Mass Spectrometry* (2nd edition). John Wiley & Sons, New York, 1985.
4. Gross, J.H., *Mass Spectrometry. A Textbook.* Springer Verlag, Berlin, Heidelberg, New York, 2004.
5. Busch, K.L., Glish, G.L. and McLuckey, S.A., *Mass Spectrometry/Mass Spectrometry: Techniques and Applications of Tandem Mass Spectrometry.* VCH Publishers, Berlin, 1989.
6. McLafferty, F.W. and Tureček, F., *Interpretation of Mass Spectra* (4th edition). University Science Books, Mill Valley, CA, 1993.

7. Siuzdak, G., *Mass Spectrometry for Biotechnology.* Academic Press, San Diego, CA, 1996.
8. Liebler, D.C., Tools for the New Biology. *Introduction to Proteomics.* Humana Press, Totova, NJ, 2002.
9. Baer, T., Ng, C.-Y. and Powis, I. (Eds.), *Large Ions: Their Vaporization. Detection and Structural Analysis.* John Wiley & Sons, Chichester, UK, 1996.
10. McMaster, M. and McMaster, C., *GC/MS: A Practical User's Guide.* Wiley-VHC, New York, NY, 1998.
11. Ardrey, R.E., *Liquid Chromatrography Mass Spectrometry: An Introduction.* John Wiley & Sons, Chichester, UK, 2003.
12. Beynon, J.H. and Gilbert, J.R., *Application to Transition State Theory to Unimolecular Reactions. An Introduction.* John Wiley & Sons, Chichester, UK; New York, NY, 1984.
13. Cooks, R.G., Beynon, J.H., Caprioli, R.M. and Lester, G.R., *Metastable Ions.* Elsevier, Amsterdam, 1973.
14. Forst, W., *Theory of Unimolecular Reactions.* Academic Press, London, UK, 1973.
15. Vékey, K., Internal energy effects in mass spectrometry. *J. Mass Spectrom.*, **31**, 445–463 (1998).
16. Drahos, L. and Vékey, K., MassKinetics: a theoretical model of mass spectra incorporating physical processes, reaction kinetics and mathematical descriptions. *J. Mass Spectrom*., **36**, 237–263 (2001).
17. Harrison, A.G., *Chemical Ionization Mass Spectrometry.* CRC Press, Boca Raton, FL, 1983.
18. Meot-Ner (Mautner), M., The proton affinity scale, and effects of ion structure and salvation. *Int. J. Mass Spectrom.*, **227**, 525–554 (2003).
19. Fenselau, C. and Cotter, R.J., Chemical aspects of fast atom bombardment. *Chem. Rev.*, **87**, 501–512 (1987).
20. Doyle, M., Mack, L.L., Hines, R.L., Mobley, R.C., Ferguson, L.D. and Alice, M.B., Molecular beams of macroions. *J. Chem. Phys.*, **49**, 2240–2249 (1968).
21. Fenn, J.B., Mann, M., Meng, C.L., Wong, S.F. and Whitehouse, C.M., Electrospray ionization for mass-spectrometry of large biomolecules. *Science*, **246**, 64–71 (1989).
22. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse, C.M., Electrospray ionization—principles and practice. *Mass Spectrom. Rev.*, **9**, 37–70 (1990).
23. Cole, R. (Ed.), *Electrospray Ionization Mass Spectrometry: Fundamentals, Instrumentation, and Applications.* John Wiley & Sons, New York, NY, 1997.
24. Znamenskiy, V., Marginean, I. and Vértes, A., Solvated ion evaporation from charged water nanodroplets. *J. Phys. Chem. A*, **107**, 7406–7412 (2003).
25. Marginean, I., Nemes, P., Parvin, L., and Vértes, A., How much charge is there on a pulsating Taylor cone? *Appl. Phys. Lett.*, **89**, 064104 (2006).
26. Marginean, I., Nemes, P. and Vértes, A., Order-chaos-order transitions in electrosprays: the electrified dripping faucet. *Phys. Rev. Lett.*, **97**, 064502 (2006).
27. McKay, A.R., Ruotolo, B.T., Ilag, L.L. and Robinson, C.V. Mass measurements of increased accuracy resolve heterogeneous populations of intact ribosomes. *J. Am. Chem. Soc.*, **128**, 11433–11442 (2006).
28. Jones, C.M., Beardsley, R.L., Galhena, A.S., Dagan, S., Cheng, G. and Wysocki, V.H., Symmetrical gas-phase dissociation of noncovalent protein complexes via surface collisions. *J. Am. Chem. Soc.*, **128**, 15044–15045 (2006).
29. Takáts, Z., Wiseman, J.M., Gologan, B. and Cooks, R.G., Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science*, **306**, 471–473 (2004).
30. Cody, R.B., Laramée, J.A. and Durst, H.D., Versatile new ion source for the analysis of materials in open air under ambient conditions. *Anal. Chem.*, **77**, 2297–2302 (2005).
31. Wang, G., Hsieh, Y. and Korfmacher, W.A., Comparison of atmospheric pressure chemical ionization, electrospray ionization, and atmospheric pressure photoionization for the determination of cyclosporin A in rat plasma. *Anal. Chem.*, 77, 541–548 (2005).

32. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y. and Yoshida, T., Protein and polymer analyses up to m/z 100000 by laser desorption ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, **2**, 151–153 (1988).
33. Karas, M. and Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons. *Anal. Chem.*, **60**, 2299–2301 (1988).
34. Dashtiev, M., Frankevich, V. and Zenobi, R., Kinetic energy of free electrons affects MALDI positive ion yield via capture cross-section. *J. Phys. Chem. A*, **110**, 926–930 (2006).
35. Vértes, A., Luo, G., Ye, L., Chen, Y. and Marginean, I., Laser pulse length dependence of internal energy transfer in UV-MALDI-MS. *Appl. Phys. A*., **79**, 823–825 (2004).
36. Go, E.P., Apon, J.V., Luo, G., Saghatelian, A., Daniels, R.H., Sahi, V., Dubrow, R., Cravatt, B.F., Vértes, A. and Siuzdak, G., Desorption/ionization on silicon nanowires. *Anal. Chem.*, **77**, 1641–1646 (2005).
37. Wei, J., Buriak, J.M. and Siuzdak, G., Desorption/ionization mass spectrometry on porous silicon. *Nature*, **399**, 243–246 (1999).
38. Yanes, O., Nazabal, A., Wenzel, R., Zenobi, R. and Aviles, F.X., Detection of noncovalent complexes in biological samples by intensity fading and high-mass detection MALDI-TOF mass spectrometry. *J. Proteome Res.*, **5**, 2711–2719 (2006).
39. Cotter, R.J., *Time-of-Flight Mass Spectrometry: Instrumentation and Applications in Biological Research. (ACS Professional Reference Books)*. American Chemical Society, Washington D.C., 1997.
40. Makarov, A., Electrostatic axially harmonic orbital trapping: a high performance technique of mass analysis. *Anal. Chem.*, **72**, 1156–1162 (2000).
41. Hu, Q. and Noll, R.J., Li, H., Makarov, A., Hardman, M. and Cooks, R.G., The orbitrap: a new mass spectrometer. *J. Mass Spectrom.*, **40**, 430–443 (2005).
42. Marshall, A.G. and Verdun, F.R., *Fourier Transforms in NMR, Optical, and Mass Spectrometry*. Elsevier, Amsterdam, 1990.
43. Miller, P.E. and Denton, M.B., The quadrupole mass filter: basic operational concepts. *J. Chem. Ed.*, **63**, 617–623 (1986).
44. Hager, J.W., A new linear ion trap mass spectrometer. *Rapid Commun. Mass Spectrom.*, **16**, 512–526 (2002).
45. Senko, M.W., Speir, J.P. and McLafferty, F.W. Collisional activation of large multiply-charged ions using Fourier-transform mass-spectrometry. *Anal. Chem.*, **66**, 2801–2808 (1994).
46. Gauthier, J.W., Trautman, T.R. and Jacobson, D.B., Sustained off-resonance irradiation for collision-activated dissociation involving Fourier-transform mass-spectrometry—collision-activated dissociation technique that emulates infrared multiphoton dissociation. *Anal. Chim. Acta*, **246**, 211–225 (1991).
47. Bowers, W.D., Delbert, S.S. and McIver, R.T., Consecutive laser-induced photodissociation as a probe of ion structure. *Anal. Chem.*, **58**, 969–972 (1986).
48. Little, D.P., Speir, J.P., Senko, M.W., O'Connor, P.B. and McLafferty, F.W., Infrared multiphoton dissociation of large multiply-charged ions for biomolecule sequencing. *Anal. Chem.*, **66**, 2809–2815 (1994).
49. Woodin, R.L., Bomse, D.S. and Beauchamp, J.L., Multiphoton dissociation of molecules with low-power continuous wave infrared-laser radiation. *J. Am. Chem. Soc.*, **100**, 3248–3250 (1978).
50. Flora, J.W. and Muddiman, D.C., Determination of the relative energies of activation for the dissociation of aromatic versus aliphatic phosphopeptides by ESI-FT-ICR-MS and IRMPD. *J. Am. Soc. Mass Spectrom.*, **15**, 121–127 (2004).
51. Zhang, J., Schubothe, K., Li, B., Russell, S. and Lebrilla, C.B., Infrared multiphoton ionization of O-linked mucin-type olicosaccharides. *Anal. Chem.*, **77**, 208–214 (2005).
52. Dunbar, R.C. and McMahon, T.B., Activation of unimolecular reactions by ambient blackbody radiation. *Science*, **279**, 194–197 (1998).

53. Price, W.D., Schnier, P.D. and Williams, E.R., Tandem mass spectrometry of large biomolecule ions by blackbody infrared radiative dissociation. *Anal. Chem.*, **68**, 859–866 (1996).
54. McLafferty, F.W., Horn, D.M., Breuker, K., Ge, Y., Lewis, M.A., Cerda, B., Zubarev, R.A. and Carpenter, B.K., Electron capture dissociation of gaseous multiply charged ions by Fourier**-**transform ion cyclotron resonance. *J. Am. Soc. Mass Spectrom.*, **12**, 245–249 (2001).
55. Zubarev, R.A., Kelleher, N.L. and McLafferty, F.W., Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.*, **120**, 3265–3266 (1998).
56. Zubarev, R.A., Horn, D.M., Fridriksson, E.K., Kelleher, N.L., Kruger, N.A., Lewis, M.A., Carpenter, B.K. and McLafferty, F.W., Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal. Chem.*, **72**, 563–573 (2000).
57. Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J. and Hunt, D.F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Acad. Natl. Sci. U.S.A.*, **101**, 9528–9533 (2004).
58. Mabud, Md. A., DeKrey, M.J. and Cooks, R.G., surface-induced dissociation of molecular ions. *Int. J. Mass Spectrom. Ion Proc.*, **67**, 285–294 (1985).
59. Dongré, A.R., Somogyi, Á. and Wysocki, V.H., Surface-induced dissociation: an effective tool to probe structure, energetics and fragmentation mechanisms of protonated peptides. *J. Mass Spectrom.*, **31**, 339–350 (1996).
60. Laskin, J. and Futrell, J.H., Activation of large ions in FT-ICR mass spectrometry. *Mass Spectrom. Rev.*, **24**, 135–167 (2005).
61. Komáromi, I., Somogyi, Á. and Wysocki, V.H., Proton migration and its effect on the MS fragmentation of N-Acetyl OMe Proline: MS/MS experiments and ab initio and density functional calculations. *Int. J. Mass Spectrom.*, **241**, 315–323 (2005).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 7

Chemoinformatics—multivariate mathematical–statistical methods for data evaluation

KÁROLY HÉBERGER*

Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary

1. Introduction

Chemometrics, or chemoinformatics, was established at the beginning of the 1970s by Svante Wold, Bruce L. Kowalski, and D.L. Massart. The term ‘chemometrics’ was first coined by S. Wold, who applied for funding from the government of

*Tel.: +361 438 1103; Fax: +361 438 1143. E-mail: heberger@chemres.hu (K. Héberger).

Sweden and thought it would be much easier to receive it for a new discipline. Since then a lot of definitions of chemometrics have been proposed. We apply here the definition given in *Chemometrics and Intelligent Laboratory Systems*—the leading journal in the field:

> Chemometrics is the chemical discipline that uses mathematical, statistical, and other methods employing formal logic to design or select optimal measurement procedures and experiments, and to provide maximum relevant chemical information by analyzing chemical data.

Despite the advent of disciplines such as biometrics, chemometrics is not dying out; on the contrary, it achieved maturity around the millennium.

The features of the chemometric approach can perhaps be best understood by comparing it with the classical approach. The classical approach aims to understand effects—which factors are dominant and which ones are negligible—whereas the chemometric approach gives up the necessity to understand the effects, and points out other aims such as prediction, pattern recognition, classification, etc.

The classical approach is reductionist, one factor examined at a time; the effects are separated as much as possible. The chemometrics approach uses multivariate methods, i.e., all variables are considered at the same time. In this way, the model is fit to the data. When building a model to fit the data, the conclusions should be in harmony with the information present in the data. This is sharply different from the classical approach, where the model is derived from theory and the data are searched to show the validity of the model. For many scientists, the theory is the nonplus ultra; it cannot be criticized. They can rather measure what should be measured according to the theory. In this way, however, the conclusions drawn may be in contradiction with the information present in the data.

As a result, the classical approach determines new (causal) relationship(s) and discovers new natural laws, whereas the chemometric approach finds usually a formal relationship, which has the elements of causality. Moreover, prediction and classification are possible by applying these "formal" models.

Naturally the classical approach has the advantage of being successful, accepted, and well based; the constants in the models have definite physical significance. The disadvantage, however, is that the factors are correlated and their effects cannot be separated. Nature is not orthogonal unlike the mathematical description. The advantage of the chemometric approach is that correlations among variables can be utilized. The disadvantage is that the constants in models do not necessarily have physical relevance.

As can be seen, the two approaches are complementary. The modern, newer approach cannot be substituted by the older, classical one and *vice versa*. The chemometric approach simply provides information not otherwise accessible.

Two new ideas have to be introduced before going into details: prediction and pattern recognition.

Prediction means declaration in advance, especially foretelling on the basis of observation, experience, or scientific reason. Prediction concerns not only temporal processes but also, for example, the toxicity of a compound on the basis of similar compounds. Even if you do not have a causal model, prediction is valuable using black box models.

Pattern recognition is to unravel patterns in the data. Although patterns are perceived automatically, the process is difficult to define: A pattern is a natural or chance configuration, reliable sample of traits, tendencies, or other observable characteristics of data. In chemometrics, patterns are usually simplified to groupings (clusters) and outliers.

Consider the blood test of healthy and ill patients. If you consider one feature at a time, all features may be within the given limits of healthiness but still a patient might be ill. On the contrary, some of the healthy patients can provide extreme values. If you consider large number of patients and all features at once, usually the healthy and ill patients can be distinguished using multivariate chemometric methods.

2. Data types and data pretreatment

2.1. Data types

It is expedient to distinguish the variables on the basis of three scales: nominal, ordinal, and numeric.

The nominal scales are categorical in nature, i.e., qualitative only. They can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories. Each category is "different" from others but cannot be quantitatively compared to others. Two kinds of nominal scales are differentiated: binary and grouping scales. The binary scales can have only two values (e.g., yes or no, zero or one, ill or healthy, etc.). Group scales can have several categories (e.g., integer numbers or strings are assigned to groups, e.g., several types of cancer or seasonal differences—spring, summer, winter, etc.—can be distinguished).

The ordinal scales are also qualitative, but they can rank (order) the items measured in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say "how much more." A typical example of an ordinal variable is the toxicity: A compound can be classified as highly toxic, moderately toxic, hardly toxic, or nontoxic. Although the compounds can be ordered according to the toxicity, how much more toxic they are cannot be established. The ordinal scale provides more information than nominal scale, but "how much more" cannot be established.

The numerical scale is quantitative in nature. This scale of measurement allows us not only to rank the items that are measured but also to quantify and compare the sizes of differences between them. Some authors distinguish interval and ratio scales, whether an absolute zero point is defined or not, but this is not mandatory. For example, the temperature measured in Celsius is on an interval scale, whereas in Kelvin it is on a ratio scale. Interval scales do not have the ratio property.

2.2. Arrangement of data

One single number, called a *scalar,* is not appropriate for data analysis.

Vectors: A series of scalars can be arranged in a column or in a row. Then, they are called a column or a row vector. If the elements of a column vector can be attributed to special characteristics, e.g., to compounds, then data analysis can be completed. The chemical structures of compounds can be characterized with different "numbers" called descriptors, variables, predictors, or factors. For example, toxicity data were measured for a series of aromatic phenols. Their toxicity can be arranged in a column arbitrarily: Each row corresponds to a phenolic compound. A lot of descriptors can be calculated for each compound (e.g., molecular mass, van der Waals volume, polarity parameters, quantum chemical descriptors, etc.). After building a multivariate model (generally one variable cannot encode the toxicity properly) we will be able to predict toxicity values for phenolic compounds for which no toxicity has been measured yet. The above approach is generally called searching quantitative structure – activity relationships or simply QSAR approach.

Matrices: Column vectors when put one after one form a matrix. Generally two kinds of matrices can be distinguished denoted by $\boldsymbol{X}$ and $\boldsymbol{Y}$. The notation $\boldsymbol{X}$ is used for the matrix of independent variables. The notation $\boldsymbol{Y}$ is used for the matrix of dependent variables; their values are to be predicted. If we can arrange our data into one ($\boldsymbol{X}$) matrix, still we can unravel patterns in the data in an unsupervised way, i.e., we do not use the information of groupings present in the data. Such matrices are suitable for a principal component analysis (PCA).

Matrices (arrays) can be multidimensional; three-dimensional matrices are also called tensors. Analysis of tensors is frequently called *3-way analysis*. Typical example is the data from a hyphenated technique, e.g., gas chromatography–mass spectrometry (GC–MS) data; one direction (way) is the mass spectrum, second direction is the chromatographic separation (time, scan), and the third direction is the samples (of different origin, repetitions, calibration series, etc.). The 3-way analyses can easily be generalized into *n*-way analysis including more directions.

3-Way analyses require routine use of matrix operations; besides this, they can be unfolded into 2-way arrays (matrices). Therefore, we deal with analysis of matrices further on.

2.3. Data pretreatment

The data are arranged in a matrix form; the column vectors are called variables and the row vectors are called mathematical–statistical cases (objects or samples):

$$\underset{n,m}{X} = \begin{matrix} x_{11} & x_{12} & \text{L} & x_{1m} \\ x_{21} & x_{22} & \text{L} & x_{2m} \\ \text{M} & \text{M} & & \text{M} \\ x_{n1} & x_{n2} & \text{L} & x_{nm} \end{matrix} \tag{1}$$

with m columns and n rows, $j = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, n$, respectively.

Centering means to subtract the column averages from each matrix element:

$$x'_{ij} = x_{ij} - \bar{x}_j \tag{2}$$

Standardization means to divide each centered matrix element with the column standard deviations:

$$x''_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{3}$$

where x_{ij} is a matrix element, $\bar{x}_j$ is the column average, and s_j is the column standard deviation.

Sometimes standardization is termed as normalization, which must not be confused with normalization to unit length. Of course, other scaling options also exist, e.g., range scaling:

$$x'''_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{4}$$

and block scaling (scaling only a part of the matrix).

Covariance and correlation matrix can be formed from the original input matrix $\boldsymbol{X}$:

$$C = \text{cov}(x_j) = \frac{1}{n-1}(X'^T X') \tag{5}$$

$$R = \text{cor}(x_j) = \frac{1}{n-1}(X''^T X'') \tag{6}$$

Centering and standardization leads to some information loss. However, some of the statistical methods require (or at least works better; the interpretation is easier) the use of standardized data. Therefore, standardization is always recommended for beginners. The purpose and use of centering and scaling are discussed in depth in ref. [1].

In practice, it often happens that some items are missing from the matrix. The best way is to substitute missing data with column means (or partial mean of an interval if applicable). Alternatively, a random number within the range can be put instead of an empty place. If the empty places are numerous and they are not randomly located, then substitution is not recommended. Similarly, putting zeros in empty places is never the advocated practice. If a measurement value is below the detection limit, then half of the detection limit is a much better choice than zeros.

Generally, no data can be eliminated from the matrix without well-documented reasons. However, constant "variables" are not useful, and similarly, some of the highly correlated variables may be cancelled as they do not represent independent information. The qualitative statement "highly correlated" can hardly be quantified as it depends on the problem.

3. Multivariate methods

Let us group the methods in the simplest way. If the data can only be arranged in one matrix ($\boldsymbol{X}$), then unsupervised pattern recognition can be carried out. Such methods are PCA and cluster analysis (CA). It is relatively easy to assign a dummy variable to the rows (objects, cases) in the matrix. Supervised pattern recognition methods aimed to predict the dummy variable also called grouping variable ($\boldsymbol{Y}$). All prediction methods can be applied in a supervised way, i.e., to predict the grouping variable(s). What is the use of employing supervised pattern recognition when the aim is to group the data into classes and class memberships have to be known before the analysis? However, we can build models on known samples (training or learning data sets) and make predictions on unknown, not yet measured samples or compounds.

The most frequently used supervised pattern recognition method is the linear discriminant analysis (LDA), not to be confused with its twin brother canonical correlation analysis (CCA) or canonical variate analysis (CVA). Recently, classification and regression trees (CART) produced surprisingly good results. Artificial neural networks (ANNs) can be applied for both prediction and pattern recognition (supervised and unsupervised).

If two matrices can be defined, matrices of dependent variables ($\boldsymbol{Y}$) and independent variables ($\boldsymbol{X}$), then prediction methods are applicable. The simplest

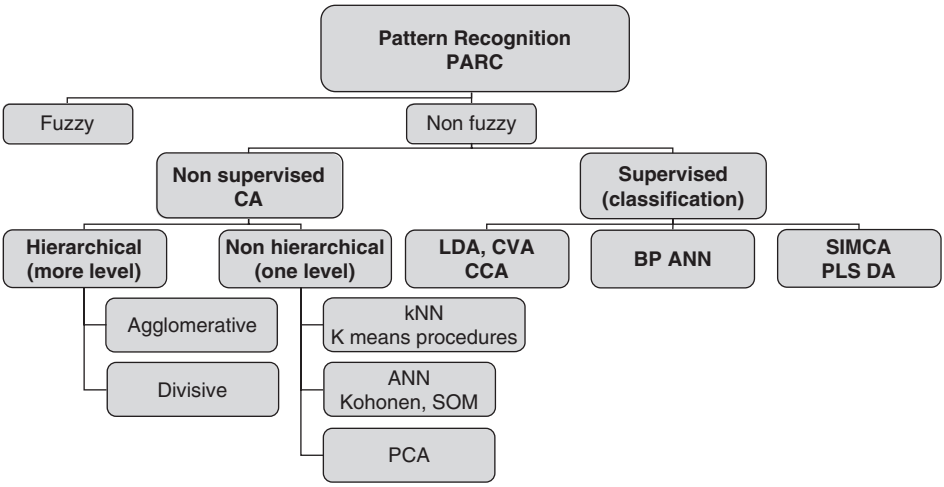


Fig. 1. Pattern recognition methods. ANN, artificial neural networks; BP ANN, back-propagation ANN; CA, cluster analysis; CART, classification and regression trees (recursive partitioning); CCA, canonical correlation analysis; CVA, canonical variate analysis; kNN, *k*-nearest neighbor methods; LDA, linear discriminant analysis; PCA, principal component analysis; PLS DA, partial least squares regression discriminant analysis; SIMCA, soft independent modeling of class analogy; SOM, self-organizing maps.

and best-understood prediction method is the multiple linear regression (MLR). It uses only one ***Y*** variable measured on a numerical scale (one at a time). Principal component regression (PCR) and partial least squares projection of latent structures (PLS) can have more ***Y*** vectors of numerical scale. The use of all pattern recognition and prediction methods is connected to variable (feature) selection. Many variables are not useful for prediction; they encode irrelevant information or even noise. Canceling uninformative variables ensures the successful application of chemometric techniques. The variable selection is implemented in the algorithm of prediction methods; here, only two general variable selection methods are mentioned: genetic algorithm (GA) and generalized pairwise correlation method (GPCM).

There are virtually endless number of methods for prediction and pattern recognition. The methods are frequently abbreviated. All of them have advantages and disadvantages; some of them have found use in special cases. In chemistry, "landscape" matrices often emerge, i.e., matrices having more columns than rows; LDA cannot handle such a situation but PLS can. Similarly, MLR cannot tolerate highly correlated variables but PLS can.

The methods of pattern recognition are summarized in Fig. 1.

3.1. Principal component analysis (PCA)

An $n \times m$ matrix can be considered n points in the m-dimensional space (or m points in the n-dimensional space). The points can be projected into a smaller dimensional subspace (smaller than n or m, whichever is the smaller) using proper techniques as PCA. Therefore, PCA is often called as a projection method. Projecting the points, dimension reduction of the data can be achieved. The principal components are often called underlying components; their values are the scores. The principal components are, in fact, linear combinations of the original variables. PCA is an unsupervised method of pattern recognition in the sense that no grouping of the data has to be known before the analysis. Still the data structure can be revealed easily and class membership is easy to assign.

The principal components are uncorrelated and account for the total variance of the original variables. The first principal component accounts for the maximum of the total variance, the second is uncorrelated with the first one and accounts for the maximum of the residual variance, and so on, until the total variance is accounted for. For practical reasons, it is sufficient to retain only those components that account for a large percentage of the total variance.

In summary, PCA decomposes the original matrix into multiplication of loading ($\boldsymbol{P}$) and score ($\boldsymbol{T}$) matrices:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} \tag{7}$$

PCA will show which variables and compounds are similar to each other, i.e., carry comparable information, and which one is unique. The schematic representation of PCA can be found in Fig. 2.

The algorithm of PCA can be found in standard chemometric articles and textbooks [2–4]. Fig. 3 shows an example of PCA. The separation in gas–liquid chromatography is ensured by stationary phases (liquids bound to chromatographic columns). These liquids have various separation abilities. Generally the polarity of

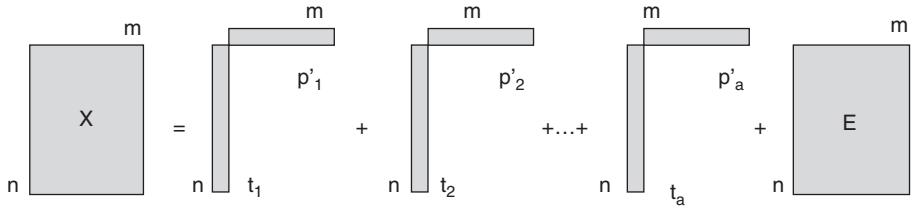


Fig. 2. Schematic representation of principal component analysis. ($\boldsymbol{X}$ original input matrix is decomposed into sum of several matrices (E is the residua, i.e., the error matrix); each matrix is calculated as outer product of two vectors. *t,* score; *p,* loading; *p′,* transpose of *p;* "*a*" is the number of principal components to be retained in the model.)

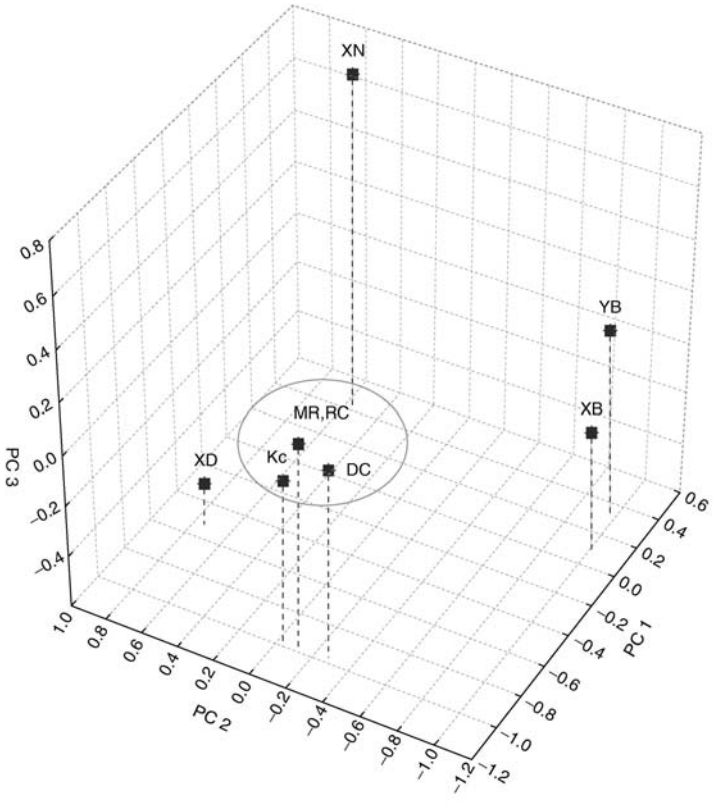


Fig. 3. Characterization of polarity (and selectivity) in gas chromatography. Principal component analysis of eight different polarity parameters: *DC*, *MR*, *Kc*, and *RP* are polarity parameters; *XB*, *YB*, *XD*, and *XN* are selectivity parameters. Notably, *MR* and *RP* carry exactly the same information ($R = 0.9999561$); *XB* and *YB* show close resemblance. The original eight-dimensional problem can be simplified into three dimensions without observable information loss.

stationary phases is used to characterize the separation. However, the polarity is not a unique, well-defined characteristic. Different authors define it differently. Eight polarity parameters are used to characterize the polarity of stationary phases [5]. The information they carry is redundant. Which polarity parameter is similar to others can be seen in the figure. Proximity of points means similarity: The closer a point, the more similar its polarity parameter is. The four polarity parameters form a dense cluster; the points for *MR* and *RP* are identical and *XN* is an outlier.

3.2. Cluster analysis (CA) [6, 7]

Two kinds of CA can be differentiated: hierarchical and nonhierarchical. Tree clustering producing dendrograms is a good example for hierarchical clustering,

whereas the *k*-nearest neighbor (kNN) method is for nonhierarchical ones. In fact, cluster analyses incorporate different algorithms. The common feature in the clustering algorithms is that they use distances for groupings (close objects form a cluster). CA helps to organize observed data into meaningful structures, that is, to

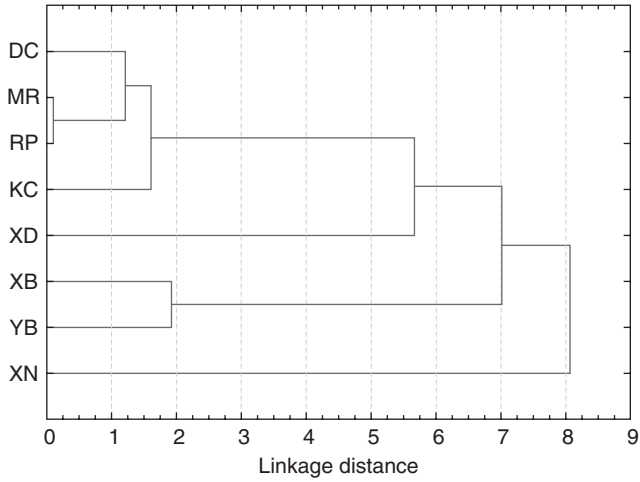


Fig. 4. Characterization of polarity (and selectivity) in gas chromatography. Cluster analysis of eight polarity parameters (cf. Fig. 3) using Euclidian distance and simple linkage. The polarity (*DC*, *MR*, *Kc*, *RP*) and selectivity (*XB*, *YB*, *XD* and *XN*) parameters are well distinguished. The close resemblance of *MR* and *RP* and to a lesser extent of *XB* and *YB* can also be seen.

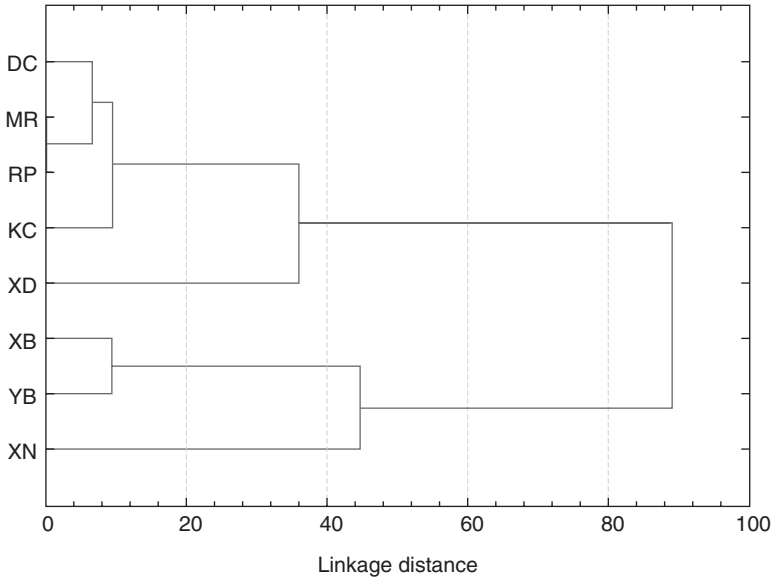


Fig. 5. Characterization of polarity (and selectivity) in gas chromatography. Cluster analysis of eight polarity parameters (cf. Fig. 3) using city block (Manhattan) distance and Ward's method.

develop taxonomies. Similarly, the correct diagnosis of a group of symptoms such as paranoia, schizophrenia, etc., is essential for successful therapy in the field of psychiatry. As said, an $\boldsymbol{X}$ matrix can be considered as $\boldsymbol{m}$ points in an $\boldsymbol{n}$-dimensional space (column-wise) or $\boldsymbol{n}$ points in an $\boldsymbol{m}$-dimensional one (row-wise).

Clustering algorithms differ from each other in how they define the distance measure and the distances among groups. Several measures for distance exist, e.g., Euclidian, Mahalanobis, city block (Manhattan), etc. Similarly, a number of linkage (amalgamation) rules have been defined: simple linkage, complete linkage, Ward's method, etc.

As the distance measure can be combined with various linkage rules, the results of CA are different; the unraveled pattern depends on the methods used. If all techniques provide the same pattern, the clustering, the classification can be accepted, otherwise it is not clear why the given linkage rule or distance measure provides acceptable/explainable groupings. Except using Mahalanobis distance, all clustering methods require standardization of data.

Figs. 4 and 5 embody CA results of exactly the same problem solved by PCA and shown in Fig. 3.

The results change using other distance measure and different linkage rule. Although the closeness of *MR* and *RP* and *XB* and *YB* remained, the polarity–selectivity distinction suffers: *XD* got into the cluster of polarity parameters; *XN* is not an outlier any more.

3.3. Multiple linear regression (MLR)

It is perhaps the most frequently applied chemometric method. One $\boldsymbol{Y}$ vector is related to the $\boldsymbol{X}$ matrix. The implicit assumption in MLR is the uncorrelatedness of variables (vectors of $\boldsymbol{X}$). It works well with long and lean (portrait) matrices, if the ratio of object exceeds at least five times the number of variables.

The basic regression equation to be solved is:

$$\boldsymbol{Y} = \boldsymbol{Xb} + \boldsymbol{\varepsilon} \tag{8}$$

where $\boldsymbol{b}$ is the vector of parameters to be fitted. Each $\boldsymbol{b}$ vector element corresponds to a variable (column) in $\boldsymbol{X}$. Variables that have $\boldsymbol{b}$ parameters not significantly different from zero should be eliminated from the model (variable selection). An estimation for $\boldsymbol{b}$ can be calculated by:

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{9}$$

where the superscripts "T" and "-1" mean transpose and inverse of the matrix, respectively.

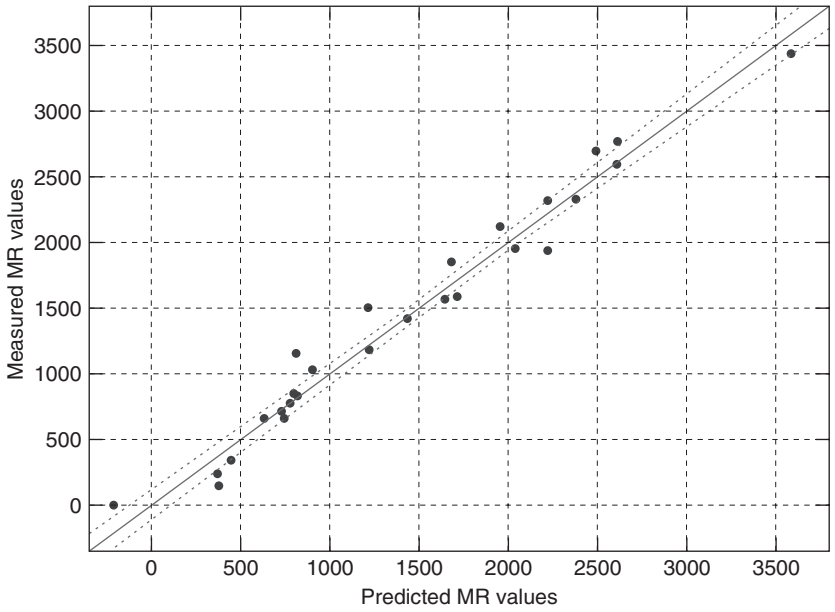


Fig. 6. Results of multiple linear regression for description of polarity in gas chromatography. Predicted and measured McReynolds polarity values.

Several model-building techniques were elaborated: forward selection and backward elimination, both in stepwise manner, all possible regressions, etc. [8].

Using the polarity example as mentioned earlier, it can be demonstrated how the various polarity variables are related. The most frequently used McReynolds polarity (*MR*) served as dependent variable. The following model can be built:

$$MR = 501.7 + 429.3\,DC - 3601\,XB$$
$$R = 0.9860;\ F(2,27) = 472.4;\ p = 0.0000;\ s = 157.6$$

where R is the multiple correlation coefficient, F is the overall Fisher statistic, p is the significance of the equation, and s is the standard error of the estimate. (The above example shows the standard way of providing regression results.)

Backward elimination procedure has kept ***X*** variables: *DC* and *XB* from among six polarity variables. The remaining variables (*YB*, *XN*, *XD*, and *Kc*) are not significant at the 5% level.

Fig. 6 shows a typical result of MLR. Dotted lines are the 95% confidence interval for the regression line.

3.4. Linear discriminant analysis (LDA) and canonical correlation analysis (CCA)

LDA allows us to classify samples with *a priori* hypothesis to find the variables with the highest discriminant power. This analysis is used to determine whether

the model (with all variables) leads to significant differences between the *a priori* defined groups, and which variables have significantly different means across the group. The selected variables are submitted to linear combinations to give rise to discriminant canonical functions, whose number is equal to the number of groups minus one (or equal the number of variables if it is smaller than the number of groups minus one). The first function provides the most overall discrimination between groups, the second provides the second most, and so on.

The discriminant power of the variables has been evaluated using Wilk's λ, F (Fisher statistics), and p-level parameters. The Wilk's λ is computed as the ratio of the determinant of the within-group variance/covariance matrix to the determinant of the total variance/covariance matrix: Its values ranges from 1 (no discriminatory power) to 0 (perfect discriminatory power).

LDA is perhaps the most frequently used supervised pattern recognition technique. It is supervised, that is, the class membership has to be known for the analysis. LDA, similarly to PCA, can be considered as a dimension reduction method. For feature reduction, we need to determine a smaller dimension hyperplane on which the points will be projected from the higher dimension space. While PCA selects a direction that retains maximal structure in a lower dimension among the data, LDA selects a direction that achieves maximum separation among the given classes. The latent variable obtained in this way is a linear combination of the original variables. This function is called the canonical variate; its values are the roots. In the method of LDA, a linear function of the variables is to be sought, which maximizes the ratio of between-class variance and minimizes the ratio of within-class variance. Finally, a percentage of correct classification is given. A variant of this method is the stepwise discriminant analysis that permits the variables with a major discriminant capacity to be selected. The description of LDA algorithm can be found in refs. [6,9,10].

Description of the discriminant analysis modules of Statistica™ program package [11]: A discrimination model will be built with the forward stepwise (forward selection) module of discriminant analysis step by step. Specifically, at each step the Statistica program will review all variables and evaluate which one will contribute to the most of the discrimination between groups. This variable will then be included into the model, and Statistica will proceed to the next step. In the general discriminant analysis module, a significance limit ($1 - \alpha$, say 95%) can be predefined. All variables that do not surpass the error limit (α, say 5%) will be included in the model, and all variables that surpass it will be eliminated.

Again, the earlier polarity example is utilized. The 30 stationary phases were classified into three categories: slightly, moderately, and highly polar according to variable "*DC*." LDA procedure in forward selection mode has selected the following variables: *DC*, *YB*, *XN*, *MR*, *Kc*, and *XB* at the 10% level. Only *XD* is not informative besides the other five. (*RP* was excluded from the analysis as it is highly correlated to *MR*.) Four stationary phases were classified into wrong groups

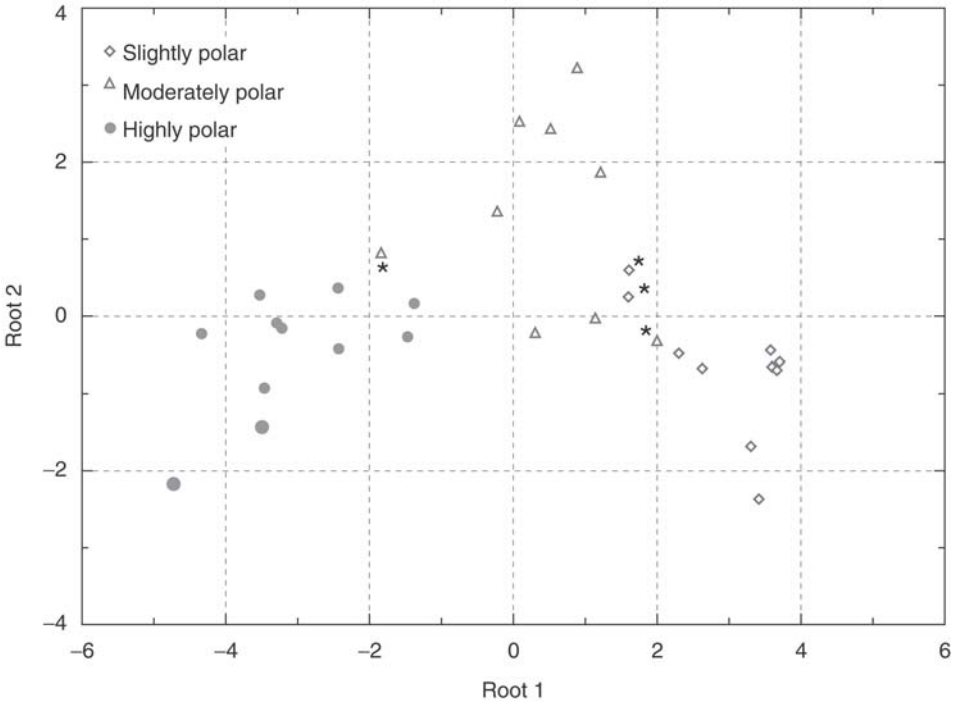


Fig. 7. Classification of stationary phases in gas chromatography. Canonical variates are plotted against each other. The misclassified stationary phases are marked with asterisks.

(misclassified). Considering that the classes were arbitrary and we used five discriminating variables instead of one, the results are satisfactory. The canonical variates show the discrimination of the classes. Stationary phase nos. 8, 9, 12, and 19 were misclassified (indicated with asterisks in Fig. 7).

3.5. Partial least squares projection of latent structures (PLS)

Partial least squares projection of latent structures (PLS) is a method for relating the variations in one or several response variables (***Y*** variables or dependent variables) to the variations of several predictors (***X*** variables), with explanatory or predictive purposes [12–14]. PLS performs particularly well when the various ***X*** variables express common information, i.e., when there is a large amount of correlation or even collinearity among them. PLS is a bilinear method where information in the original ***X*** data is projected onto a small number of underlying ("latent") variables to ensure that the first components are those that are most relevant for predicting the ***Y*** variables. Interpretation of the relationship between ***X*** data and ***Y*** data is then simplified, as this relationship is concentrated on the smallest possible number of components [15].

The partial least squares regression discriminant analysis (PLS DA) is a classification method based on modeling the differences between several classes with PLS [16–18]. If there are only two classes to separate, the PLS model uses one response variable, which codes for class membership as follows: 1 for the members of one class, 0 (or −1) for members of the other class (dummy variables) [18]. If there are three classes (or more), three dummy variables (or more) are needed. From the predicted ***Y*** values, we assigned the groups using the closest distance (maximum probability) approach. The maximum predicted values were assigned to unity (to the given class); smaller ones (all negative values) were assigned to zeros.

Again, the earlier polarity example is used for prediction of three dummy variables composed of zeros and unities showing the class memberships. The 30 stationary phases were classified into three categories: slightly, moderately, and highly polar according to *DC*. As PLS is not sensitive to collinearity of ***X*** variables, all variables were used including *RP*. Only *XD* is not informative besides the other five. Again, as in the case of LDA, four stationary phases were misclassified: stationary phase nos. 12, 13, 19, and 20, but they were different from the earlier misclassified ones (Fig. 8).

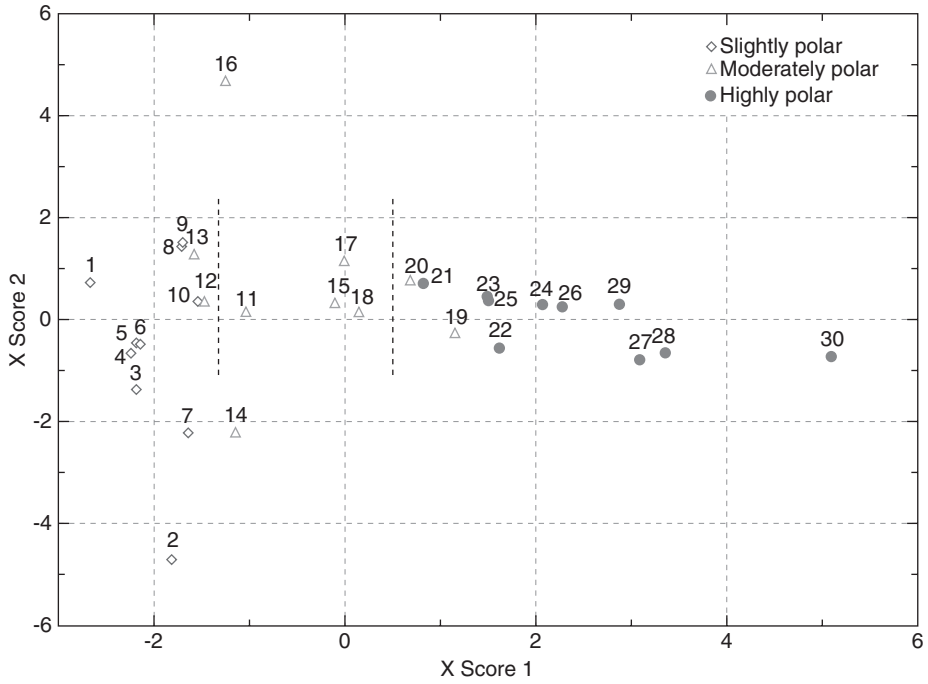


Fig. 8. Classification of stationary phases in gas chromatography. Partial least squares ***X*** scores are plotted against each other, whereas three PLS components were retained. Dotted lines show the separation of slightly, moderately, and highly polar phases. The misclassified stationary phases are nos. 12, 13, 19, and 20 (cf. Fig. 7).

The second ***X*** score expresses important features related to polarity: Stationary phase nos. 2 and 16 exert hydrogen donating and accepting ability, which might be important in many applications.

The misuse of chemometric methods is well summarized in ref. [19].

3.6. Classification and regression trees (CART)

Classification and regression tree (CART, eventually C&RT) is a tree-shaped structure that represents a set of decisions. These decisions generate rules for the classification of a data set. CART provides a set of rules that can be applied to a new (unclassified) data set to predict which records will have a given outcome [20,21]. It is easy to conjure up the image of a decision "tree" from such rules. A hierarchy of questions is asked and the final decision that is made depends on the answers to all the previous questions. Similarly, the relationship of a leaf to the tree on which it grows can be described by the hierarchy of splits of branches (starting from the trunk) leading to the last branch from which the leaf hangs. The recursive, hierarchical nature of classification trees is one of their most basic features; CART is also called recursive partitioning.

The final results of using tree methods for classification can be summarized in a series of (usually few) logical if–then conditions (tree nodes). Therefore,

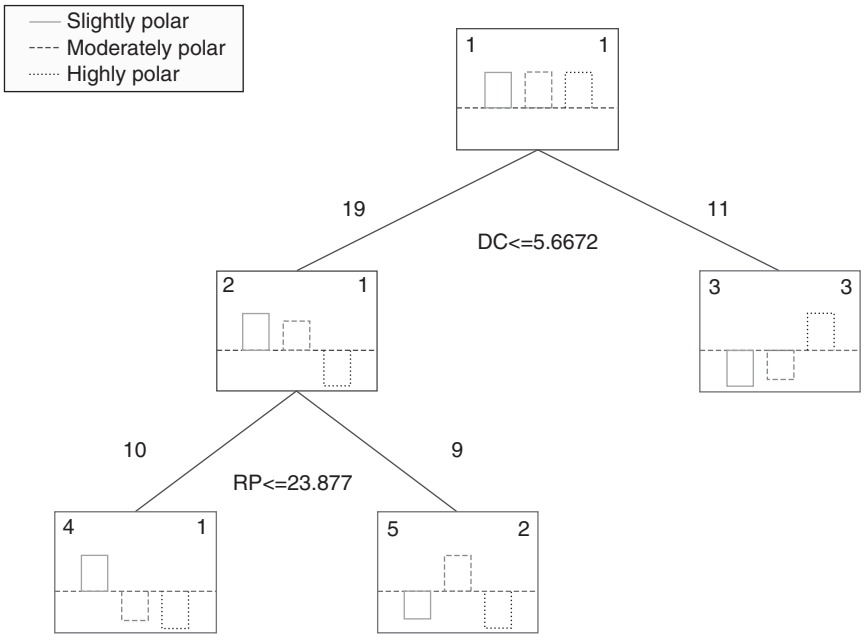


Fig. 9. Classification and regression trees using univariate split and prune on misclassification error. Using two variables (*DC* and *RP*) better classification can be achieved than using LDA or PLS. From among the three misclassified phases, one can only be seen at the first split.

there is no implicit assumption that the underlying relationships between the predictor variables and the dependent variable are linear or follow some specific nonlinear link function. The interpretation of results summarized in a tree is very simple. This simplicity is useful for purposes of rapid classification of new observations.

Unlike LDA, CART works well not only with numerical descriptors but also with categorical descriptors.

The Statistica program package implements three basic algorithms for CART: univariate split, linear combination split, and exhaustive search. Similarly, three stopping options can be chosen: prune on misclassification error, prune on deviance, and direct stop [11].

Fig. 9 shows the results of CART using univariate split and prune on misclassification error. Apart from the trivial solution (the polarity was classified using monotonous increase of *DC*), two variables (*DC* and *RP*) provide better classification than using five variables in LDA or eight variables in PLS. Three stationary phases were misclassified: nos. 11, 12, and 20. However, the CART solution does not reflect the complexity of the problem.

3.7. Artificial neural networks (ANN) [22–24]

A bunch of different methods of artificial intelligence are grouped under the term ANN. They can be used for pattern recognition both for supervised and unsupervised manner and for prediction purposes. ANNs are among the best available fitting methods; they can be applied for highly complex and strongly nonlinear relationships.

Artificial neural networks consist of groups of interconnected processing elements called neurons. The neurons are organized in layers producing "architecture." The first layer is termed the input layer, and each of its neurons receives information from outside (generally the independent variables are used as inputs). The last layer is the output layer; the layers of neurons between the input and output layers are called hidden layers. Input and output data (***X*** and ***Y*** matrices) are used to train the networks, e.g., change the weights for each connection; sum of all inputs for individual neuron transfers the information using appropriate transfer function (e.g., sigmoid, tangent hiperbolicus) and passes the results forward. Feed-forward neural networks connect the neurons in the upward direction, i.e., connections are not allowed among the neurons themselves (loops) and within one layer; they are consecutive (i.e., no jumps are allowed between layers). The weights are adjusted in such a way that the difference of measured and calculated outputs should decrease. The error propagates backwards during the training of feed-forward neural networks (80% of the ANN applications use back propagation learning). Kohonen's maps are self-organizing neural networks and have two layers only; they can unravel pattern in the data without using dependent variables (unsupervised pattern recognition).

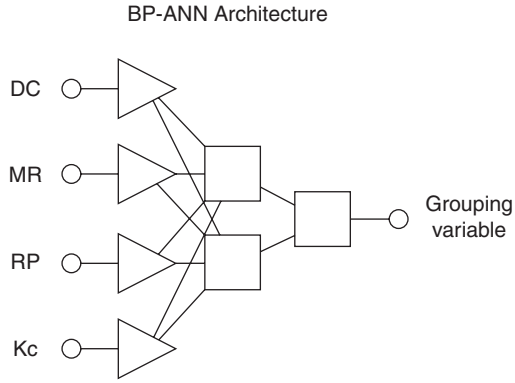


Fig. 10. Artificial neural network architecture for prediction of the polarity of columns in GC (optimized, but not fully optimal).

ANN is adaptive, i.e., it can learn from data and recognize dominant patterns. ANN is able to generalize; however, it has serious drawbacks as well. As a black box model, ANN can hardly be interpreted. It tends to overfit the data; it is not applicable for extrapolation but just for interpolation. There is no guarantee that a given architecture and training will find the global minimum. The selected variables depend on the initial random weights used for training. Careful cross-validation (CV) is needed to prove that the learned pattern is real and does not contain idiosyncrasy from noise.

Fig. 10 shows the architecture for the well-known polarity example. There is no need to use all variables to predict the grouping variable for polarity, except four variables (*DC*, *MR*, *Kc*, and *RP*). Selectivity parameters are not necessary for a proper classification (stationary phase no. 21 was misclassified with two hidden neurons; two phases (nos. 19 and 20) were misclassified with one hidden neuron). However, one variable *DC* classifies the phases as slightly, moderately, and highly polar.

3.8. Some methods of variable selection

3.8.1. Genetic algorithms

GAs are implementations of various search paradigms inspired by natural evolution. At a very general level, a GA may be any (chromosome-type) population-based model that uses selection and recombination operators to generate new sample points in a research space. Each input parameter (e.g., independent variables) is uniquely associated with a chromosome gene. The first step is to choose the size of chromosomes and to put in place an encoding scheme, uniquely mapping combinations of model parameters of the same size with chromosomes. The next step is to generate the initial populations of the chromosomes.

Implementing genetic competition requires definition of a fitness function for the chromosome population, e.g., R^2 (correlation coefficient) or R^2_{CV} (correlation coefficient for CV).

The GA starts with one or more current populations. The next population(s) is the result of genetic manipulation of the chromosomes through recombination (exchange of genes between chromosomes) and/or mutation (randomly replacing genes with genes not present in the chromosome; its role is to restore lost genetic material). The chromosomes are evaluated after each cycle using fitness function. Generation of new populations is represented until a satisfactory solution is identified [24–26].

Leardi and Gonzalez have addressed the issue of the use of GAs for extraction of the most relevant variables for PLS analysis. The critical point is summarized in their paper: ". . . a variable/objects ratio equal to 5 has been found to be the critical point, beyond which using GA will be very dangerous" [27].

The Moby digs software of Todeschini [28] applies the GA for variable selection in a MLR algorithm. The variable pool consists of 2000 variables at maximum two populations are allowed at a time. As a fitness function, R^2_{CV} can be selected.

3.8.2. Generalized pairwise correlation method

The pairwise correlation method (PCM) [29,30] utilizes a portion of information present in the data but overlooked till now. PCM selects from two independent variables ($\boldsymbol{X_1}$ and $\boldsymbol{X_2}$) which one is superior, i.e., "correlates" better to the dependent variable $\boldsymbol{Y}$. Three vectors are defined: $\boldsymbol{Y}$ (dependent variable), $\boldsymbol{X_1}$, and $\boldsymbol{X_2}$ (independent variables). The task is to choose the superior one from $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$. First, it is assumed that both of the independent variables correlate positively with the dependent variable $\boldsymbol{Y}$. Other cases are discussed in refs. [29,30] exhaustively. All the possible element pairs of the $\boldsymbol{Y}$ vector are considered that can occur when the differences $\Delta\boldsymbol{X_1}$ for $\boldsymbol{Y}$ vs. $\boldsymbol{X_1}$, and $\Delta\boldsymbol{X_2}$ for $\boldsymbol{Y}$ vs. $\boldsymbol{X_2}$ are determined. Only the signs of the differences are taken into account. There will be $m = \binom{n}{2} = n(n-1)/2$ point pairs and differences $\Delta\boldsymbol{X_1}$ as well as $\Delta\boldsymbol{X_2}$. The frequencies for the four possible different signs of $\Delta\boldsymbol{X_1}$ and $\Delta\boldsymbol{X_2}$ are arranged in a 2×2 contingency table. If both differences are positive (and both are negative), the distinction cannot be made between $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$. However, if the frequency value for opposite signs of differences for $\boldsymbol{X_1}$ is significantly greater, then $\boldsymbol{X_1}$ is termed as superior, otherwise $\boldsymbol{X_2}$. Whether the frequency value is significant or not, this can be determined using suitable statistical tests: the Williams' *t*-test as a parametric test and the McNemar's, the Chi-square and the conditional Fisher's tests as nonparametric statistical tests [29].

In its generalized form (GPCM), all possible independent variable pairs are compared and the number of "superiority" is determined. The number of "superiority"

Table 1
Results of generalized pairwise correlations for the polarity example: Dependent variables were (a) *MR* and (b) *XB*

(a)	*RP*	*Kc*	*DC*	*XD*	*XN*	*YB*	*XB*
No. of wins	6	4	4	2	1	1	0
No. of losses	0	1	1	3	3	4	6
No. of decisions	0	1	1	1	2	1	0
Rank ordering wins–losses	**1**	**2**	**3**	4	5	6	7
	α (user)	=0.05	α (emp.)	0	Crit. sum	11.4	12
(b)	*YB*	*XD*	*XN*	*Kc*	*MR*	*RP*	*DC*
No. of wins	6	4	4	3	1	1	0
No. of losses	0	1	1	3	4	4	6
No. of decisions	0	1	1	0	1	1	0
Rank ordering wins–losses	**1**	**2**	**3**	4	5	6	7
	α (user)	=0.05	α (emp.)	0	Crit. sum	11.4	12

Conditional exact Fisher test as selection criterion and "ranking according to the number of wins minus losses" were used in both cases. Bold numbers indicate the variables selected.

is termed as the number of wins: How many times a given ***X*** variable was "superior" to the other ***X*** variables. The number of "inferiority" is termed as the number of losses: How many times an ***X*** variable was "inferior" to the other ***X*** variables. The number of wins is simply summed for all variable pair comparisons. Several ranking methods, namely (i) simple ranking according to the number of wins, (ii) ranking according to the differences in wins and losses, and (iii) probability weighted ranking according to the differences in wins and losses, were elaborated for GPCM [31,32].

GPCM needs an independent variable; then, it rank orders all the remaining variables. If *MR* was used as supervisor, the next most similar variable to it is *RP*, second next is *Kc*, and so on. (Table 1a). If *XB* were selected as supervisor, the next most similar variable to it is *YB*, second next is *XD*, and so on (Table 1b). In both cases, the polarity and selectivity parameters are well distinguished.

3.8.3. Other aspects of data analysis

In the exploratory phase of any data analysis it is expedient to calculate means, standard deviation medians, skewness, and kurtosis of variables. They are important indicators of the distribution of data. Box and Whisker plot using median reveals easily the asymmetry in the distribution. If the number of data makes it possible, it is worth to plot the histogram of each variable and test the normality (Kolmogorov-Smirnov, Shapiro-Wilk's test, etc.). If the number of variables is

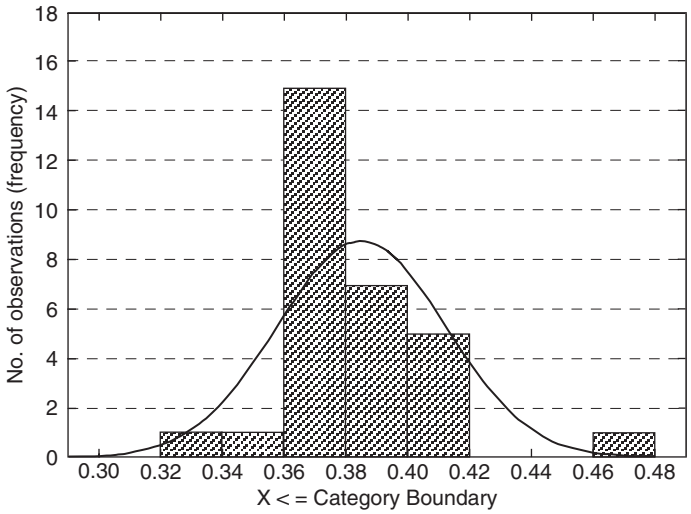


Fig. 11. Histogram for *XN* (Kolmogorov-Smirnov $d = 0.18221$, $p > 0.20$; Lilliefors $p < 0.05$; Shapiro-Wilk $W = 0.89464$, $p = 0.00623$).

small (<8–10), matrix plot can show strong patterns and outliers in the data immediately. Similarly, calculating the correlation matrix is certainly useful. If the data are not normally distributed, nonparametric alternatives of correlation coefficients [32] should be used, such as Spearman ρ and Kendall τ. Fig. 11 shows that the distribution of *XN* is far from being normal. Although the Kolmogorov-Smirnov test is conservative, the two other tests indicate the nonnormal distribution equivocally, at the 5% level.

Correlation coefficient is applied most frequently to reveal relationships and connections between variables. However, its use is seldom correct. Misuse and abuse of correlation coefficient is well spread in all scientific fields. First of all, its value without the degrees of freedom says nothing. Even $r = 0.997$ is not significant at the 5% level if $n = 3$; in contrast, $r = 0.300$ is significant if $n > 44$ [33]. Large correlation coefficient does not mean causal relationships between the variables necessarily. Two increasing series of numbers are always correlated. Even zero correlation coefficient does not mean that there are no relationships between the variables, but then it means that no linear relationship exists. Clustering the data in two groups can provide high correlation coefficient with the illusion of definite relationship.

3.8.4. Validation of model building techniques

The most frequently used technique for model validation is no doubt the CV. It can be applied in several variants: leave-one-out (LOO), leave-multiple-out

(leave-*n*-out) splitting the data set into training an test sets [34]. Unfortunately, there is no agreed method how to split data set into training, calibration, and test sets.

There is a widespread perception among statisticians that CV is a poor method of verifying the fit of a model. Especially, LOO method is damned. On the contrary, Miller recommends splitting the data into three sets: One is used for model selection, the second one for parameter estimation (calibration), and the third one for external validation (CV is a poor alternative instead) [35]. Tropsha and Gramatica also support the view in insisting to the external validation [36].

The prediction error is estimated using CV almost unbiasedly in case no feature selection has been made. However, CV is heavily biased when the variables are selected from a large number of variables. The indicators of the fit are deceptively overoptimistic in such cases [37].

A correct LOO cross-validation can be done by moving the delete-and-predict step inside the subset search loop. In other words, we take the sample of size n, remove one case, search for the best subset regression on the remaining $n - 1$ cases, and apply this subset regression to predict the holdout case. Repeat for each of the cases in turn, getting a true holdout prediction for each of the cases. Use these holdouts as a measure of the fit [38,39].

A fast and effective way is to estimate the performance of a fit by using generated matrices consisting of random numbers. The same number of variables should be generated as was used in the modeling, prediction step for real measured data. The same procedure should be followed as in the real case, variable selection, model building, etc., and the indicators of the fit (correlation coefficients for the training and prediction sets, prediction errors) should be compared with the same values of the real case. If the random numbers indicate approximately the same fit and/or prediction, the variables selected, the models built on real data are of little value even if physical significance can be found for the parameters of the model.

4. Selected applications of chemometrics

There are numerous applications evaluating results of instrumental–analytical methods (like mass spectrometry, NMR spectroscopy, chromatography, etc.) using chemoinformatics. Combinations involving NMR spectroscopy and chromatography have many applications in the biomedical field. On the other hand, although chemometric techniques have frequently been applied to analyze mass spectral data, applications in the biomedical field are rare. Multivariate data analysis has been applied to mass spectrometry [40], especially revealing relationships of mass spectral data and chemical structure [41]. The state of the art for structure elucidation can be found in critical reviews and evaluations [42, 43]. Even the differentiation of stereoisomers can be solved using mass spectral data coupled with chemometric data evaluation [44].

Mass spectrometry and chemometric methods cover very diverse fields: Different origin of enzymes can be disclosed with LC–MS and multivariate analysis [45]. Pyrolysis mass spectrometry and chemometrics have been applied for quality control of paints [46] and food analysis [47]. Olive oils can be classified by analyzing volatile organic hydrocarbons (of benzene type) with headspace–mass spectrometry and CA as well as PCA [48]. Differentiation and classification of wines can similarly be solved with headspace–mass spectrometry using unsupervised and supervised principal component analyses (SIMCA = soft independent modeling of class analogy) [49]. Early prediction of wheat quality is possible using mass spectrometry and multivariate data analysis [50].

Pyrolysis mass spectrometry and chemometrics have been coupled to analyze the adulteration of orange juice quantitatively [51], to test the authenticity of honey [52], and to discriminate the unfractionated plant extracts [53].

GC–MS coupled with chemometric techniques has been used to characterize roasted coffees [54], to detect adulterants in olive oils [55], and to determine fatty acids in fish oils [56]. GC–MS data have also been used in toxicology assessments to reveal patterns in complex chemical mixtures with the help of multivariate analyses [57,58].

Novel fast developing fields are metabonomics, metabolomics, proteomics, and genomics (the "omics world"). The connection to chemometric methods can easily be established without going into details and discussing their definition. However, the utilization of mass spectral data is relatively rare in these fields: More than 300 compounds can be distinguished with GC–MS after deconvolution of overlapping peaks [59]. Screening biomarkers in rat urine have been solved using LC–MS data (electrospray ionization) and 2-way data analysis [60]. The useful methods of chemoinformatics have been summarized in the review [17].

Time of flight mass spectrometry has also provided data for chemometric analyses, e.g., for PCA [61,62] and for trilinear (3-way) analysis [63].

The chemometric approach has been applied on diverse field of mass spectral data evaluation: peak resolution and quantification [64], calibration [65], instrument standardization [66], fast interpretation [67], and evaluation of rate constants [68].

Finally some sources are mentioned, which are not necessarily bound to mass spectrometry, but they illustrate well the usefulness of chemometric methods in medical diagnosis: multilevel component analysis of metabolomic fingerprinting data [69], artificial neural network applications for the clinical diagnosis of tumors [70,71], lipoprotein analysis [72], searching cardiovascular markers [73], differentiation of heroin samples [74], and quantification of pollution levels by multiway modeling [75]. Patients with low, normal, and high plasma cholesterol and statin therapy level were classified using LDA: Variables (markers) with highest discrimination power were selected [76]. The inherent accuracy of ^{1}H-NMR spectroscopy to quantify plasma lipoproteins is subclass dependent [77].

Terms and terminology

"All possible regressions" is a model-building method, when all possible variable combinations are examined in the model.

Backward elimination is a variable selection algorithm for multiple linear regression; it starts with all variables in the model and eliminates all nonsignificant variables; see forward selection as well.

Calibration data set is a part of the data on which the estimation of model parameters is carried out.

Canonical variate is a linear combination of the original variables for the highest discrimination power.

City block (Manhattan) distance equals the sum of absolute distances for each variable.

Class membership information shows groups or clusters in the data.

Complete linkage defines the distance between clusters as the distance between the two farthest objects.

Confidence intervals for the regression line are limits within the estimated regression line that can be found with a certain, say 95%, probability.

Cross-validation is the collective term for a bunch of validation techniques.

Dendrogram or branched diagram is a diagram showing the relationships of items arranged like the branches of a tree.

Dimension reduction is generally achieved by combining the original variables in a linear way (defining principal components) and not using all linear combinations.

Euclidian distance is computed by finding the square of the distance between each variable, summing the squares, and finding the square root of the sum.

Fisher statistic, Fisher value: ratio of variances for two models to be compared. It can be overall or partial F value. The overall Fisher statistic tests the entire equation, whether all coefficients are significant in the model. The partial F value is used to test whether the incriminated variable is significant in the model.

Forward selection is a variable selection algorithm for multiple linear regression; it starts with no variable in the model and introduces all significant variables; see backward elimination as well.

Grouping variable ($\boldsymbol{Y}$) shows whether a given object (statistical case) belongs to a certain class (e.g., the code 0 means "ill" and the code "1" means healthy); it is also called dummy variable.

Hierarchical clustering uses algorithms, which find successive clusters using previously established clusters. Hierarchical algorithms can be agglomerative (bottom–up) or divisive (top–down). Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters.

"Landscape" matrix is a short and fat matrix, i.e., matrices having (substantially) more columns than rows.

Latent variable is a linear combination of the original variables.

Linkage (amalgamation) rules differ from each other in how they define the distances between clusters.

*Loading (**P**) matrix* consists of linear coefficients for principal component analysis (Equation 7).

Mahalanobis distance is based on correlations between variables by which different patterns can be identified and analyzed. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e., not dependent on the scale of measurements.

Nonhierarchical or partitional clustering uses algorithms, which determine all clusters at once.

Partial least squares projection of latent structures (PLS) is a method for relating the variations in one or several response variables (***Y*** variables or dependent variables) to the variations of several predictors (***X*** variables), with explanatory or predictive purposes.

"Portrait" matrix is a long and lean matrix, i.e., matrices having (substantially) more rows than columns.

Prediction set is an independent part of the data that serves to check the model performance (also called test set).

p-level parameters: significance level parameters.

Principal components are linear combinations of the original variables.

Projection methods project the points into a smaller dimensional subspace.

Roots: values of the canonical variate; cf. scores and principal components.

*Score (**T**) matrix* consists of linear combinations of the original variables (Equation 7).

Significance limit is a predefined probability: $1 - \alpha$, where α is the error limit.

Significance of an equation (p) is the limit (threshold) probability, where the equation is still significant.

Simple linkage defines the distance between clusters as the distance of the two closest objects.

Stepwise linear regression is a variant of multiple linear regression in which variables are added one at a time according to the F test.

Supervised pattern recognition methods are the methods that use the class membership information while revealing dominant pattern in the data.

Taxonomy refers to either a classification of things or the principles underlying the classification.

Training data set is a part of the data on which model building is carried out (also called learning set).

Underlying components or principal components are linear combinations of the original variables; it is also called latent variables.

Unsupervised pattern recognition methods are methods that do not use the class membership information while searching dominant pattern in the data.

Validation data set. *See* calibration data set.

Ward's method takes into account the number of objects when defining distance between clusters.

Wilk's λ is the standard statistic that is used to denote the statistical significance of the discriminatory power of the current model.

References

1. Bro, R. and Smilde, A.K., Centering and scaling in component analysis. *J. Chemom.*, **17**, 16–33 (2003).
2. Wold, S., Esbensen, K. and Geladi, P., Principal component analysis. *Chemom. Intell. Lab. Syst*., **2**, 37–52 (1987).
3. Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., Jong, S.D.E., Lewi, P.J. and Smeyers-Verbeke, J., Principal component analysis. In: *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier, Amsterdam, The Netherlands, 1998, pp. 88–149 (Chapter 31.1).
4. Otto, M., *Chemometrics Statistics and Computer Application in Analytical Chemistry*. Wiley-VCH, Weinheim, Germany, 1999. Chapter 5.2. Principal component analysis, pp. 125–132.
5. Héberger, K., Evaluation of polarity indicators and stationary phases by principal component analysis in gas-liquid chromatography. *Chemom. Intell. Lab. Syst*., **47**, 41–49 (1999).
6. Otto, M., *Chemometrics Statistics and Computer Application in Analytical Chemistry*. Wiley-VCH, Weinheim, Germany, 1999. Chapter 5.2.2. Cluster analysis, pp. 148–158.
7. Hartigan, J.A., *Clustering Algorithms*. Wiley, New York, 1975.
8. Draper, N.R. and Smith, H., *Applied Regression Analysis* (2nd edition). Wiley, New York, 1981.
9. Hastie, T., Tibshirani, R. and Friedman, J., Linear methods for classification. In: *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, New York, 2001, pp. 79–113 (Chapter 4).
10. Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., Jong, S.D.E., Lewi, P.J. and Smeyers-Verbeke, J., Supervised pattern recognition. In: *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier, Amsterdam, The Netherlands, 1998, pp. 207–223 (Chapter 33).
11. Statistica 6 software package, StatSoft Inc., Tulsa, OK, USA.
12. Wold, S., PLS for multivariate linear modeling. In: van de Waterbeemd, H. (Ed.), *Chemometric Methods for Molecular Design.* VCH, Weinheim, 1995, pp. 195–218 (Chapter 4.4).
13. Geladi, P. and Kowalski, B.R., Partial least squares regression: a tutorial. *Anal. Chim. Acta*, **185**, 1–17 (1986).
14. Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., Jong, S.D.E., Lewi, P.J. and Smeyers-Verbeke, J., Partial least squares regression. In: *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier, Amsterdam, The Netherlands, 1998, pp. 329–342 (Chapter 35.7).
15. Esbensen, K.H., *Multivariate Data Analysis—in Practice* (5th edition). CAMO Process AS, Oslo, 2002.
16. Barker, M. and Rayens, W., Partial least squares for discrimination. *J. Chemom.*, **17**, 166–173 (2003).
17. Eriksson, L., Anti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., Long, I., Lundstedt, T., Trygg, J. and Wold, S., Using chemometrics for navigating in the large data sets of genomics, proteomics and metabonomics (gpm). *Anal. Bioanal. Chem*., **380**, 419–429 (2004).
18. Ståhle, L. and Wold, S., Partial least squares analysis with cross validation for the two-class problem: a Monte Carlo study. *J. Chemom.*, **1**, 185–196 (1987).

19. Defernez, M. and Kemsley, E.K., The use and misuse of chemometrics for treating classification problems. *Trends Anal. Chem.*, **16**, 216–221 (1997).
20. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1984.
21. Mazzatorta, P., Benfenati, E., Lorenzini, P. and Vighi, M., QSAR in ecotoxicity: an overview of modern classification techniques. *J. Chem. Inf. Comput. Sci*., **44**, 105–112 (2004).
22. Zupan, J. and Gasteiger, J., *Neural Networks for Chemists*. VCH Verlagsgesellschaft, Weinheim, Federal Republic of Germany, 1993.
23. Masters, T., *Practical Neural Network Recipes in C++*. Academic Press, Inc., Academic Press Inc., San Diego, CA and Academic Press Inc., London, UK, 1993.
24. Niculescu, S.P., Artificial neural networks and genetic algorithms in QSAR. *J. Mol. Struct. (Teochem)*, **622**, 71–83 (2003).
25. Lucasius, C.B. and Kateman, G., Understanding and using genetic algorithms: Part 1. Concepts, properties and context. *Chemom. Intell. Lab. Syst*., **19**, 1–33 (1993).
26. Lucasius, C.B. and Kateman, G., Understanding and using genetic algorithms: Part 2. Representation, configuration and hybridization. *Chemom. Intell. Lab. Syst*., **25**, 99–145 (1994).
27. Leardi, R. and González, A.L., Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst*., **41**, 195–207 (1998).
28. Todeschini, R., Ballabio, D., Consonni, V., Mauri, A. and Pavan, M., MobyDigs Professional version 1.0 2004 Milano Chemometrics and QSAR Research Group.
29. Rajkó, R. and Héberger, K., Conditional Fisher's exact test as a selection criterion for pair-correlation method. Type I and type II errors. *Chemom. Intell. Lab. Syst*., **57**, 1–14 (2001).
30. Héberger, K. and Rajkó, R., Generalization of pair-correlation method (PCM) for nonparametric variable selection. *J. Chemom.*, **16**, 436–443 (2002).
31. Héberger, K. and Rajkó, R., Variable selection using pair-correlation method. Environmental applications. *SAR QSAR Environ. Res.*, **13**, 541–554 (2002).
32. Conover, W.J., Measures of rank correlation. In: *Practical Nonparametric Statistics* (2nd edition). John Wiley & Sons, New York, 1980, pp. 250–256 and pp. 256–263 (Chapter 5.4).
33. Bevington, P.R., *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill Book Co., New York, 1969.
34. Baumann, K., Cross-validation as the objective function for variable-selection techniques. *Trends Anal. Chem.*, **22**, 395–406 (2003).
35. Miller, A.J., *Subset Selection in Regression*. Chapman and Hall, London, 1990, pp. 43–82.
36. Tropsha, A., Gramatica, P. and Gombar, V.K., The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci*., **22**, 69–77 (2003).
37. Baumann, K., Chance correlation in variable subset regression: influence of the objective function, the selection mechanism, and ensemble averaging. *QSAR Comb. Sci*., **24**, 1033–1046 (2005).
38. Hawkins, D.M., Assessing model fit by cross validation. *J. Chem. Inf. Comput. Sci*., **43**, 579–586 (2003).
39. Hawkins, D.M., The problem of overfitting. *J. Chem. Inf. Comput. Sci*., **44**, 1–12 (2004).
40. Varmuza, K., Penchev, P., Stancl, F. and Werther, W., Systematic structure elucidation of organic compounds by mass spectra calssification. *J. Mol. Struct*., **408/409**, 91–96 (1997).
41. Varmuza, K., Recognition of relationships between mass spectral data and chemical structures by multivariate data analysis. *Anal. Sci.*, **17(Suppl.)**, 1467–1470 (2001).
42. Varmuza, K., Chemical structure information from mass spectrometry, *Encyclopedia of Spectroscopy and Spectrometry*. Academic Press, San Diego, 2000.

43. Demuth, W., Karlovits, M. and Varmuza, K., Spectral similarity versus structural similarity: mass spectrometry. *Anal. Chim. Acta*, **516**, 75–85 (2004).
44. Le Bizec, B., Antignac, J-P., Bertrand, D., Qannari, E.M. and Andre, F., Multidimensional statistical analysis applied to electron ionization mass spectra to determine steroid stereochemistry. *Rapid Commun. Mass Spectrom*., **19**, 509–518 (2005).
45. Bylund, D., Samskog, J., Markides, K.E. and Jacobsson, S.P., Classification of lactate dehydrogenase of different origin by liquid chromatography-mass spectrometry and multivariate analysis. *J. Am. Soc. Mass Spectrom*., **14**, 236–240 (2003).
46. Wilcken, H. and Schulten, H.-R., Quality control of paints; Pyrolysis—mass spectrometry and chemometrics. *Anal. Chim. Acta*, **336**, 201–208 (1996).
47. Guillou, C., Lipp, M., Radovic, B., Reniero, F., Schmidt, M. and Anklam, E., Use of pyrolysis-mass spectrometry in food analysis: application in the food analysis laboratory of the European Commissions' Joint Research Centre. *J. Anal. Appl. Pyrolysis*, **49**, 329–335 (1999).
48. Peña, F., Cárdenas, S., Gallego, M. and Valcárcel, M., Direct screening of olive oil samples for residual benzene hydrocarbon compounds by headspace-mass spectrometry. *Anal. Chim. Acta*, **526**, 77–82 (2004).
49. Marti, M.P., Busto, O. and Guasch, J., Application of a headspace mass spectrometry system to the differentiation and classification of wines according to their origin, variety and ageing. *J. Chromatogr. A*, **1057**, 211–217 (2004).
50. Ghirardo, A., Sorensen, H.A., Petersen, M., Jacobsen, S. and Sondergaard, I., Early prediction of wheat quality: analysis during grain development using mass spectrometry and multivariate data analysis. *Rapid Commun. Mass Spectrom*., **19**, 525–532 (2005).
51. Goodacre, R., Hammond, D. and Kell, D.B., Quantitative analysis of the adulteration of orange juice with sucrose using pyrolysis mass spectrometry and chemometrics. *J. Anal. Appl. Pyrolysis*, **40–41**, 135–158 (1997).
52. Radovis, B.S., Goodacre, R. and Anklam, E., Contribution of pyrolysis-mass spectrometry (Py-MS) to authenticity testing of honey. *J. Anal. Appl. Pyrolysis*, **60**, 79–87 (2001).
53. Goodacre, R., York, E.V., Heald, J.K. and Scott, I.M., Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemsitry*, **62**, 859–863 (2003).
54. Zambonin, C.G., Balest, L., De Benedetto, G.E. and Palmisano, F., Solid-phase microextraction-gas chromatography mass spectrometry and multivariate analysis for the characterization of roasted coffees. *Talanta*, **66**, 261–265 (2005).
55. Lorenzo, I.M., Pavón, J.L.P., Laespada, M.E.F., Pinto, C.G. and Cordero, B.M., Detection of adulterants in olive oil by headspace-mass spectrometry. *J. Chromatogr. A*, **945**, 221–230 (2002).
56. Jalali-Heravi, M. and Vosough, M., Characterization and determination of fatty acids in fish oil using gas chromatography-mass spectrometry coupled with chemometric resolution techniques. *J. Chromatogr. A*, **1024**, 165–176 (2004).
57. Eide, I., Neverdal, G., Thorvaldsen, B., Grung, B. and Kvalheim, O.M., Toxicological evaluation of complex mixtures by pattern recognition: correlating chemical fingerprints to mutagenecity. *Environ. Health Perspect.*, **110(6 Suppl.)**, 985–988 (2002).
58. Eide, I., Neverdal, G., Thorvaldsen, B., Arneberg, R., Grung, B. and Kvalheim, O.M., Toxicological evaluation of complex mixtures: fingerprinting and multivariate analysis. *Environ. Toxicol. Pharmacol*., **18**, 127–133 (2004).
59. Jonsson, P., Gullberg, J., Nordström, A., Kusano, M., Kowalczyk, M., Sjöström, M., and Moritz, T., A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Anal. Chem*., **76**, 1738–1745 (2004).
60. Idborg-Björkman, H., Edlund, P-O, Kvalheim, O.M., Schuppe-Koistinen, I. and Jacobsson, S.P., Screening of biomarkers in rat urine using LC/electrospray ionization-MS and two-way data analysis. *Anal. Chem*., **75**, 4784–4792 (2003).

61. Sinha, A.E., Hope, J.L., Prazen, B.J., Fraga, C.G., Nilsson, E.J. and Synovec, R.E., Multivariate selectivity as a metric for evaluating comprehensive two-dimensional gas chromatography—time-of-flight mass spectrometry subjected to chemometric peak deconvolution. *J. Chromatogr A*, **1056**, 145–154 (2004).
62. Sinha, A.E., Fraga, C.G., Prazen, B.J. and Synovec, R.E., Trilinear chemometric analysis of two-dimensional comprehensive gas chromatography—time-of-flight mass spectrometry data. *J. Chromatogr. A*, **1027**, 269–277 (2004).
63. Pacholski, M.L., Principal component analysis of TOF-SIMS spectra, images and depth profiles: an industrial perspective. *Appl. Surf. Sci.*, **231/232**, 235–239 (2004).
64. Fraga, C.G., Chemometric approach for the resolution and quantification of unresolved peaks in gas chromatography—selected-ion mass spectrometry data. *J. Chromatogr. A*, **1019**, 31–42 (2003).
65. Araujo, P., Nguyen, T-T. and Froyland, L., Chemometric approaches in calibration experiments of trilinolenoylglycerol by liquid chromatography ion-trap mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **16**, 388–396 (2005).
66. Goodacre, R., Timmins, É.M., Jones, A., Kell, D.B., Maddock, J., Heginbothom, M.L. and Magee, J.T., On mass spectrometer instrument standardization and interlaboratory calibration transfer using neural networks. *Anal. Chim. Acta*, **348**, 511–532 (1997).
67. Windig, W., Smith, W.F. and Nichols, W.F., Fast interpretation of complex LC/MS data using chemometrics. *Anal. Chim. Acta*, **446**, 467–476 (2001).
68. Carpignano, R., Operti, L., Rabezzana, R. and Vaglio, G.A., Chemometric approach to evaluate the parameters affecting the determination of reaction rate constants by ion trap mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **9**, 938–944 (1998).
69. Jansen, J.J.; Hoefsloot, H.C.J., Greef, J.V.D., Timmermann, M.E. and Smilde, A.K., Multilevel component analysis of time-resolved metabolic fingerprinting data. *Anal. Chim. Acta*, **530**, 173–183 (2005).
70. Shan, Y., Zhao, R., Xu, G., Lieblich, H.M. and Zhang, Y., Application of probabilistic neural network in the clinical diagnosis of cancers based on clinical chemistry data. *Anal. Chim. Acta*, **471**, 77–86 (2002).
71. Zhao, R., Xu, G., Yue, B., Lieblich, H.M., Zhang, Y., Artificial neural network classification based on capillary electrophoresis of urinary nucleosides for the clinical diagnosis of tumors. *J. Chromatogr. A*, **828**, 489–496 (1998).
72. Dyrby, M., Petersen, M., Whittaker, A.K., Lambert, L., Norgaard, L., Bro, R. and Engelsen, S.B., Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics. *Anal. Chim. Acta*, **531**, 209–216 (2005).
73. Balla, B., Mocak, J., Pivovarnikova, H. and Balla, J., Comparative study of cardiovascular markers data by various techniques of multivariate analysis. *Chemom. Intell. Lab. Syst.*, **72**, 259–267 (2004).
74. Klemenc, S., In common batch searching of illicit heroin samples—evaluation of data by chemometrics methods. *Forensic Sci. Int.*, **115**, 43–52 (2001).
75. Flåten, G.R., Grung, B. and Kvalheim, O.M., Quantification of pollution levels by multiway modelling. *J. Chemom.*, **18**, 173–182 (2004).
76. Nagy, K., Jakab, A., Pollreisz, F., Bongiorno, D., Ceraulo, L., Averna, M.R., Noto, D. and Vékey, K., Analysis of sterols by high-performance liquid chromatography/mass spectrometry combined with chemometrics, *Rapid Commun. Mass Spectrom.*, **20**, 2433–2440 (2006).
77. Ala-Korpela, M., Lankinen, N., Salminen, A., Suna, T., Soininen, P., Laatikainen, R., Ingman, P., Jauhiainen, M., Taskinen, M.-R., Héberger, K. and Kaski, K., The inherent accuracy of ^{1}H NMR spectroscopy to quantify plasma lipoproteins is subclass dependent. *Atherosclerosis*, **190**, 352–358 (2007) (available online via *Science Direct*, 30 May 2006).

This page intentionally left blank

Part III
Biomolecules

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 8

Mass spectrometry in proteomics

AKOS VERTES*

W. M. Keck Institute for Proteomics Technology and Applications, Department of Chemistry, The George Washington University, Washington, DC, USA

1. Introduction

In recent decades, rapidly expanding knowledge in molecular biology has provided the biochemical framework for the functioning of all eukaryotic organisms. The core principles of this framework are based on three fundamental classes of molecules: nucleic acids, proteins, and metabolites. In a living organism a gene, coded in the DNA, is transcribed into an RNA molecule. Through processing, the noncoding regions of the RNA are removed and a messenger RNA, mRNA, is spliced. Genes in the DNA are studied by genomics, whereas their expression in the form of mRNA is explored by transcriptomics. The past 20 years witnessed the

*Tel.: +1-202-994-2717; Fax: +1-202-994-5873. E-mail: vertes@gwu.edu.

sequencing of the human genome [1,2]; thus, discovering the genetic basis of certain diseases became feasible.

According to current estimates, there are approximately 25,000 genes in the human genome. Owing to alternative splicing and other mechanisms, the transcriptome of an organism is much more complex than the genome. As transcription and processing are influenced by the condition of the organism, disease states can be reflected in expression level changes in the transcriptome. Analysis of the transcribed mRNAs is typically carried out using DNA microarrays.

The second group of molecules, proteins, is produced through the ribosome-mediated translation of the mRNAs. Proteins serve as the general actors in carrying out most cell functions from motility to mitosis. The nature and activity of these functions are regulated by multitudes of posttranslational modifications, e.g., by acetylation, phosphorylation, or ubiquitination, of the proteins. These modifications emerge as the main regulators of protein functions. Proteomics, a vigorously developing field, is the systemic study of all proteins produced by an organism.

Owing to posttranslational modifications, there are many more proteins than mRNAs. It is estimated that approximately one million different proteins correspond to the ~25,000 human genes. In addition, protein concentrations vary greatly in space, time, and expression level. Therefore, it is not sufficient to ascertain that a particular protein is present in the organism; the spatial and temporal distributions of its concentration also have to be established. Spatial variations of protein expression in an organism are traditionally imaged using quantitative autoradiography and fluorescent labeling methods, including tagging with green fluorescent protein. These approaches, however, require the development of labels for every individual protein. Therefore, their utility for high-throughput systemic studies is very limited.

Importantly, the proteome can change in response to a disease. The altered expression levels can be used in diagnostics or form the basis of treatment strategies. Conventional methods of expression profiling were largely based on two-dimensional gel electrophoresis (2-DE). However, because of the limited accuracy, resolution, and specificity of this method, positive protein identification had to rely on additional forms of analysis. As a result of these complicating factors, proteomics presents an even greater challenge than genomics.

Some of the common objectives in proteomics include identification of proteins in a particular tissue or biological fluid (through peptide mapping, sequence tags, *de novo* sequencing, etc.), secondary, tertiary, or quaternary structure analysis of known proteins, function analysis through epitope mapping, quantitation of protein expression levels, and imaging of their distributions. The main method used for protein identification and quantitation in proteomics is mass spectrometry. An introduction to the established methods of mass spectrometry in proteomics is the subject of this chapter.

Mass spectrometry is uniquely positioned among the large variety of analytical techniques to achieve the outlined objectives [3]. Chapter 6 presents a thorough introduction to the principles and instrumentation of mass spectrometry. Mass spectrometric methods provide a better sensitivity, dynamics range, and selectivity than nuclear magnetic resonance (NMR) techniques. Mass spectra are more specific and less complex than many forms of optical spectroscopy and, given the right ionization technique, they can provide structural information. With the discovery of electrospray ionization [4] (ESI) and matrix-assisted laser desorption/ionization [5,6] (MALDI) in the late 1980s, the ion sources with the necessary capabilities (no high mass limit and adjustable amount of fragmentation) became available and the stage was set for the birth of proteomics. For their respective role in developing these enabling technologies, John Fenn and Koichi Tanaka received the 2002 Nobel Prize in Chemistry [7].

The third class of molecules, metabolites, is a diverse collection of typically smaller species ($<$1500 Da) that participate in cellular energy production and in the synthesis and degradation of macromolecules. The systematic study of the human metabolome has started only recently. By early 2007 already over 2000 endogenous metabolites have been identified, quantitated, and catalogued [8]. There is clearly a large diversity for this class of molecules; for example, the number of different metabolites in the plant kingdom is estimated to be $\sim$200,000. In addition to the endogenous metabolites, molecules introduced from the environment through nutrition or as drugs and their degradation products are also present in living organisms. A simplified view of the three major molecular classes, their hierarchy and interactions, and the corresponding disciplines devoted to their study are presented in Fig. 1.

Most biomedical samples contain thousands of biochemical components and thus are too complex even for mass spectrometry. Separation methods are needed to reduce this complexity by selecting smaller groups of components from the original specimens (see Chapter 5). The most commonly used separation methods in proteomics are affinity chromatography with its high selectivity, multidimensional techniques, such as 2-DE, and the combination of ion exchange (IEX) and high-performance liquid chromatography (HPLC). More recently, ion mobility spectrometry was used to separate the polypeptide components before mass analysis.

These separation methods and especially their combinations with mass spectrometry are capable of producing data in large volumes. Curated archiving and interpretation of these data require sophisticated computational resources. Bioinformatics aims to manage and mine the rapidly growing information from genomic, proteomic, and metabolomic investigations including the discovery of reaction networks (see Chapter 10). There are numerous bioinformatics databases and tools available on the Internet (e.g., http://www.ncbi.nlm.nih.gov/, http://www.expasy.ch/, and http://prospector.ucsf.edu/) and from commercial sources. The leading mass spectrometer manufacturers integrate their data acquisition systems with these tools to provide comprehensive solutions for proteomics research.

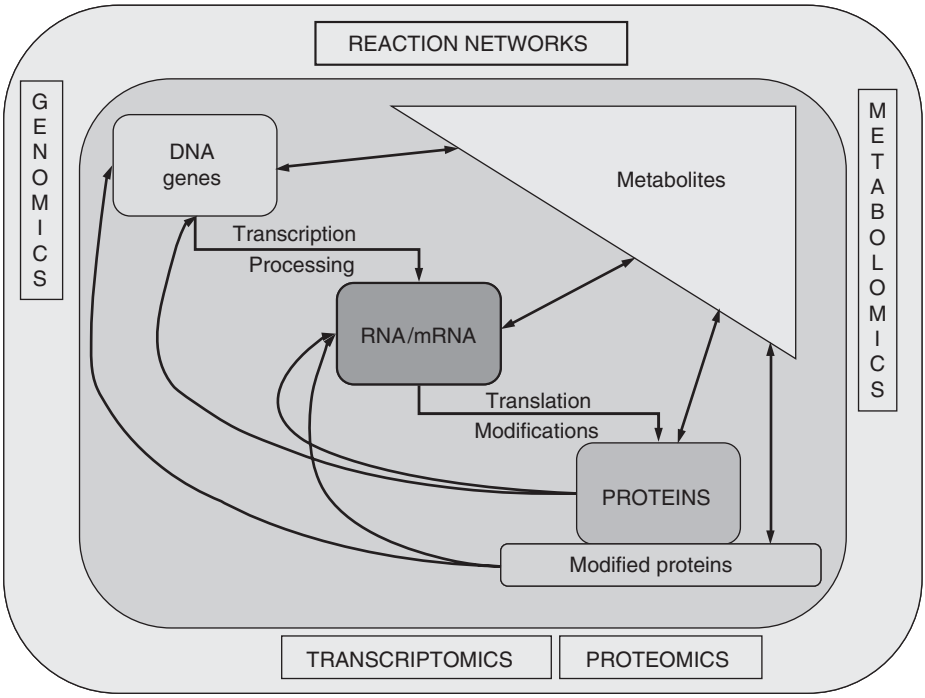


Fig. 1. Fundamental molecular classes in eukaryotic organisms and their interactions. Subdisciplines devoted to studying particular classes are shown on the perimeter. Genomics gives an unprecedented glimpse into the DNA-based molecular design of life. Proteomics studies the translated and modified proteins, the main actors of cellular processes, and metabolomics tracks the dynamic changes in the makeup of small molecules brought about by inherent and environmental conditions. Ultimately, all the molecular constituents, their interactions, and the knowledge of the entire reaction network are needed to understand the basic processes in physiology.

2. Methods in proteomics

A common task in proteomic analysis is to identify a subset of proteins in a biomedical sample. In principle, this can be accomplished through two different routes. The first, and most common, approach is to break down the proteins into peptide segments of manageable size through enzymatic digestion and analyze these building blocks using mass spectrometry. This is the so-called bottom-up approach. The other, less common, method that relies on the analysis of intact proteins is the top-down approach. The top-down strategy requires high-performance mass spectrometers (e.g., ion cyclotron resonance, ICR, or orbitrap systems; see Chapter 6) with exceptional mass resolution and accuracy in combination with powerful fragmentation techniques (such as electron capture dissociation, ECD) to

enable sequence readout. In the following sections we briefly review the methods used in the bottom-up and top-down approaches.

2.1. Peptide mapping

Peptide mapping takes advantage of the accurate mass measurement of unique protein fragments produced by highly specific enzymatic digestion. Typically, trypsin is used due to its high fidelity in producing peptides in the size range most efficient for protein identification ($400 < m/z < 5000$). This range corresponds to ~4–45 amino acid residues; thus, the corresponding peptides exhibit sufficient specificity. It also coincides with the m/z range where some common mass analyzers (e.g., quadrupoles or ion traps) show their best performance.

Accurate mass measurement of the resulting peptides produces a set of m/z values that can be compared against a database of protein fragment masses [9,10]. These fragment databases are produced by the *in silico* digestion of all the entries in large protein databases. Several fragment databases are available online with the necessary searching tools. For example, as of January 9, 2007, the SwissProt protein database contained 252,616 entries. Their *in silico* digestion using trypsin with a single missed cleavage allowed the production of 10,225,094 peptides [11]. The search algorithm finds the proteins with enzymatic fragments in this database that match the measured peptide masses within a predefined tolerance. Usually there are multiple possible matches and a review is required to further narrow the set and ultimately identify the unknown protein.

The efficiency of identification greatly depends on the performance of the mass spectrometer. Most notably, the mass accuracy of the instrument, usually determined by studying standards, has a dramatic effect. Clearly, the more accurate the measured masses are the narrower is the set of proteins that produce fragments with masses within the tolerance. The number of peptides identified is correlated with the amino acid residue coverage of the original protein.

We demonstrate the mechanics of peptide mapping using the example of the α-chain of human hemoglobin. This protein is composed of 141 residues: VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLS HCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR. Unfragmented, it appears in the MALDI mass spectrum as a protonated ion with a molecular weight of 15,126.5 Da. This single number is clearly not specific enough to identify the protein. There are many other proteins with the same m/z, e.g., the ones with any permutation of the residues. Tryptic digestion with no missed cleavages produces characteristic fragments in the $400 < m/z < 5000$ range. Table 1 shows these fragments, their location in the original protein molecule, and the corresponding calculated monoisotopic and average masses.

Table 1
Monoisotopic (mi) and average (av) peptide masses from tryptic digestion of human hemoglobin α chain [11]

m/z (mi)	*m/z* (av)	Start	Sequence	End
461.2718	461.5416	8	TNVK	11
532.2878	532.6235	12	AAWGK	16
729.4141	729.8564	1	VLSPADK	7
818.4407	818.9537	93	VDPVNFK	99
1071.5543	1072.3195	32	MFLSFPTTK	40
1252.7147	1253.4903	128	FLASVSTVLTSK	139
1529.7343	1530.6470	17	VGAHAGEYGAEALER	31
1833.8919	1835.0415	41	TYFPHFDLSHGSAQVK	56
2996.4894	2998.3651	62	VADALTNAVAHVDDMPNALSALSDLHAHK	90
3038.6496	3040.6206	100	LLSHCLLVTLAAHLPAEFTPAVHASLDK	127

In our first example we use a low-performance mass spectrometer. Assuming that five peptides (*m/z* 729.86, 818.95, 1072.32, 1253.49, and 1530.65) appear in the mass spectrum (e.g., as commonly observed, due to the ion-suppression effect we do not detect all tryptic peptides), the average masses are determined with 2000 ppm mass accuracy, and the search in the SwissProt database is restricted to the proteins of *Homo sapiens,* the MS-Fit searching tool of Protein Prospector [11] finds 114 entries that are more or less consistent with this data. The relevant section of the mass spectrum is shown in Fig. 2. Note that in this example no impurities complicate the spectrum.

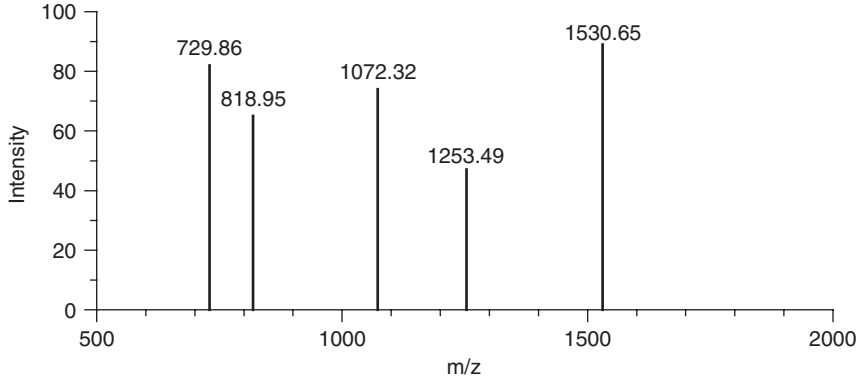


Fig. 2. Five fragment masses determined from the mass spectrum of the mock unknown protein (human hemoglobin α chain) tryptic digest form the basis of peptide mapping. The mass-to-charge ratio is labeled *m/z* on the horizontal axis.

The top-ranked hit is human hemoglobin α subunit with all five masses matched, but with only 35.5% coverage (in light gray below).

```
1 VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNAL
81SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

The other proteins on the list showed fewer number of matching peptides or lower degree of coverage.

There are several ways to increase the fidelity of protein identification. Chief among them are to use better performing instrumentation [12] (nowadays a typical high-performance mass spectrometer can achieve ~5–10 ppm mass accuracy) and to identify more peptides. Improving the mass accuracy to 50 ppm for the same set of *m*/*z* values does not increase the coverage, but it reduces the number of hits from 114 to a single one, human hemoglobin α subunit.

Increasing the number of peptides used in the search to 10 (*m*/*z* 461.54, 532.62, 729.86, 818.95, 1072.32, 1253.49, 1530.65, 1835.04, 2998.37, and 3040.62) without improving mass accuracy (keeping it at 2000 ppm) actually increases the number of hits in the search to 1151, but the coverage of the top scoring human hemoglobin α subunit increases to 93.6%.

```
1 VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNAL
81SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

Enhanced mass accuracy (50 ppm) for this set of peptides reduces the number of hits to one, the human hemoglobin α subunit, with 93.6% coverage. Thus, the right sample preparation and ionization method in combination with species information and reasonable instrument performance enabled us to identify a single protein in a database of over 250,000 entries.

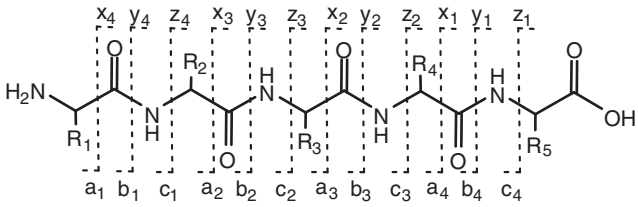
2.2. Peptide fragmentation

Peptide mapping does not require any knowledge about the primary structure of the protein or of its fragments. Owing to peptide fragmentation, however, parts of the primary structure might become known from the mass spectra. The spontaneous fragmentation of peptides is relatively slow; it mostly takes place in the postsource region of the mass spectrometer. In time-of-flight instruments equipped with an ion reflector, the ions produced by postsource decay (PSD) become observable in the mass spectrum at appropriate reflector voltage settings. This gives rise to peptide-sequencing capabilities [13].

More energetic ionization methods (in-source decay, ISD) or collisions with inert (collision-activated dissociation, CAD, also known as collision-induced

dissociation, CID) or reacting species (ECD and electron transfer dissociation, ETD) also produce structural data. In these techniques, fragmentation of the peptide backbone is induced through increasing the internal energy of the ions (ISD and CAD) or through ion chemistry (ECD and ETD). Thus, the presence of particular fragments in the spectrum is the function of the different ionization methods, e.g., MALDI and ESI, and more recently desorption/ionization on silicon [14] (DIOS) and laser-induced silicon microcolumn arrays [15] (LISMA) as well as instrument types (TOF, ion trap, ICR, etc.). There is more control over fragmentation patterns in tandem mass spectrometers (e.g., MS/MS and MS^n), where the primary ion internal energy can be adjusted by, for example, CAD.

Depending on the actual bond that breaks in the peptide backbone (C–C, C–N, or N–C) and on the partitioning of the charge on the resulting fragments (amino or carboxyl side fragment), there are six major fragment types. Their nomenclature for a pentapeptide is shown below.



Other less common fragmentation pathways, e.g., resulting in internal fragments or neutral loss ions, are not discussed here. As an example we can look at the neuropeptide leucine enkephalin, which has a sequence of YGGFL and a protonated monoisotopic mass of *m/z* 556.28. The fragmentation of this ion in a collision cell through CAD might produce a tandem mass spectrum similar to the one in Fig. 3.

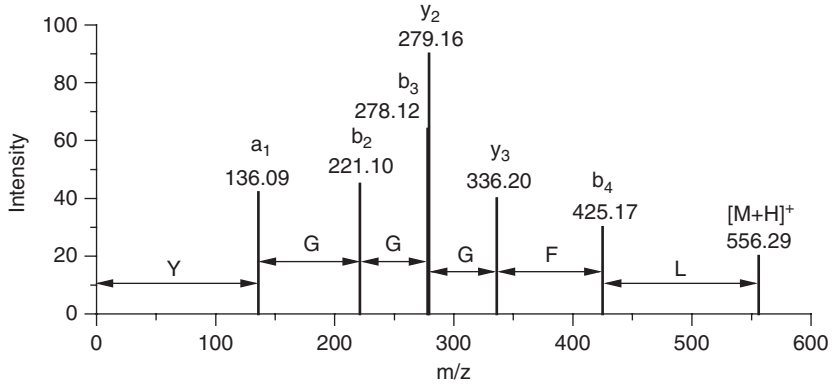


Fig. 3. Fragmentation of the protonated leucine enkephalin molecular ion via CAD in a tandem mass spectrometer.

Table 2
Monoisotopic masses of major fragment ions for leucine enkephalin

N-terminal ions					C-terminal ions				
N	1	2	3	4	N	4	3	2	1
a_n	136.08	193.10	250.12	397.19	x_n	419.19	362.17	305.15	158.08
b_n	–	221.09	278.11	425.18	y_n	393.21	336.19	279.17	132.10
c_n	–	238.12	295.14	442.21	z_n	377.19	320.17	263.15	116.08

In this simplified case the identity of amino acid residues in the peptide can be inferred from the mass differences of successive peaks by comparing them with the known masses of the amino acids. In real-world samples, the presence of other ions and the absence of certain fragments make this task fairly complex. Comparison of the measured m/z values in the spectrum with the calculated fragment masses in Table 2 enables the assignment of the peaks.

In addition to a_1 and the molecular ion, parts of the b_n and y_n series are present in Fig. 3. The sequence can be read as YGGFL. Note that the y series reads the sequence from right to left, whereas the b series reports it from left to right. Coincidentally, the mass difference between b_2 and b_3 and between y_2 and y_3 identify the same residue.

Changing the internal energy of the ions through CAD can reveal more about the primary structure. This can be induced by changing the collision energy of the primary ions or by changing the collision gas pressure in the tandem mass spectrometer [16]. With the emergence of new laser desorption/ionization platforms based on nanostructured silicon, simpler instrumentation can also yield similar data. Fig. 4 shows the spectrum of a vasodilator peptide, bradykinin (RPPGFSPFR), as a function of relative laser intensity. At low laser power the molecular ion dominates the spectrum. This can be advantageous in complex mixtures, where the molecular weights of the different components can be identified. Increasing the laser power from 95 to 145 relative value resulted in enhanced structure-specific fragmentation. Although the entire primary structure cannot be inferred from this spectrum, the identity of the N-terminal residues is revealed.

Even in the case of unmodified residues, entire peptide sequences are rarely revealed by fragment spectra induced by CAD. The task is even more complex when posttranslational modifications are present. Phosphorylation, for example, is prevalent due to its role in signal transduction and in the regulation of protein function. In eukaryotic cells, as much as 30% of the proteins can be phosphorylated. Histone protein functions are believed to be regulated by acetylation, phosphorylation, methylation, and ubiquitination. These modifications play an important part in fundamental biological functions, e.g., gene silencing. Identifying

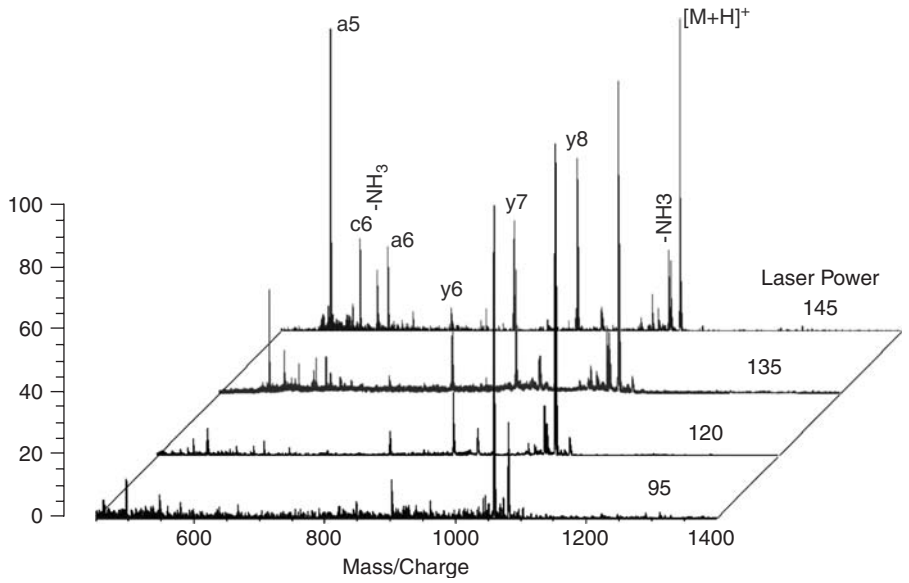


Fig. 4. Laser desorption/ionization of 1 pmol of bradykinin from a LISMA surface produces increasing amount of structure-specific fragmentation as the relative laser power increases.

the modified residues by mass spectrometry requires comprehensive fragmentation of the protein domains of interest [17].

2.3. Sequence tags

Although comprehensive sequence information on protein domains is not available on most instruments, shorter segments are often revealed by PSD, CAD, or other techniques. The concept of a sequence tag is based on using the partial sequence of a peptide digestion product, usually composed of a few residues, in combination with the masses of the adjoining N- and C-terminal fragments to efficiently search protein databases for the identity of unknown proteins [18,19].

For example, let us assume that we find three *b* series fragment ions, *m*/*z* 908.4, 1021.5, and 1108.5 in the CAD spectrum from the tryptic digest of the human hemoglobin α subunit that belong to the peptide parent ion with *m*/*z* 1833.9 (see Fig. 5). This is the peptide between residues 41 and 56 in Table 1.

The mass differences in the *b* series reveal the presence of L/I followed by S in the sequence. This information is sufficient to attempt a sequence tag search. Searching the SwissProt database for *H. sapiens* proteins by entering *m*/*z* 1833.9 for the parent ion and 1108.5, 1021.5, and 908.4 for the *b* series fragments in the MS-Seq searching tool of Protein Prospector [11] turns up a single protein, human hemoglobin α subunit with primary accession number P69905.

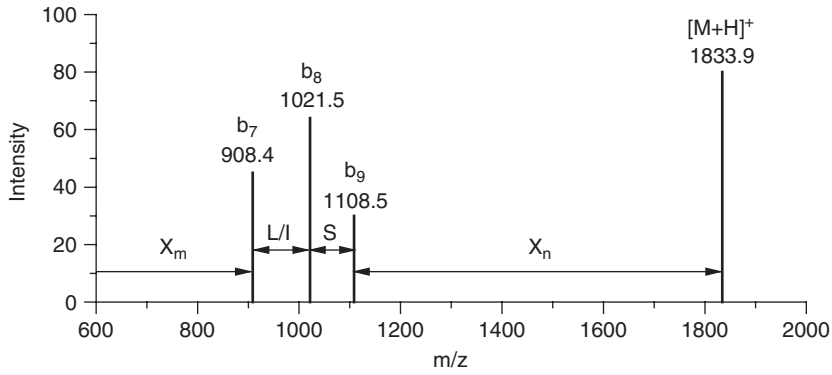


Fig. 5. MS/MS mass spectrum reveals the partial sequence of a tryptic peptide from the human hemoglobin α subunit. This information is sufficient to successfully perform a sequence tag search and identify the protein. L/I stands for leucine or isoleucine, whereas X_m and X_n denote unknown sequences.

The initial 252,616 entries in the database are reduced to 1328 by the parent mass filter. Using the three fragment masses the number of matching proteins for all species is 80. At this point, all the hits are related to the hemoglobin α subunit. Introducing the information on the species produces a single hit. Note, however, that the sequence coverage of the protein is only 11.3%. This limitation curtails the value of sequence tag identifications in the presence of multiple posttranslational modifications.

2.4. De novo *sequencing*

We have seen powerful methods to identify proteins in a sample based on mass spectra and information from large protein databases. These strategies require that the protein of interest exists in the database. Protein databases contain information that was originally produced by traditional Edman sequencing or by meticulous mass spectrometric methods commonly known as *de novo* sequencing. These approaches are necessary if the protein of interest is undescribed or substantially modified. Although both Edman degradation and tandem mass spectrometry can provide sequences with acceptable accuracy, recently mass spectrometry seems to have come out on top due to its dramatically higher throughput and better sensitivity.

There are two major approaches to *de novo* sequencing by mass spectrometry. The first one is based on a number of empirical rules obtained by observing typical peptide fragmentation schemes [20]. Current versions of this approach rely on computerized expert systems that are built on the dozens of empirical rules and factors. These include general observations on the prevalence of certain fragments in spectra produced by the used fragmentation methods and in typical instruments. For example, CAD is known to produce predominantly *y*- and *b*-type ions. There

are other rules related to neutral losses and relative intensities of spectral features, and on determining the presence of certain amino acid residues based on immonium ions formed by the combination of *a*- and *y*-type cleavages.

Furthermore, it is imperative to recognize the ambiguities resulting from identical or indistinguishable masses (isobars). Common examples are leucine and isoleucine or lysine and glutamine with only 0.0364 Da mass difference for the latter. Similar problems arise when dipeptide masses are isobaric with single amino acids or with other dipeptides. These challenges can only be resolved by using instrumentation of sufficiently high mass accuracy or by residue-specific chemical derivatization. The expert systems can successfully call sequences of over 10 residues, including posttranslational modifications.

The other approach to *de novo* sequencing is based on a systematic treatment of tandem mass spectrometric data and database search. An excellent description of these methods is available in Chapter 9; thus, we refrain from the detailed discussion here.

As the exploration of the human proteome advances from better known proteins to more and more obscure ones, the significance of *de novo* sequencing as the primary source of information is likely to grow. Similarly, the identification of splice variants, mutations, and modifications calls for increasing number of *de novo* investigations.

2.5. Electron capture and electron transfer dissociations

As we pointed out in Section 2.2, the *y* and *b* series ions induced by CAD, or other methods of gradually producing elevated internal energy, rarely reveal even the majority of the residues. For example, only ~25% of the 76-residue ubiquitin sequence can be identified through CAD. This incomplete information leaves the primary structure unresolved.

The problem with gradually energizing these polypeptide ions seems to be the rapid redistribution of internal energy, which leads to the preferential breakage of a low number of the weakest bonds. After several years of searching for a method to produce more complete fragmentation, great improvement was achieved by reacting low-energy electrons and the multiply charged peptide ions, $[M + nH]^{n+}$, produced by ESI [21]. This method, termed electron capture dissociation (ECD), produced a radical cation, $[M + nH]^{(n-1)+\bullet}$, that in turn rapidly dissociated into *c* and *z* series ions with the degree of fragmentation approaching 80% and without preference to bond strength [22]. An alternative fragmentation pathway can also produce *a*- and *y*-type ions. Not only the fragments in ECD provide higher coverage than CAD but also the information in the two methods is complementary. Thus, a mass spectrometric method to sequence large peptides and small proteins in their entirety became feasible. This also presented a realistic approach to top-down

proteomics, i.e., to the analysis of intact protein components without enzymatic cleavage.

A comparison of the fragments produced by CAD and ECD shows the advantages of the latter in phosphopeptide analysis. Quadruply charged molecular ions of a 28-mer phosphopeptide, atrial natriuretic peptide substrate (ANPS), SLRRSpSCFGGRIDRIGAQSGLGCNSFRY, were fragmented by the two methods [23]. The resulting patterns showed incomplete fragmentation (20 of the 27 peptide bonds) for CAD with significant loss of the phosphorylation site information. The corresponding ECD spectrum showed complete sequence coverage and the location of the phosphorylation site (see Fig. 6).

ETD takes the concept of ECD to the next level [24]. Owing to the conditions required to trap the thermalized electrons that produce ECD, it can only be performed in ICR mass spectrometers. These systems are large and expensive; thus, this technical requirement limits the availability of ECD to a relatively small number of laboratories. To make the benefits of ECD available on more common instrumentation (e.g., ion traps), heavier electron-donating agents, i.e., low electron affinity anions are needed that can be trapped together with the peptide ions. Anthracene [24] and fluoranthene [17] radical anions as ETD agents were shown to generate primarily *c*- and *z*-type ions from multiply charged large peptide, phosphopeptide, and small protein species. Like ECD, ETD produces close to complete fragmentation and thus enables the elucidation of primary structures.

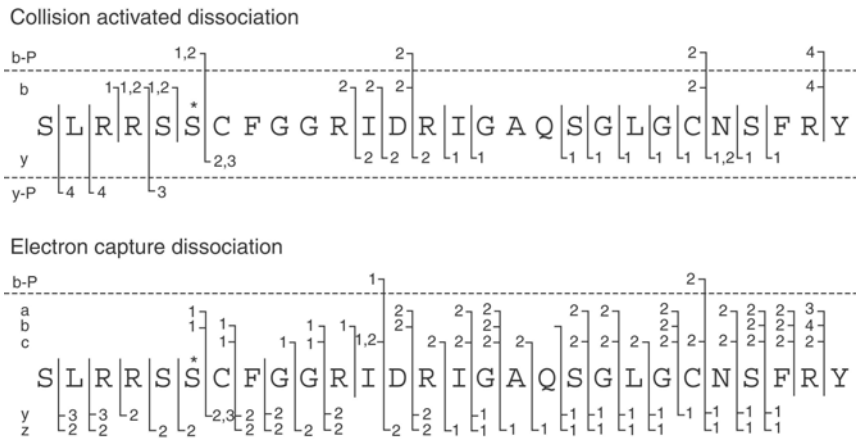


Fig. 6. Comparison of fragmentation patterns for a 28-mer phosphopeptide. In the top pattern produced by CAD, incomplete backbone fragmentation and extensive phosphate loss (denoted by −P) can be observed. Numbers indicate the charge carried by a particular fragment. Complete sequence readout and identification of the phosphorylation site are straightforward for ECD (bottom pattern). (Reprinted with permission from: Shi, S.D.H., Hemling, M.E., Carr, S.A., Horn, D.M., Lindh, I. and McLafferty, F.W., *Anal. Chem.*, **73**, 19–22 (2001). Copyright 2001. American Chemical Society.)

A fascinating application of ETD is the analysis of posttranslational modifications on the H3.1 histone tail [17]. Histones are proteins found in chromatin and serve as the core for DNA coils. It is hypothesized that particular combinations of posttranslational modifications, e.g., acetylation, methylation, and phosphorylation, on the histone tail at the amino terminus, form a code that is directly involved in gene regulation [25]. A 50-mer peptide from the amino terminus of H3.1 was isolated from human cells and subjected to ETD by fluoranthene anions in an ion-trap mass spectrometer. To reduce the charge state of the produced fragments, proton-transfer reactions were performed by benzoic acid anions. The resulting mass spectra showed a unique pattern of methylation sites that showed systematic variations during chromatographic separation. Correlating these modifications with gene expression data is instrumental in understanding the role of histone modifications in gene regulation.

2.6. Quantitative proteomics

Unlike nucleic acids, proteins in an organism are present at very different concentration levels. Thus, it is not sufficient to demonstrate that a particular protein is present; we also need to know its concentration. From the high-concentration globulins in blood to the low-copy-number proteins that are represented by only a few molecules per cell, there is an enormous dynamic range. This presents a challenge to the utilized analytical methods because of the potential interferences, especially when quantitating the proteins at low concentration. For example, the high-abundance proteins can compete in the ionization process and suppress the ion formation from the low-level species. This ion suppression effect is quite common in MALDI and ESI ion sources.

Common approaches to minimize these problems include extensive separation before mass spectrometric analysis. Typical separation protocols consist of an orthogonal combination of affinity chromatography, 2-DE, IEX, HPLC, and ion mobility techniques. If these steps can reduce the sample complexity to a single component, the signal from the separation method (e.g., chromatographic peak area) can be used for quantitation. Frequently this is not achievable or verifiable. Relative quantitation in these instances can be performed by stable isotope labeling methods.

A common example of relative quantitation is used in comparative proteomics. For example, to uncover the differences in protein makeup and concentration levels between the healthy state and a particular disease (e.g., protein expression in normal vs. HIV-infected cells [26]), stable isotope labeling can be applied to one or the other. A frequently used variant of this approach is the isotope-coded affinity tag (ICAT) method [27]. Fig. 7 shows how an ICAT reagent is used to tag the cysteine residues of a peptide, human insulin chain B in this example. First, the reactive end of the ICAT reagent covalently attaches to the

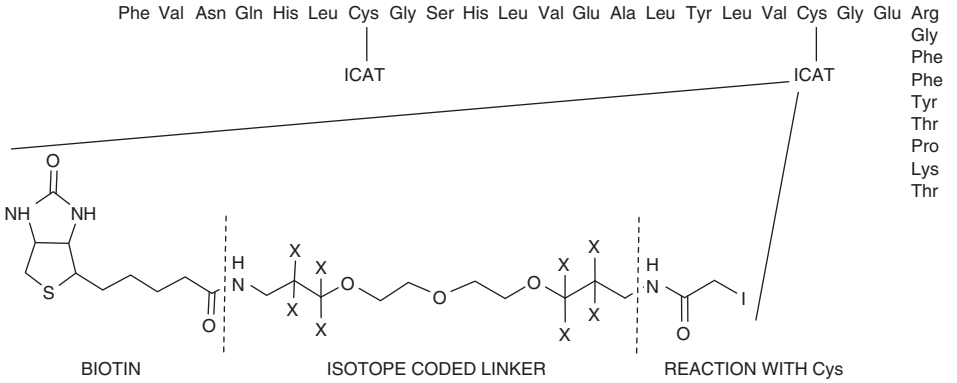


Fig. 7. Cysteine residues of human hemoglobin chain B are tagged with ICAT reagent. The coding of the linker, X, can be hydrogen (d_0-ICAT) for the normal sample and deuterium (d_8-ICAT) for the diseased sample.

cysteine residues through thiol chemistry. One form of the reagent, d_0-ICAT with no deuterium atoms, can be used to label the sample from the healthy source, whereas the other, d_8-ICAT with eight hydrogens in the linker replaced by deuterium, can designate the diseased sample. As a result the tagged peptides in the healthy and the diseased samples will exhibit a mass difference of 8 or its multiples depending on the number Cys residues. In the next step the two samples are combined and the biotin end of the ICAT reagent is used to separate the tagged peptides through affinity capture with avidin. This results in significantly reduced sample complexity.

The mass spectrum of the captured mixture exhibits the peptide peaks as doublets with a mass shift of 8 or, in case of multiple cysteine residues, its multiples between the normal and the diseased sample. The abundance ratios of these doublets characterize the relative quantity of a particular protein in the two samples. As both the d_0- and the d_8-tagged components are in the same matrix and differ only in isotope composition, the relative peak intensities are a true reflection of the protein level changes in disease. The ICAT method is limited to cysteine-containing proteins, but other tagging protocols (e.g., through proteolytic ^{18}O labeling) are being developed to eliminate this restriction [28].

2.7. Higher order structures

The efficiency of mass spectrometric methods in determining primary protein structure naturally leads to the question of their utility to characterize secondary, tertiary, and quaternary structures as well as the formation of noncovalent complexes. The success of mass spectrometry in approaching these problems is more limited. For example, there are some legitimate questions about the correspondence

of these structures between the native solution state and their ionized form in the gas phase. Are there significant structure changes as the molecule is ionized? How does the structure change when the molecule loses its solvation shell during volatilization?

It was noticed in the early 1990s that conformation changes, for example, due to pH changes, resulted in altered charge-state distributions in ESI spectra [29]. Although there are literature reports on successful deconvolution of these charge distributions to assess the relative weight of coexisting conformations (both secondary and tertiary structures) [30], the method is far from being routinely applicable. This approach hinges on the differences in the available protonation sites in a multiply charged ion in its folded and stretched conformations. When the molecule is folded, only the protonation sites exposed on its surface are accessible, whereas in its stretched conformation, at least in principle, all amenable sites should be ionized. Thus, unfolding of the molecule is reflected in a charge state distribution shifted to lower m/z values. Under limited conditions, folding and unfolding kinetics can also be followed measuring the time dependence of charge state distributions following a chemical perturbation (e.g., pH change) of the system.

Another method to study higher order protein structure is hydrogen–deuterium exchange [31]. When a protein molecule is dissolved in deuterium oxide, D_2O ("heavy water"), deuterium atoms start to exchange their accessible hydrogens. The resulting mass difference in the mass spectrum of the protein and its digestion products can reveal which part of the folded protein is accessible for the D_2O molecules.

Carbon-bound hydrogens do not exchange, whereas the exchange on the side chains of certain residues (e.g., Arg, Asn, Cys, and Trp) is very fast, essentially immediate on the timescale of the experiment. The exchange rate of amide hydrogens on the peptide backbone is between the two extremes and can be used to explore protein structure. The exchange rates of these amide hydrogens also depend on the pH and the temperature, so adjusting these parameters gives additional control. A typical experiment starts with exchanging the solvent to D_2O at pH 7.0 and at room temperature. This initiates the exchange of accessible amide hydrogens at the surface of the protein to deuterium. Changing the pH to 2.5 and the temperature to 0ºC arrests the exchange process and gives enough time to perform enzymatic digestion (typically with pepsin) followed by HPLC separation and mass spectrometry. A complicating factor is back exchange that can replace the deuterium already in the peptide fragments with hydrogen. This effect can be estimated and the results corrected for it.

The hydrogen–deuterium exchange method can be used to study secondary, tertiary, and even quaternary structures. Amide hydrogens in the hydrophobic core of the protein or at the interface of attached subunits are less accessible for the exchange reaction. Studying the kinetics of the exchange can reveal unfolding

dynamics and the association of partners in noncovalent complexes. The advantages of mass spectrometry over competing techniques used in combination with hydrogen–deuterium exchange (e.g., NMR) are the very low amount of protein required (~1 nmol) and the ability to tackle very large proteins including an entire 2.5 MDa ribosome and its subunits [32]. In addition, protein mixtures can also be studied with mass spectrometry.

Molecular recognition and noncovalent complexes are at the core of reaction networks in biology. Molecular complexes are often associated with the proliferation of disease (see, for example, the Tax-associated complexes in human T-cell leukemia type 1, HTLV-1 [33]). Along with other competing techniques (e.g., surface plasmon resonance), mass spectrometry can be successfully used to detect noncovalent complex formation. The corresponding ions can be present in both MALDI [34] and ESI [35] spectra, although the latter is used more often. A wide variety of protein–protein interactions as well as protein interactions with other species (nucleotides, carbohydrates, etc.) have been studied. The spectra can reveal the components of the complex and in some cases the association constant.

2.8. Mapping protein function

From the biomedical perspective, structural and kinetic studies are incomplete without determining the function of the protein. In the discussion of posttranslational modifications and noncovalent complexes, we have already indicated their important role in regulating the role a protein plays. In addition to biological function, protein-based drug and vaccine design also requires the elucidation of their mechanism of action. From heart disease to cancer, there are many examples in this volume showing the variety of implicated proteins [36]. Conversely, structural discrepancies in proteins are shown to result in disease states.

An interesting example of using mass spectrometry to unravel protein function is epitope mapping. In broad terms, an epitope is the binding site on the surface of a protein that attaches to another molecule; for example, to a monoclonal antibody. There are two general strategies to identify the epitope. In the first one the protein is attached to the antibody. Then, proteolytic digestion is performed that removes the nonattached parts of the protein. Mass spectrometric analysis of the removed fragments and the segment retained on the antibody can reveal the epitope. In the second strategy, the studied protein is digested first and the resulting mixture is affinity separated by the monoclonal antibody. The protein fragment that contains the epitope is preferentially captured [37].

Even if the participating protein segments are discontinuous, epitopes can also be identified by hydrogen–deuterium exchange. The components of the noncovalent complex are deuterated in D_2O environment and allowed to react. When the

solvent is changed to water, the amide deuterium atoms on the exposed surface of the formed complex are exchanged with hydrogen. The epitope region, however, is not affected because it is not exposed. Displacing the protein from the complex followed by pepsin digestion produces peptides that are deuterated at the epitope. The resulting mass differences can be detected by MALDI mass spectrometry [38].

Although epitope mapping can contribute an important piece of the puzzle, identifying protein function requires a more complex approach. The available subset of genetic, X-ray diffraction, NMR, and mass spectrometric data has to be considered in its entirety to shed light on the function of newly discovered proteins [39]. Often similarity searches in genomic and proteomic databases can provide an initial hypothesis based on homology with proteins of known function. For example, proteomic analysis of the *Torpedo californica* electric organ, a large-scale model for the neuromuscular junction, identified 11 human open reading frames coding for proteins of unknown function [40]. When similarity is not found, high-resolution structures (X-ray and NMR data) as well as mass spectrometric study of noncovalent complexes can be used to identify active sites and infer the possible functions of the protein.

3. Outlook

In the past few years we have witnessed the explosive growth in the field of proteomics. During this period, proteomics has captured the attention of academia, government, and industry alike. At the universities, new courses are being introduced to teach the related technologies and applications for the emerging generation of biomedical professionals. Government funding in developed countries is increasingly available in the proteomics field. The landscape of mass spectrometer manufacturing has been reordered by the technological demands of proteomics; reagent, diagnostic, and pharmaceutical vendors gear up to take advantage of the new market opportunities.

This dramatic new focus was already clearly discernable from the presentations at the 2002 symposium organized by the U.S. National Academies, *Defining the Mandate of Proteomics in the Post-Genomics Era* as well as from the launching of three dedicated journals, *Journal of Proteome Research, Molecular and Cellular Proteomics,* and *Proteomics*. Learning from the lessons of the Human Genome Project, it was clear from the outset that international efforts had to be coordinated. In 2001 an international consortium, the Human Proteome Organization (HUPO), was launched to facilitate several initiatives, including projects related to the proteomes of the liver, brain, and plasma, to the development of proteomics standards, and to mouse models of human disease [41].

Despite its short history, the field of proteomics has already started to differentiate. Beyond the basic distinction between methods, including instrumentation and

bioinformatics, and applications to biomedical problems of interest, more or less coherent subfields are beginning to appear. Among them are proteomics within the subdisciplines of biology (e.g., proteomics in cell biology and microbiology, plant proteomics, and animal proteomics) as well as proteomics in the medical fields (e.g., the proteomics of a certain organ or disease). As the discovery of disease-related protein biomarkers continues, proteomics is poised to become an everyday tool in clinical diagnostics and serve as a basis for new therapies.

References

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H.M., Yu, J., Wang, J., Huang, G.Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S.Z., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H.Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W.H., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J.R., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A. and Morgan, M.J., Initial sequencing and analysis of the human genome, *Nature*, **409**, 860–921 (2001).

2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.Q.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J.H., Miklos, G.L.G., Nelson, C., Broder, S., Clark, A.G., Nadeau, C., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z.M., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W.M., Gong, F.C., Gu, Z.P., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z.X., Ketchum, K.A., Lai, Z.W., Lei, Y.D., Li, Z.Y., Li, J.Y., Liang, Y., Lin, X.Y., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B.X., Sun, J.T., Wang, Z.Y., Wang, A.H., Wang, X., Wang, J., Wei, M.H., Wides, R., Xiao, C.L., Yan, C.H., Yao, A., Ye, J., Zhan, M., Zhang, W.Q., Zhang, H.Y., Zhao, Q., Zheng, L.S., Zhong, F., Zhong, W.Y., Zhu, S.P.C., Zhao, S.Y., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H.J., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H.Y., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X.J., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M.Y., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X.H., The sequence of the human genome, *Science*, **291**, 1304–1351 (2001).
3. Siuzdak, G., *The Expanding Role of Mass Spectrometry in Biotechnology* (2nd edition). MCC Press, San Diego, CA, 2003.
4. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse, C.M., Electrospray Ionization for Mass-Spectrometry of Large Biomolecules, *Science*, **246**, 64–71 (1989).
5. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y. and Yoshida, T., Protein and Polymer Analyses up to m/z 100000 by Laser Ionization Time-of-flight Mass Spectrometry, *Rapid Commun. Mass Spectrom.*, **2**, 151–153 (1988).
6. Karas, M. and Hillenkamp, F., Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10000 Daltons, *Anal. Chem.*, **60**, 2299–2301 (1988).
7. See the web site at http://nobelprize.org/nobel_prizes/chemistry/laureates/2002/index.html (accessed February 28, 2007).

8. Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., MacInnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J. and Querengesser, L., HMDB: the human metabolome database, *Nucleic Acids Res.*, **35**, D521–D526 (2007).
9. Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C. and Watanabe, C., Identifying Proteins from 2-Dimensional Gels by Molecular Mass Searching of Peptide-Fragments in Protein-Sequence Databases, *Proc. Nat. Acad. Sci. U.S.A.*, **90**, 5011–5015 (1993).
10. Yates, J.R., Speicher, S., Griffin, P.R. and Hunkapiller, T., Peptide Mass Maps—a Highly Informative Approach to Protein Identification, *Anal. Biochem.*, **214**, 397–408 (1993).
11. See the web site at http://prospector.ucsf.edu/ (accessed March 14, 2007).
12. Clauser, K.R., Baker, P. and Burlingame, A.L., Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS MS and database searching, *Anal. Chem.*, **71**, 2871–2882 (1999).
13. Spengler, B., Kirsch, D., Kaufmann, R. and Jaeger, E., Peptide Sequencing by Matrix-Assisted Laser-Desorption Mass-Spectrometry, *Rapid Comm. Mass Spectrom.*, **6**, 105–108 (1992).
14. Wei, J., Buriak, J.M. and Siuzdak, G., Desorption-ionization mass spectrometry on porous silicon, *Nature*, **399**, 243–246 (1999).
15. Chen, Y. and Vertes, A., Adjustable fragmentation in laser desorption/ionization from laser-induced silicon microcolumn arrays, *Anal. Chem.*, **78**, 5835–5844 (2006).
16. Baeten, W., Claereboudt, J., Vandenheuvel, H. and Claeys, M., Tandem Mass-Spectrometry of Leucine Enkephalin and Physalaemin Using a Hybrid Instrument, *Biomed. Environ. Mass Spectrom.*, **18**, 727–732 (1989).
17. Coon, J.J., Ueberheide, B., Syka, J.E.P., Dryhurst, D.D., Ausio, J., Shabanowitz, J. and Hunt, D.F., Protein identification using sequential ion/ion reactions and tandem mass spectrometry, *Proc. Nat. Acad. Sci. U.S.A.*, **102**, 9463–9468 (2005).
18. Mann, M. and Wilm, M., Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags, *Anal. Chem.*, **66**, 4390–4399 (1994).
19. Mortz, E., Oconnor, P.B., Roepstorff, P., Kelleher, N.L., Wood, T.D., McLafferty, F.W. and Mann, M., Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases, *Proc. Nat. Acad. Sci. U.S.A.*, **93**, 8264–8267 (1996).
20. Hunt, D.F., Henderson, R.A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A.L., Appella, E. and Engelhard, V.H., Characterization of Peptides Bound to the Class-I Mhc Molecule Hla-A2.1 by Mass-Spectrometry, *Science*, **255**, 1261–1263 (1992).
21. Zubarev, R.A., Kelleher, N.L. and McLafferty, F.W., Electron capture dissociation of multiply charged protein cations. A nonergodic process, *J. Am. Chem. Soc.*, **120**, 3265–3266 (1998).
22. Zubarev, R.A., Horn, D.M., Fridriksson, E.K., Kelleher, N.L., Kruger, N.A., Lewis, M.A., Carpenter, B.K. and McLafferty, F.W., Electron capture dissociation for structural characterization of multiply charged protein cations, *Anal. Chem.*, **72**, 563–573 (2000).
23. Shi, S.D.H., Hemling, M.E., Carr, S.A., Horn, D.M., Lindh, I. and McLafferty, F.W., Phosphopeptide/phosphoprotein mapping by electron capture dissociation mass spectrometry, *Anal. Chem.*, **73**, 19–22 (2001).
24. Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J. and Hunt, D.F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry, *Proc. Nat. Acad. Sci. U.S.A.*, **101**, 9528–9533 (2004).
25. Jenuwein, T. and Allis, C.D., Translating the histone code, *Science*, **293**, 1074–1080 (2001).

26. Berro, R., de la Fuente, C., Klase, Z., Kehn, K., Parvin, L., Pumfrey, A., Agbottah, E., Vertes, A., Nekhai, S. and Kashanchi, F., Identifying the membrane proteome of HIV-1 latently-infected cells, *J. Biol. Chem.*, **282**, 8207–8218 (2007).
27. Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, *Nat. Biotechnol.*, **17**, 994–999 (1999).
28. Yao, X.D., Freas, A., Ramirez, J., Demirev, P.A. and Fenselau, C., Proteolytic O-18 labeling for comparative proteomics: Model studies with two serotypes of adenovirus, *Anal. Chem.*, **73**, 2836–2842 (2001).
29. Chowdhury, S.K., Katta, V. and Chait, B.T., Probing Conformational-Changes in Proteins by Mass-Spectrometry, *J. Am. Chem. Soc.*, **112**, 9012–9013 (1990).
30. Mohimen, A., Dobo, A., Hoerner, J.K. and Kaltashov, I.A., A chemometric approach to detection and characterization of multiple protein conformers in solution using electrospray ionization mass spectrometry, *Anal. Chem.*, **75**, 4139–4147 (2003).
31. Wales, T.E. and Engen, J.R., Hydrogen exchange mass spectrometry for the analysis of protein dynamics, *Mass Spectrom. Rev.*, **25**, 158–170 (2006).
32. Benjamin, D.R., Robinson, C.V., Hendrick, J.P., Hartl, F.U. and Dobson, C.M., Mass spectrometry of ribosomes and ribosomal subunits, *Proc. Nat. Acad. Sci. U.S.A.*, **95**, 7391–7395 (1998).
33. Wu, K.L., Bottazzi, M.E., de la Fuente, C., Deng, L.W., Gitlin, S.D., Maddukuri, A., Dadgar, S., Li, H., Vertes, A., Pumfery, A. and Kashanchi, F., Protein profile of tax-associated complexes, *J. Biol. Chem.*, **279**, 495–508 (2004).
34. Tang, X.D., Callahan, J.H., Zhou, P. and Vertes, A., Noncovalent Protein-Oligonucleotide Interactions Monitored By Matrix-Assisted Laser Desorption/Ionization Mass-Spectrometry, *Anal. Chem.*, **67**, 4542–4548 (1995).
35. Loo, J.A., Electrospray ionization mass spectrometry: a technology for studying noncovalent macromolecular complexes, *Int. J. Mass Spectrom.*, **200**, 175–186 (2000).
36. Banks, R.E., Dunn, M.J., Hochstrasser, D.F., Sanchez, J.C., Blackstock, W., Pappin, D.J. and Selby, P.J., Proteomics: new perspectives, new biomedical opportunities, *Lancet*, **356**, 1749–1756 (2000).
37. Zhao, Y.M. and Chait, B.T., Protein Epitope Mapping by Mass-Spectrometry, *Anal. Chem.*, **66**, 3723–3726 (1994).
38. Baerga-Ortiz, A., Hughes, C.A., Mandell, J.G. and Komives, E.A., Epitope mapping of a monoclonal antibody against human thrombin by R/D-exchange mass spectrometry reveals selection of a diverse sequence in a highly conserved protein, *Protein Sci.*, **11**, 1300–1308 (2002).
39. Eisenstein, E., Gilliland, G.L., Herzberg, O., Moult, J., Orban, J., Poljak, R.J., Banerjei, L., Richardson, D. and Howard, A.J., Biological function made crystal clear-annotation of hypothetical proteins via structural genomics, *Curr. Opin. Biotechnol.*, **11**, 25–30 (2000).
40. Nazarian, J., Hathout, Y., Vertes, A. and Hoffman, E.P., The proteome survey of an electricity-generating organ (***Torpedo californica*** electric organ), *Proteomics*, **7**, 617–627 (2007).
41. See the web site at http://www.hupo.org/ (accessed March 30, 2007).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 9

De novo sequencing of peptides

MATTHEW T. OLSON, JONATHAN A. EPSTEIN, and ALFRED L. YERGEY*

National Institute of Child Health and Human Development, NIH, Bethesda, MD, USA

As described elsewhere in this volume and in a number of excellent reviews including the recent one by Steen and Mann [1], the typical strategy for mass spectrometric identification of proteins employs a combination of peptide mass fingerprinting (PMF) and peptide fragment ion spectra to search databases of protein sequences and look for high-probability matches. Modern algorithms search protein databases for sequences that most consistently match the spectral data. Resulting sequences are ranked according to the statistical significance assigned to the proteins found by the search [2,3].

One must ask however, "What if the protein I am seeking to identify is not likely to be found in a database?" Potential reasons for such an absence include proteins isolated from organisms for which genomes have not been sequenced at this time—such as sea urchin, *S. purpuratus,* and the polyploidal frog *X. laevis*—as well as mutations and splice variants of otherwise well-characterized proteins. It might be argued that the last two cases could possibly be addressed by the so-called homology exploration options of the better known search algorithms, but proteins from organisms with unsequenced genomes are not likely to be reliably identified by this approach. Furthermore, independent validation of sequences may be necessary for novel or rare peptides. In order to address these and similar essentially insoluble problems, it becomes the task of investigators to deduce a peptide's sequence purely from mass spectral fragmentation data. This is *de novo* sequencing.

Fig. 1 is an illustration of the need for *de novo* sequencing. The portion of the protein sequence shown in the box corresponds to the mass of a peptide produced by proteolysis. Above this sequence is shown a series of three amino acid residues that have been found to be present in a fragmentation spectrum, a sequence tag [4]. Note that, in general, the order of the residues found is not known, i.e., they could

*Corresponding author. Fax: +1-301-480-5793. E-mail: aly@helix.nih.gov.

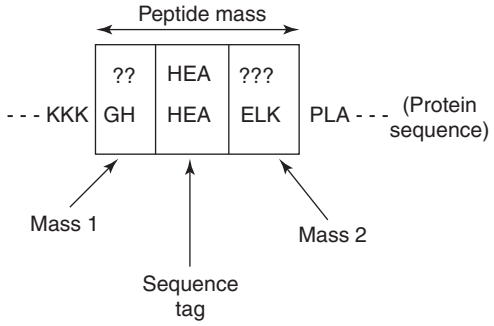


Fig. 1. Schematic of matching peptide fragments.

be in the order shown or the reverse of that order. Nevertheless, this tag is a powerful tool in that it not only provides information about a series of amino acid residues but also contains information about the peptide's total mass and the mass of those portions of the peptide on both the C- and N-terminal portions of the three residue tags. The tag is used to search a database for the best match to all of the information, i.e., sequence, peptide mass, and the mass of the portions of the peptide outside of the identified residues. For well-defined protein systems, this is a very effective approach. However, close inspection of the figure shows several potential problems, and even if we assume that the peptide matched is the "correct one," we must continue to be aware of potential problems with the "hit." First, since there are apparently no fragmentation data present in the spectrum to extend the area of the tag, there can be no definitive proof for the sequence of the residues on either terminal of the peptide. Second, because of the same lack of evidence in the spectrum, it is possible to imagine a number of isobaric alternatives to the sequence "found" by a search algorithm. For example, the two residues GH that are taken to be part of the peptide "identified" have nearly the same mass as that of the two residues PP, 213.124 and 213.099 Da, respectively. This 0.025 Da difference in mass can be used very effectively as shown below, but the most commonly used instruments producing MS/MS spectra are ion traps and triple quadrupoles, neither of which have mass accuracies better than 0.5–1 Da. Thus, what is routinely ranked as an identified peptide is seen to have a number of weak points in terms of its true identity.

There are a number of substantial technical issues involved in *de novo* sequencing that arise from the fact that peptides do not fragment in an ideal manner. One result of this is that skilled spectral interpreters have devised sets of rules that can be used to convert mass differences between ions of a spectrum into amino acid sequences; unfortunately, these rules are complicated and are not always followed. In addition, the fragmentations do not occur in a manner that gives rise to uniform ion intensities. This phenomenon results in spectra that have a substantial range of intensities which, depending on the ionization and mass analyzer used,

can lead to spectra in which important information cannot be distinguished from baseline noise. As a consequence of these physical realities, *de novo* sequencing remains a challenging problem, and interpretations unaided by a computer can take substantial time with no guarantee of a correct result (or even a meaningful one). As instruments developed the capability to generate very large numbers of spectra, often without trained mass spectrometrists, the possibility for such manual interventions became much smaller. It became clear in the late 1980s that it would be desirable to incorporate the knowledge of skilled mass spectrometrists into computer programs that would not only save them time but also provide solutions for investigators who were less experienced in this field.

There are roughly two ways to view the relationship between *de novo* sequencing and database search algorithms—the first is complementary, and the second is alternative. Approaches that emphasize the complementary relationship between *de novo* sequences and database searches treat *de novo* sequences as a means to enhance the quality and reliability of database searches. Regimes originating from this approach utilize the spectra to derive one or more highly reliable sequence tags. These tags guide the database matching process, and since they ostensibly represent the most prominent and reliable features of the spectra, the tags also impart a higher degree of confidence to the database results. These methods do not require a complete *de novo* formulation of the peptide sequence prior to database searching, but a complete sequence may be yielded occasionally by these approaches and is highly desirable. In contrast, approaches that attempt to preclude the requirement for a database search focus exclusively on the spectral data. This approach requires that the *de novo* algorithm generate a complete sequence, hence the name "complete *de novo* sequencing." In general, regimes originating from this approach are more difficult to implement in the end because they lack a database search to verify the sequence information. Nevertheless, they remain an ultimate goal for some of the challenges of proteomics research described earlier. However, due to the difficulty of obtaining independent, complete, and reliable *de novo* sequences, essentially all of the mainstream *de novo* sequencing packages are implementations of the partial sequencing approach and are used to complement database searches.

De novo sequencing approaches also differ according to the way in which they determine the fragment type (y or b, etc.) and score the sequence information. These differences in approach tend to generate great differences between *de novo* sequences obtained even from the same spectra. Because of the diverse, incomplete, and extremely complex fragmentation patterns, better understanding of these fundamental issues is of central importance to ranking the accuracy of *de novo* sequencing regimes. If the fragmentation process were uniform and complete or even simple, there would be no problem because a score could be derived from a normalized sum of total spectral intensity. Perhaps the simplest and most widely used solution is the assignment of arbitrary weighting factors to the intensities of

peaks based on the ion type they are determined to be. For example, if the peptide is fragmented by unimolecular decomposition, then y and b ions are far more common than the multitude of other possibilities (a, x, w, c, or d ions), so the intensities of peaks determined to be of the rare ion types are decreased by an arbitrary scalar in the sum of peak intensities which eventually factors into the score of the sequence [5,6]. Another means of scoring the *de novo* sequence involves the use of an empirical function that adjusts the measured intensities according to the intensities observed in other spectra [7]. Yet another method involves simulated fragmentations of peptides and subsequent matching of the observed for a match of the fragmentation pattern [8,9]. Unfortunately, while each of these methods reports high efficacy and accuracy, a broad comparison to determine a superior approach or optimal usage criteria has not been performed. The complex nature of arbitrary scores, empirical functions, and fragmentation simulations infers a degree of instrumental specificity that confounds the difficulty of such comparisons. Although such complexity will undoubtedly prove useful in a universal solution, if and when it exists, the authors of this chapter are of the opinion that the weighting schemes of the aforementioned methods, though they are complex, do not fully model or predict the spectra and lead to a high frequency of sequence errors. One approach [6] is noteworthy because, while it uses arbitrary ion weighting, it circumvents the problem with accurate mass evaluation of ion type. The value of this remains to be proven in coming years.

There have been two principal approaches to the implementation of partial *de novo* sequencing, and these have been well defined by Pevzner and colleagues [10] and termed by them to be the global and local paradigms. In the most general terms, the global implementations are those in which theoretical spectra for all peptides of a given mass are generated initially and then the observed spectra matched against them for the best fit. This approach was described initially by Sakurai et al. [11]. Clearly the generation of theoretical spectra for all possible peptides for a given mass is a huge task that increases exponentially in complexity with peptide mass. This reality led later workers to devise methods to prune the number of theoretical possibilities [12–14], typically by calculating a small subset of possible extensions to ions present in the spectrum, matching observed ions in the mass range of these new subsequences and then computing further extensions to the highest scoring subsequences. Scoring of the matches was typically done by incorporating some subset of the knowledge-based rules for peptide fragmentation into their programs. Perhaps the most successful of these approaches was that developed by Johnson and Biemann that demonstrated the ability to sequence peptides from a variety of sources without regard to proteolysis method [14].

The local approaches tend to be somewhat less computationally intensive in that they filter the spectral data in some fashion prior to any evaluation of candidate sequences. The various local approaches [15,16,5,10,17] then employ

an algorithm to implement a graph theory approach to determine the amino acid sequence. The filtered spectral data peaks as vertices in a graph with the edges of the graph as the connecting links between them; each peak in the spectrum could possibly originate from a different ion type, i.e., y, b, a, neutral loss, etc., and so it might be possible to have several ion-type graphs within a single spectrum. This possibility can be eliminated by converting all of the peaks into an ion of a specific type, i.e., C-terminal (y series) or N-terminal (b series). Fig. 2 illustrates the overall concept with a graph for y series ions in a hypothetical spectrum. Thus, the task of an algorithm is to find the longest possible acyclic path among the spectral vertices.

From this brief discussion it is clear that, though high-quality spectra are not absolutely necessary to obtain at least some results, the very best results will be obtained from spectra that have the highest possible values of signal-to-noise ratio, which provide fragmentations that are as complete as possible representations of full coverage of the peptide(s) being considered. Until very recently, another fundamental aspect of mass spectra has not been given an appropriate level of attention. That is, previously algorithms used in the construction of graphs or for the calculations of the global approaches tended to ignore the deviations of measured masses from integer values, i.e., the mass defect. Although there was

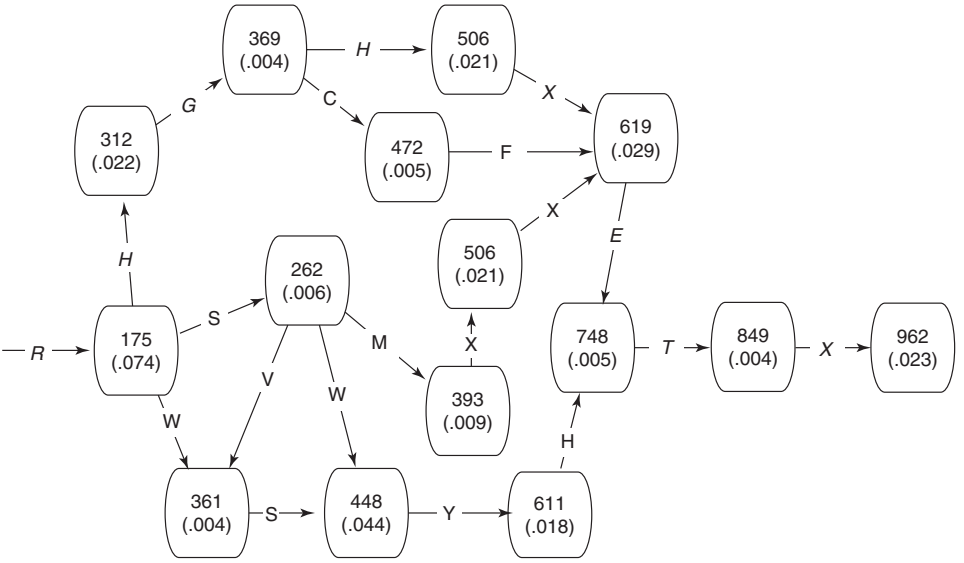


Fig. 2. Graphical solution of *de novo* sequence. Peptide masses, corresponding to graph vertices, are shown within boxes with relative intensities in parentheses. Graph edges for different paths are represented by the amino acid residues allowing vertices to be connected. All paths shown are complete, but the one with the highest score is shown by the edges labeled in italics, XTEXHGHR, where X is either Ile or Leu.

never doubt that good mass accuracy was an important aspect of any sequence determination, it was not until the recent work of Spengler [6] that this parameter came to the fore as one to be employed prospectively in sequence determinations. In this work, the approach taken was to use the very high mass accuracy possible with Fourier transform mass spectrometry in conjunction with what might be termed a hybrid local–global approach. That is, a set of potential sequences is developed from using an algorithm based on filtered spectra. These sequences are then evaluated using a semiglobal approach in which each sequence has its fragmentation pattern calculated to a level of mass accuracy of about 1 ppm, i.e., to 0.001 Da at *m/z* 1000. This approach appears to give substantial improvement in the confidence level of a sequence generated *de novo,* but, as described in the original paper, has been used to date principally for improving the confidence level of database search matches.

More recently, the authors of this chapter have developed an extension of Spengler's approach in that it is also something of a local–global hybrid approach that employs the mass defect of fragment ions. This new approach has been demonstrated to be effective at somewhat lower levels of mass accuracy [18] and is to some extent also an extension of earlier work from this group [19]. This approach has been employed to date only on MALDI fragment ion spectra that are generated by a tandem TOF instrument. The somewhat more extensive and complete fragmentation resulting from this technique permits the use of mass accuracies of about 0.05 Da. More fundamental to this approach, however, is the use of a database consisting of an exhaustive listing of all amino acid combinations giving rise to peptides up to and including 2000 Da. By using a combination of prefiltering of the spectra and an extension of bit-mapping algorithm, the authors have shown the capability of generating reliable sequences *de novo*. Although the utility of this approach is yet to be fully evaluated, it appears from preliminary evaluations that it may prove very useful for generating complete peptide sequences.

De novo peptide sequencing has been shown to be a useful tool particularly with regard to improving the reliability of database searching algorithms, but in many respects it remains an open problem with a great deal of work yet to be done in order to make it widely useful for the characterization of peptides from organisms with incomplete or poorly characterized genomes and as a robust technique for probing novel posttranslational splicing patterns. At this point in time, it is not altogether clear whether the difficulty in having a completely effective algorithm for complete *de novo* sequencing of peptides not present in a database is a consequence of computational complexity, inability to achieve reliable complete peptide fragmentation, or difficulty in routinely providing adequate mass accuracy in fragmentation spectra, or indeed, some combination of all of these factors. Until a clear understanding of all of these factors is achieved, this problem is likely to remain incompletely solved.

References

1. Steen, H. and Mann, M., The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, **5**, 699–711 (2004).
2. Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567 (1999).
3. Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W. and Bryant, S.H., Open mass spectrometry search algorithm. *J. Proteome Res.*, **3**, 958–964 (2004).
4. Mann, M. and Wilm, M., Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.*, **66**, 4390–4399 (1994).
5. Taylor, J. and Johnson, R., Sequence database search via *de Novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **11**, 1067–1075 (1997).
6. Spengler, B., *De novo* sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by Fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **15**, 703–714 (2004).
7. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A. and Lajoie, G., PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **17**, 2337–2342 (2003).
8. Zhang, Z. and McElavin, J.S., *De novo* peptide sequencing by two-dimensional fragment correlation mass spectrometry. *Anal. Chem.*, 72, 2337–2350 (2000).
9. Kapp, E.A., Schutz, F., Reid, G.E., Eddes, J.S., Moritz, R.L., O'Hair, R.A., Speed, T.P. and Simpson, R.J., Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.*, **75**, 6251–6264 (2003).
10. Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E. and Pevzner, P.A., *De novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, **6**(**3/4**), 327–342 (1999).
11. Sakurai, T., Matsuo, T., Matsuda, H. and Katakuse, I., Paas 3: a computer program to determine probable sequence of peptides from mass spectrometric data. *Biomed. Environ. Mass Spectrom.*, **11**, 396–399 (1984).
12. Ishikawa, K. and Niva, Y., Computer-aided peptide sequencing by fast atom bombardment mass spectrometry. *Biomed. Environ. Mass Spectrom.*, **13**, 373–380 (1986).
13. Siegel, M. and Bauman, N., An efficient algorithm for sequencing peptides using fast atom bombardment mass spectral data. *Biomed. Environ. Mass Spectrom.*, **15**, 333–343 (1988).
14. Johnson, R. and Biemann, K., Computer Program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed. Environ. Mass Spectrom.*, **18**, 945–957 (1989).
15. Bartels, C., Fast algorithm for peptide sequencing by mass spectroscopy. *Biomed. Environ. Mass Spectrom.*, **19**, 363–368 (1990).
16. Hines, W.M., Falick, A.M., Burlingame, A.L. and Gibson, B.W., Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. *J. Am. Soc. Mass Spectrom.*, **3**, 326–336 (1992).
17. Chen, T., Kao, M.-Y., Tepel, M., Rush, J. and Church, G., A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, **8(3)**, 325–337 (2001).
18. Olson, M., Epstein, J. and Yergey, A., De novo peptide sequencing using exhaustive enumeration of peptide composition. *J. Am. Soc. Mass Spectrom.*, **17**(**8**), 1041–1049 (2006).
19. Yergey, A.L., Coorssen, J.R., Backlund, P.S., Blank, P.S., Humphrey, G.A., Zimmerberg, J., Campbell, J.M. and Vestal, M.L., *De novo* sequencing of peptides using MALDI/TOF-TOF. *J. Am. Soc. Mass Spectrom.*, **10**, 784–791 (2002).

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 10

Protein bioinformatics

PETER McGARVEY, HONGZHAN HUANG, and CATHY H. WU*

Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Reservoir Road, NW, Washington, DC 20057-1455, USA

1. Highlights for medical professionals

Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. Proteomics methods used in mass spectrometry require databases of protein sequences and post-translational modifications as well as algorithms and tools to match spectra to peptides and peptides to proteins. Following identification of a protein, further interpretation and knowledge discovery comes from the integration of protein sequence data

*Corresponding author. Fax: +1-202-687-0057. E-mail: wuc@georgetown.edu (C.H. Wu).

with all forms of additional biomedical data contained in various databases. Here we review some of the key integrated datasets, tools, and methods used in this discovery process.

The Universal Protein Resource (UniProt) provides the scientific community with a centralized, authoritative resource for protein sequences and functional information with three database components. (1) The *UniProt Knowledgebase* (UniProtKB), produced by a combination of automation and over 25 years of human curation, is the central protein sequence database with accurate, consistent, functional annotation and extensive cross-references. (2) The *UniProt Reference Clusters* (UniRef) provide clustered sets of sequences from UniProtKB (including splice variants and isoforms) in order to obtain complete coverage of sequence space at several resolutions. The UniRef100 database is particularly useful for Mass Spec identifications as it exposes known sequence variation and splice-form annotation contained in UniProtKB records. (3) The *UniProt Archive* (UniParc) provides a stable comprehensive sequence collection by storing the complete body of all publicly available protein sequence data.

The Protein Information Resource (PIR) (http://pir.georgetown.edu/) is an integrated public bioinformatics resource supporting genomic and proteomic research. PIR provides access to all the UniProt databases and complementary databases including iProClass, which provides an integrated view of protein information from over 90 databases and serves as a bioinformatics framework for data integration and associative analysis of proteins and PIRSF, an annotated family database based on the PIRSF classification system, which applies a network structure for protein classification from superfamily to sub-family levels.

Even the most up-to-date databases and tools lag behind actual research results by months or years because human reading of the scientific literature is required. In the future more automated rule-based systems will take the lead in data analysis and integration by linking existing protein knowledge to new experimental data almost as soon as they are published or even prior to publication. Some efforts under development include iProLink, which has tools and resources for automated literature mining, and iProXpress, where data produced by high-throughput proteomics research can feed into automated analysis and annotation pipeline. However, no one database or institution can keep up with the flood of new biological information. Further efforts on integration of a wider array of literature, data, and analysis tools require community efforts to develop and utilize common standards for data exchange, ontologies, and object models. Some prominent community efforts include the Human Proteome Organization (HUPO) Protein Standards Initiative (PSI) for Proteomics, the Microarray Gene Expression Data (MGED) Society for gene expression data, the National Center for Biomedical Ontology, and the National Cancer Institute's Cancer Biomedical Informatics Grid (caBIG) initiative which hopes to combine many of the current community efforts into a semantically interoperable grid of database and software resources.

2. Introduction

Bioinformatics can be defined as the field of science in which biology, computer science, and information technology merge into a single discipline. Bioinformatics contains a number of important sub-disciplines including: development of new algorithms and statistics; the analysis and interpretation of data; development of tools that mine and manage various types of information; and database development and data integration. All these sub-disciplines have played a role in developing the mass spectrometry methods reviewed in this book. As described in this volume and elsewhere the high-throughput proteomics methods used in mass spectrometry require accurate databases of both protein sequences and post-translational modifications as well as algorithms and tools to match spectra to peptides and peptides to proteins [1,2]. After identification of a protein, further interpretation and knowledge discovery come from the integration of protein sequence data with all forms of additional biomedical data. There are many approaches to data integration and the field is evolving as different approaches and data collections merge. Here we describe our bottom-up approach at data integration, starting with protein sequence information and bringing in a wide variety of structural, functional, genetic, and disease information related to proteins. We also discuss some future efforts to link this information to other data collections and broader community efforts and approaches to data integration.

High-throughput genome and proteome projects have resulted in the rapid accumulation of genome sequences for a large number of organisms. Meanwhile, scientists have begun to systematically tackle other complex regulatory processes by studying organisms at the global scale of transcriptomes (RNA and gene expression), metabolomes (metabolites and metabolic networks), interactomes (protein–protein interactions), and physiomes (physiological dynamics and functions of whole organisms). Associated with the enormous quantity and variety of data being produced is the growing number of databases that are being generated and maintained. Meta databases (databases of databases) have been compiled to catalog and categorize these databases, such as the Molecular Biology Database Collection [3]. This online collection (http://www.oxfordjournals.org/nar/database/cap/) lists over 700 key biological databases that add new value to the underlying data by virtue of curation, provide new types of data connections, or implement other innovative approaches to facilitate biological discovery. Based on the type of information they provide, these databases can be conveniently classified into sub-categories. Examples of major database categories include genomic sequence repositories (e.g., GenBank [4]), gene expression (e.g., SMD [5]), model organism genomes (e.g., MGD [6]), mutation databases (e.g., dbSNP [7]), RNA sequences (e.g., RDP [8]), protein sequences (e.g., UniProt [9]), protein family (e.g., InterPro [10]), protein structure (e.g., PDB [11]), intermolecular interactions (e.g., BIND [12]), metabolic pathways and cellular regulation (e.g., KEGG [13]), and taxonomy (e.g., National Center for Biotechnology Information (NCBI) taxonomy [14]).

To fully explore these datasets, advanced bioinformatics infrastructures must be developed for biological knowledge extraction and management. The PIR [15] is an integrated bioinformatics resource that supports genomic and proteomic research in this manner. PIR is a member of UniProt—the world's most comprehensive catalog of information on proteins, which unifies the previously separate PIR, Swiss-Prot, and TrEMBL databases [9]. The core resources and bioinformatics framework for large-scale proteomic data mining at PIR include: the UniProtKB of all known proteins; iProClass [16] database integrating information from over 90 biological databases; PIRSF classification-driven and rule-based system for protein functional annotation [17,18]; iProLINK [19] literature mining resource; and some new tools for proteomics data analysis and target identification.

3. Methodology

3.1. UniProt sequence databases

The UniProt provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information with three database components, each addressing a key need in protein bioinformatics. The *UniProtKB* is the central protein sequence database with accurate, consistent, and rich sequence and functional annotation, full classification, and extensive cross-references. Produced by a combination of automated and over 25 years of human curation, the annotations in UniProtKB include protein name and function, taxonomy, enzyme-specific information (catalytic activity, cofactors, metabolic pathway, regulation mechanisms), domains and sites, post-translational modifications, sub-cellular locations, tissue- or developmentally-specific expression, interactions, splice isoforms, polymorphisms, diseases, and sequence conflicts. The *UniParc* provides a stable and comprehensive sequence collection by storing the complete body of publicly available protein sequence data. While a protein sequence may exist in multiple databases, UniParc stores each unique sequence only once and assigns it a unique UniParc identifier. Cross-references back to the source databases are provided and include source accession numbers, sequence versions, and status (active or obsolete). The archive thus provides a history of protein sequences. The *UniRef* provides clustered sets of sequences from UniProtKB (including splice variants and isoforms) and selected UniParc records, in order to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences from view. The sequence compression is achieved by merging sequences and sub-sequences that are 100% (UniRef100), 90% (UniRef90), or 50% (UniRef50) identical, regardless of source organism. Removing sequence redundancy in UniRef90 and UniRef50 speeds sequence computational methods, e.g., similarity searches, while rendering such searches more informative. UniRef100

is particularly useful for Mass Spec identifications as it exposes known sequence variation and splice-form annotation contained in the Swiss-Prot section of UniProtKB. The UniProt databases can be accessed online at http://www.uniprot.org/ or downloaded in several formats (ftp://ftp.uniprot.org/pub). New releases are published every two weeks.

3.2. PIRSF protein family classification

The PIRSF family classification system applies a network structure for protein classification from superfamily to sub-family levels on the UniProtKB [17]. The primary PIRSF classification unit is the homeomorphic family whose members are homologous (sharing common ancestry) and homeomorphic (sharing full-length sequence similarity with common domain architecture). PIRSF classification considers both full-length similarity and domain architecture, discriminates between single- and multi-domain proteins, and shows functional differences associated with the presence or absence of one or more domains. For example, the relationship between domain architecture and function can be illustrated by the various types of response regulator proteins that share the CheY-like phosphoacceptor domain (Pfam domain PF00072) (Fig. 1) and are involved in signal transduction by two-component signaling systems. These response regulators usually consist of an N-terminal CheY-like receiver domain and a C-terminal output (usually DNA-binding) domain. In addition to the "classical" well-known response regulators (e.g., PIRSF003173 with the winged helix–turn–helix DNA-binding domain), bacterial genomes encode a variety of response regulators with other types of DNA-binding domains (e.g., PIRSF006198, PIRSF036392), RNA-binding domain (PIRSF036382), or enzymatic domains (e.g., PIRSF000876, PIRSF006638), or a combination of these types of domains (e.g., PIRSF003187).

For a biologist seeking to collect and analyze information about a protein, matching a protein sequence to a curated protein family provides a tool that is usually faster and more accurate than searching against a protein sequence database, which may only return a sequence and name submitted by a genomic sequencing project. Human curation of families provides richer information on protein structure and function, as it draws from a wider pool of information and from a classification-driven and rule-based system for automation of protein functional annotation that has been developed using PIRSF families [17,18].

The protein family classifications and associated information are stored in the PIRSF database and can be searched by a variety of methods (http://pir.georgetown.edu/pirsf). The PIRSF family reports (Fig. 2) (e.g., http://pir.georgetown.edu/cgi-bin/ipcSF?id=PIRSF000514) provide classification and annotation summaries organized in several sections—(i) *general information:* PIRSF number and general statistics (family size, taxonomy range, length range, keywords), as well as additional annotation for curated families, such as family name, bibliography,

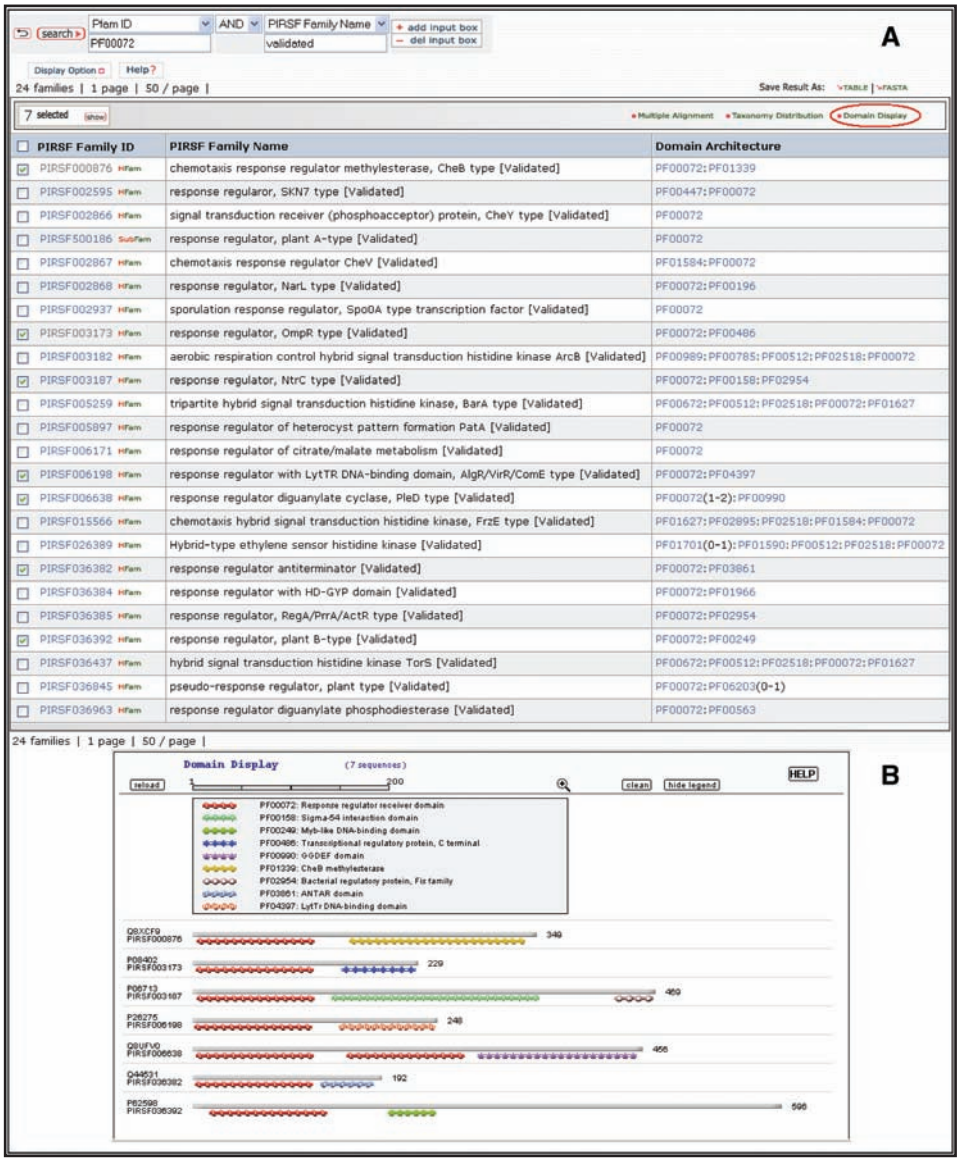


Fig. 1. (A) Selected PIRSF response regulator families all containing the CheY-like phosphoacceptor domain (Pfam domain PF00072) and (B) domain display of the selected PIRSF families.

family description, representative and seed members, and domain architecture; (ii) *membership:* lists of all members separated by major kingdoms and members from model organisms; and (iii) *function, structure, and family relationship:* enzyme classification (EC, http://www.chem.qmw.ac.uk/iubmb/enzyme/), structure hierarchy (SCOP [20]), gene ontology (GO [21]), as well as family relationships at the full-length protein, domain, and motif levels with direct mapping and links to other family, function, and structure classification schemes, such as Pfam and InterPro [10].

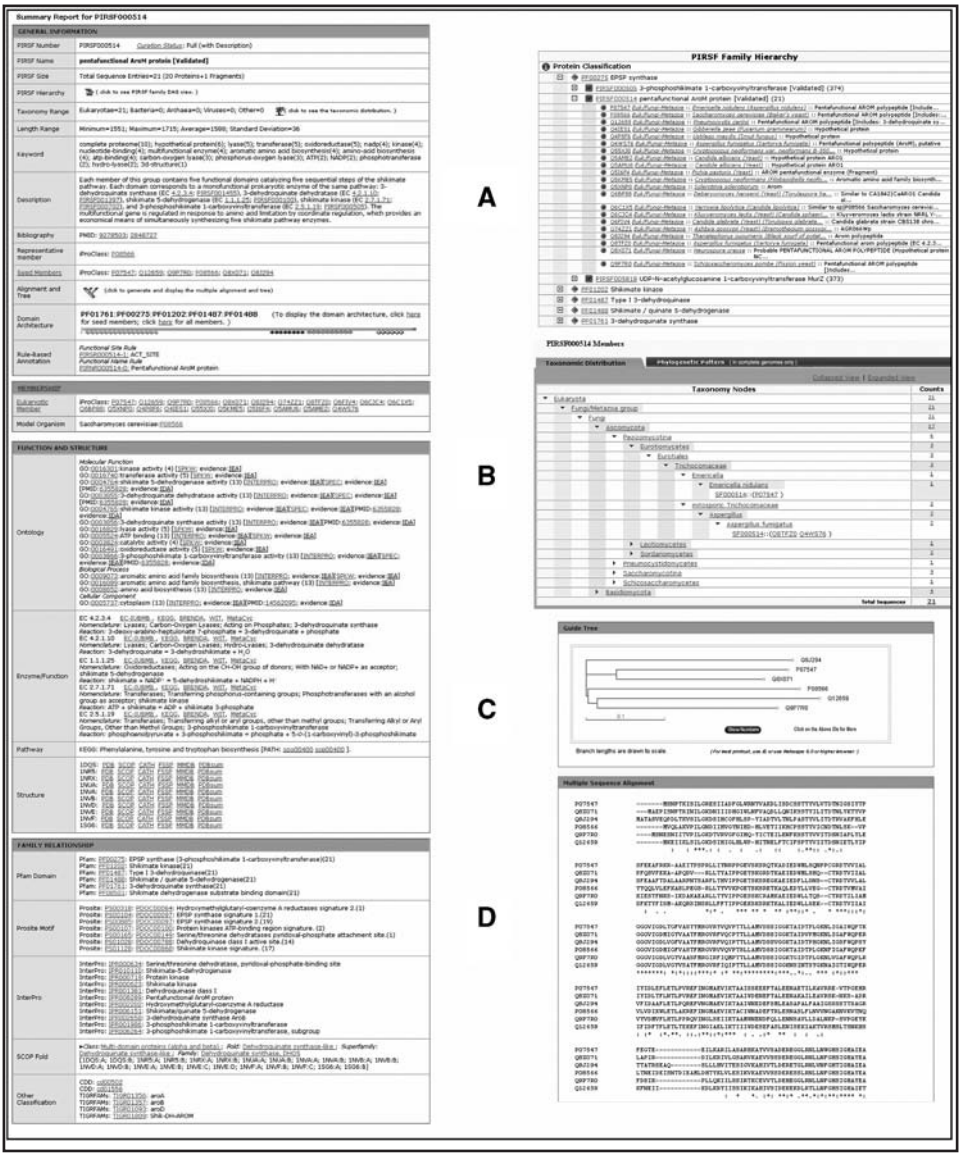


Fig. 2. PIRSF protein family report. Includes: (A) DAG browser displaying the PIRSF family hierarchy; (B) taxonomy tree browser displaying the taxonomy distribution of all family members; (C) tree viewer with neighbor-joining tree; and (D) alignment viewer displaying ClustalW multiple alignment of seed members. This report can be viewed directly at http://pir.georgetown.edu/cgi-bin/ipcSF?id=PIRSF000514.

The PIRSF reports connect to several graphical viewers, including: (i) DAG browser, which displays the PIRSF family hierarchy with Pfam domain superfamilies and protein membership in a network structure (Fig. 2A); (ii) taxonomy tree browser, which displays the taxonomy distribution of all family members and the

phylogenetic pattern of members in complete genomes (Fig. 2B); (iii) alignment and tree viewer, which displays ClustalW multiple alignment and neighbor-joining tree dynamically generated from seed members of curated families (Fig. 2C and D); and (iv) domain viewer, which displays domain architecture of seed members or all members.

3.3. iProClass integrated protein database

The iProClass database provides an integrated view of protein information [22] and serves as a bioinformatics framework for data integration and associative analysis of proteins [16]. iProClass presents value-added descriptions of all proteins in UniProtKB and contains comprehensive, up-to-date protein information derived from over 90 biological databases. Rich links to the underlying sources are provided with source attribution, hypertext links, and extracted summary information. The source databases include those for protein sequence, family, function, pathway, protein–protein interaction, complex, post-translational modification, protein expression, structure, structural classification, gene, genome, gene expression, disease, ontology, literature, and taxonomy. The iProClass protein summary report (Fig. 3) contains—(i) *general information:* protein ID and name (with synonyms, alternative names), source organism taxonomy (with NCBI taxonomy ID, group, and lineage), and sequence annotations such as gene names, keywords, function, and complex; (ii) *database cross-references:* bibliography (with PubMed ID and link to a bibliography information and submission page), gene and genome databases including RefSeq [23], Entrez gene [24], GO (with GO hierarchy and evidence tag), enzyme/function (with EC hierarchy, nomenclature, and reaction), pathway (with KEGG pathway name and link to pathway map), protein–protein interaction, structure (with PDB 3D structure image, matched residue range, and percent sequence identity for all structures matched at $\geq$30% identity), structural classes (with SCOP hierarchy for structures at $\geq$90% identity), sequence features, and post-translational modifications (with residues or residue ranges); (iii) *family classification:* PIRSF family, InterPro family, Pfam domain (with residue range), Prosite motif (with residue range), COG, and other classifications; and (iv) *sequence display:* graphical display of domains and motifs on the amino acid sequence.

The source attribution and hypertext links in iProClass facilitate exploration of additional information and examination of discrepancies in annotations from different sources. The data integration in iProClass allows identification of interesting relationships between protein sequence, structure, and function. It supports analyses of proteins in a "systems biology" context and has led to novel functional inference for uncharacterized proteins in the absence of sequence homology [25]. Furthermore, iProClass is used to support an ID mapping service that associates gene and protein IDs (such as NCBI's gi number and Entrez Gene ID) to UniProtKB identifiers. ID cross-referencing is fundamental to support data

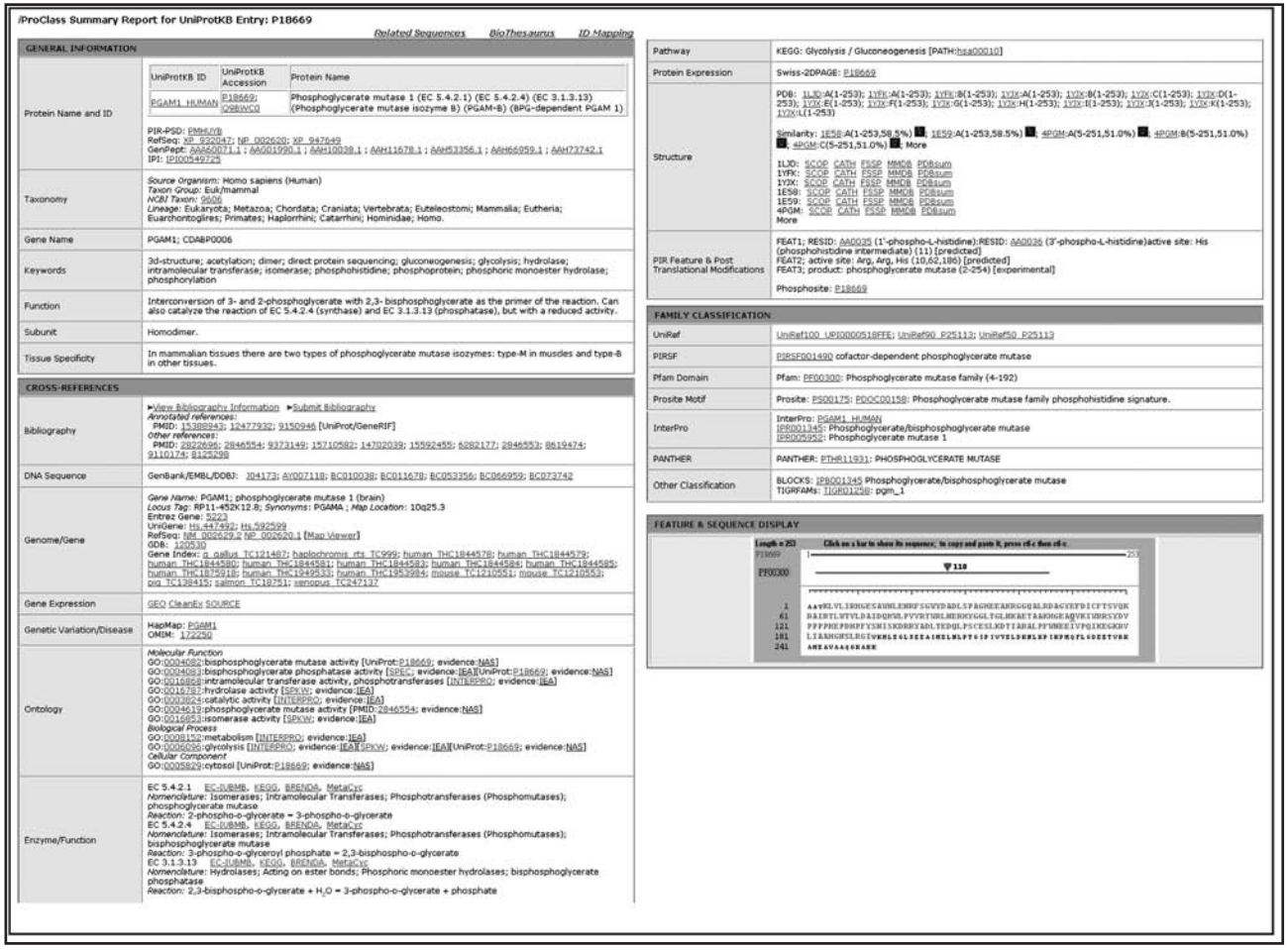


Fig. 3. iProClass protein sequence report. This report, for a human phosphoglycerate mutase, can be viewed directly at http://pir.georgetown.edu/cgi-bin/ipcEntry?id=P18669.

interoperability among disparate data sources and to allow integration and querying of data from heterogeneous molecular biology databases. The ID mapping service, accessible from http://pir.georgetown.edu/pirwww/search/idmapping.shtml, currently maps between UniProtKB identifiers and over 90 other database identifiers.

3.4. NIAID proteomic bioinformatics resource

The identification of proteins expressed in tissue, serum, cell lines, and other biological samples provides a mechanism for the discovery of novel biomarkers, particularly where contrasting samples can be derived, such as from healthy and diseased tissues or cells. Even when the biological mechanism of disease is poorly understood, proteomics studies can provide insight into the proteins and their isoforms that show evidence of heightened or suppressed abundance in one context or the other. With the advent of high-throughput proteomics technologies, ever-increasing amounts of proteomic data are being generated. The challenge is to link relevant experimental data to other information on the proteins.

The National Institute of Allergy and Infectious Diseases (NIAID) Biodefense Proteomic Research Program has funded seven centers to work on NIAID Category A–C priority pathogens and other microorganisms responsible for emerging and/or re-emerging diseases. In addition, they have funded a Resource Center for Biodefense Proteomics Research (http://www.proteomicsresource.org/) of which PIR is a member. The Administrative Resource is charged with making the data, methods, and conclusions from Proteomic Research Centers available to the scientific community.

For the NIAID project PIR has developed several data integration tools. (1) The Master Protein Directory is a complete compilation of proteins and reagents identified by the NIAID Biodefense Proteomics Research Centers. The directory links protein sequence and functional annotation to experimental data generated by the project and eventually to metabolic information. (2) Complete Predicted Proteomes Tool (Fig. 4) that allows users to view and search selected proteomes being studied by the NIAID Biodefense Proteomics Research Centers. Over 50 fields are searchable, a customizable display of functional annotation is provided, and proteins are linked to the Master Protein Directory of experimental data. (3) Core/Unique Protein Identification (CUPID) system [26] provides a list of proteins encoded by selected organisms that are unique to the query strain, species, or genus. Such proteins may serve as potential drug targets or diagnostics for pathogenic organisms. The unique protein "signatures" may be specific to the strain of interest (narrow-range targets) or may be part of the "core set" of proteins encoded by strains within the same species or genus of interest (broad-range targets). These tools are available at http://pir.georgetown.edu/proteomics/.

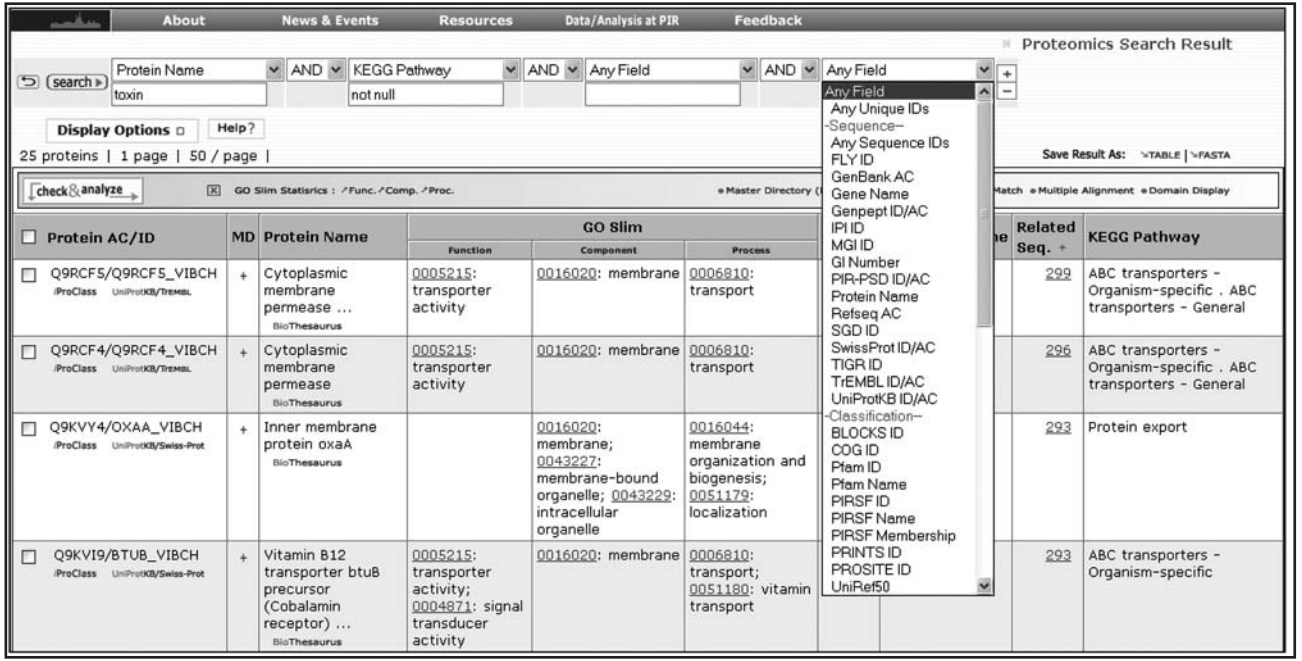


Fig. 4. Complete predicted Proteomes Tool for the NIAID Biodefense Proteomics Research Program (http://pir.georgetown.edu/pirwww/proteomics/). Tool allows interactive text mining of selected complete proteomes. Features include: over 50 fields for Boolean text searches; customizable display and export; links to master catalog of experimental data from NIAID Proteomics Research Centers; and links to various reports on additional protein information like UniProt, iProClass, BioThesaurus, and PIRSF reports.

4. Discussion

To fully utilize the increasing flood of biological data requires the creation of integrated systems for knowledge discovery and scientific exploration, including the integration of (i) disparate data sources and scientific literature and (ii) data mining and analysis tools. Here we presented our ongoing efforts at creating an integrated knowledgebase system for proteomic information. Other major groups have tried different approaches for integrating genetic information (i.e., NCBI and European Bioinformatics Institute, EBI) and cancer information (National Cancer Institute, NCI). However, the current system for funding academic bioinformatics research does not provide many options to fund broad infrastructure and integration efforts and is currently more focused on developing new tools and algorithms to address particular problems. Thus, further progress on integration of a wider array of literature, data, and analysis tools requires community-wide efforts to develop and utilize common standards for data exchange, software architecture, and interoperability.

A number of such community efforts are underway in data exchange standards for both genomics and proteomics, including the MGED Society (http://www.mged.org/), which aims to facilitate the sharing of microarray data generated by functional genomics experiments, and the HUPO PSI (http://psidev.info/), which is trying to define community standards for data representation in proteomics. These groups are developing XML data exchange standards, minimum reporting requirements, object models, and ontologies for their area of interest. Minimum reporting requirements (i.e., Minimum Information about a Proteomics Experiment, MIAPE) are an attempt to define the minimum information required to publish results on a genomic or proteomic study. Object models are a practice derived from software engineering that attempts to abstract the data objects and sometimes even analysis steps of a system independent of any implementation. Common object models can thus facilitate the development of compatible search and analysis tools regardless of platform, simplifying both the dissemination and the exchange of data. An ontology is an explicit specification of the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them. If two systems (i.e., databases) share a common ontology it means they share a common vocabulary that can be used in a consistent manner. This allows intelligent automation of information gathering and knowledge sharing via software agents. The National Center for Biomedical Ontology is one resource for tools and information on ontologies [27] (http:// bioontology.org).

The NCI-funded caBIG is a community effort of cancer centers in the United States to develop a web of interoperable data sources and tools that can seamlessly share and analyze information from a wide variety of sources including clinical cancer studies and molecular research laboratories (https://cabig.nci.nih.gov/). The evolving architecture for this system is dependent on developing common standards and practices, including object modeling, data exchange standards, and

common ontologies and vocabularies. PIR is an active participant in caBIG and is one of the service nodes available on caGrid 1.0, the first public version of the grid. The PIR grid service makes search and retrieval services available for information in the UniProtKB protein database.

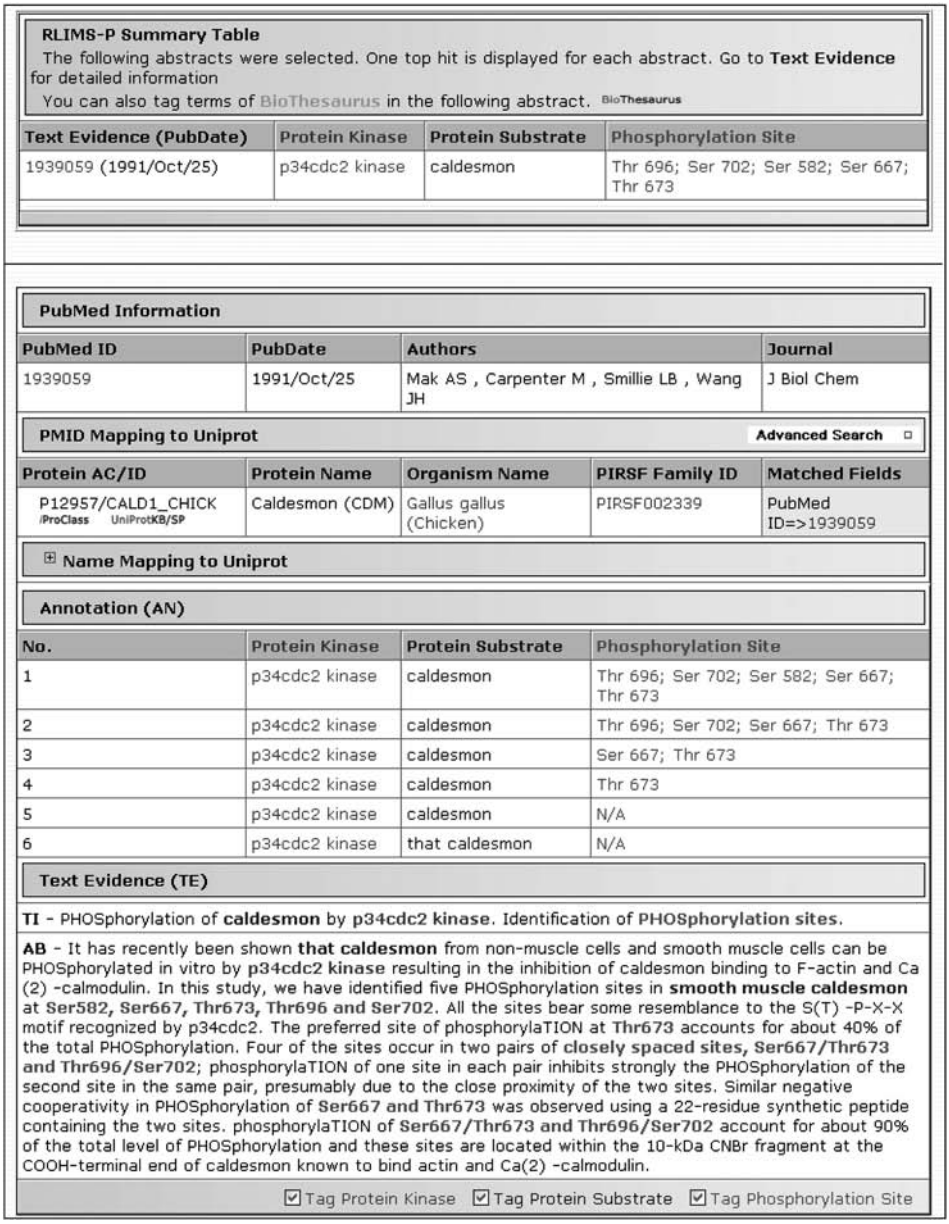
5. Future trends

Like many human endeavors, the future trend in bioinformatics data analysis and integration is to replace routine human intervention as much as possible. Currently even the most up-to-date databases and tools lag behind the actual research results by months and sometimes years if human reading and processing of the scientific literature is required. More and more automated rule-based systems will take the lead in data analysis and integration, linking existing protein knowledge to new experimental data almost as soon as they are published or even prior to publication. Several such efforts are under development at PIR including iProLink which has tools and resources for automated literature mining and iProXpress where data such as those produced by the NIAID Proteomics Centers and others can feed into automated analysis and annotation pipeline.

5.1. iProLINK literature mining resource

A large volume of protein experimental data is buried within the fast-growing scientific literature. While of great value, such information is limited in databases due to the laborious process of literature-based curation. A resource for protein literature mining, iProLINK provides curated data sources and tools to support text mining in the areas of bibliography mapping, annotation extraction, protein named-entity recognition, and protein ontology development [19]. The data sources and tools include mapped citations (mapping of annotated bibliography with PubMed IDs to protein entries), name- or annotation-tagged literature corpora (papers tagged with protein names [28] or with experimentally validated post-translational modifications), the RLIMS-P rule-based literature mining system for protein phosphorylation [29], and the BioThesaurus of protein and gene names [30]. iProLINK is freely accessible at http://pir.georgetown.edu/iprolink/, and serves as a knowledge link bridging protein databases and literature databases such as PubMed.

The RLIMS-P [29] is a text-mining program that can be used to identify papers describing protein phosphorylation from all PubMed abstracts, and to extract from these abstracts the specific information on protein phosphorylation, namely the kinases, the protein substrates, and the amino acid residues/positions being phosphorylated (Fig. 5). The system achieved an overall recall of 96% for paper retrieval and a precision of 98% for extraction of substrates and phosphorylation sites. The RLIMS-P Web site [31] provides online retrieval of phosphorylation papers using PubMed ID, followed by extraction of phosphorylation information



RLIMS-P Summary Table

The following abstracts were selected. One top hit is displayed for each abstract. Go to **Text Evidence** for detailed information

You can also tag terms of BioThesaurus in the following abstract. BioThesaurus

Text Evidence (PubDate)	Protein Kinase	Protein Substrate	Phosphorylation Site
1939059 (1991/Oct/25)	p34cdc2 kinase	caldesmon	Thr 696; Ser 702; Ser 582; Ser 667; Thr 673

PubMed Information

PubMed ID	PubDate	Authors	Journal
1939059	1991/Oct/25	Mak AS , Carpenter M , Smillie LB , Wang JH	J Biol Chem

PMID Mapping to Uniprot Advanced Search

Protein AC/ID	Protein Name	Organism Name	PIRSF Family ID	Matched Fields
P12957/CALD1_CHICK /ProClass UniProtKB/SP	Caldesmon (CDM)	Gallus gallus (Chicken)	PIRSF002339	PubMed ID=>1939059

Name Mapping to Uniprot

Annotation (AN)

No.	Protein Kinase	Protein Substrate	Phosphorylation Site
1	p34cdc2 kinase	caldesmon	Thr 696; Ser 702; Ser 582; Ser 667; Thr 673
2	p34cdc2 kinase	caldesmon	Thr 696; Ser 702; Ser 667; Thr 673
3	p34cdc2 kinase	caldesmon	Ser 667; Thr 673
4	p34cdc2 kinase	caldesmon	Thr 673
5	p34cdc2 kinase	caldesmon	N/A
6	p34cdc2 kinase	that caldesmon	N/A

Text Evidence (TE)

TI - PHOSphorylation of **caldesmon** by **p34cdc2 kinase**. Identification of **PHOSphorylation sites.**

AB - It has recently been shown **that caldesmon** from non-muscle cells and smooth muscle cells can be PHOSphorylated in vitro by **p34cdc2 kinase** resulting in the inhibition of caldesmon binding to F-actin and Ca (2) -calmodulin. In this study, we have identified five PHOSphorylation sites in **smooth muscle caldesmon** at **Ser582, Ser667, Thr673, Thr696 and Ser702**. All the sites bear some resemblance to the S(T) -P-X-X motif recognized by p34cdc2. The preferred site of phosphorylaTION at **Thr673** accounts for about 40% of the total PHOSphorylation. Four of the sites occur in two pairs of **closely spaced sites, Ser667/Thr673 and Thr696/Ser702**; phosphorylaTION of one site in each pair inhibits strongly the PHOSphorylation of the second site in the same pair, presumably due to the close proximity of the two sites. Similar negative cooperativity in PHOSphorylation of **Ser667 and Thr673** was observed using a 22-residue synthetic peptide containing the two sites. phosphorylaTION of **Ser667/Thr673 and Thr696/Ser702** account for about 90% of the total level of PHOSphorylation and these sites are located within the 10-kDa CNBr fragment at the COOH-terminal end of caldesmon known to bind actin and Ca(2) -calmodulin.

☑ Tag Protein Kinase ☑ Tag Protein Substrate ☑ Tag Phosphorylation Site

Fig. 5. RLIMS-P text-mining results show summary of protein kinase, protein substrate, and phosphorylation position information extracted from a Medline abstract. Details of information extracted and words tagged as phosphorylation objects in the abstracts text are shown below the summary.

from the Medline abstracts and tagging of the three phosphorylation objects (kinases, substrates, and sites). The Web site also allows mapping of phosphorylated proteins to UniProtKB protein entries based on PubMed ID and/or protein name.

The BioThesaurus maps a comprehensive collection of protein and gene names to all known proteins in UniProtKB [30]. Currently covering more than 3 million proteins, BioThesaurus consists of over 4 million names extracted from multiple molecular biological databases according to the database cross-references in iProClass. The BioThesaurus Web site allows the retrieval of all the various names used for a single protein and the identification of all proteins sharing the same name. The synonymous names in BioThesaurus can be used for query expansion during literature search to retrieve relevant papers and extract protein information even when non-standardized names are used.

5.2. iProXpress knowledge system for gene expression and proteomic data analysis

Identification of expressed proteins in biological samples allows the discovery of novel disease biomarkers even when the underlying biological mechanism is poorly understood. Once proteins are identified and their expression profiles defined, the protein groups can be analyzed for their functional involvement in metabolic and signaling pathways, cell cycles, apoptosis, and other cellular functions and processes. Such biological interpretation requires the data to be related to other types of information at the protein function, pathway, and network level. While numerous resources are available for processing data generated from transcriptome and proteome-wide experiments, expression data analysis is often carried out in an *ad hoc* manner, with a fragmented and inefficient use of information resources.

The iProXpress knowledge system consists of (i) a data warehouse with integrated protein information, (ii) analytical tools for protein sequence analysis and functional annotation, and (iii) a graphical user interface for categorization and visualization of expression data. The design of the iProXpress knowledge system (Fig. 6) is outlined below.

5.2.1. Gene/peptide to protein mapping

Gene or protein probes are mapped to the corresponding entries in UniProtKB of all known proteins, based on gene/protein IDs, names, or sequences. Genes are mapped using iProClass cross-references that connect gene identifiers such as GenBank or Entrez Gene IDs to UniProtKB identifiers. If a common gene or protein ID is not used for the probe set, the mapping is based on direct sequence comparison or on name matching if sequence is not available. Peptide data are mapped by matching peptide sequences to protein sequences with subsequent assembly. The name matching is assisted by BioThesaurus; however, often there are ambiguous identifications due to the lack of gene and protein name standards.

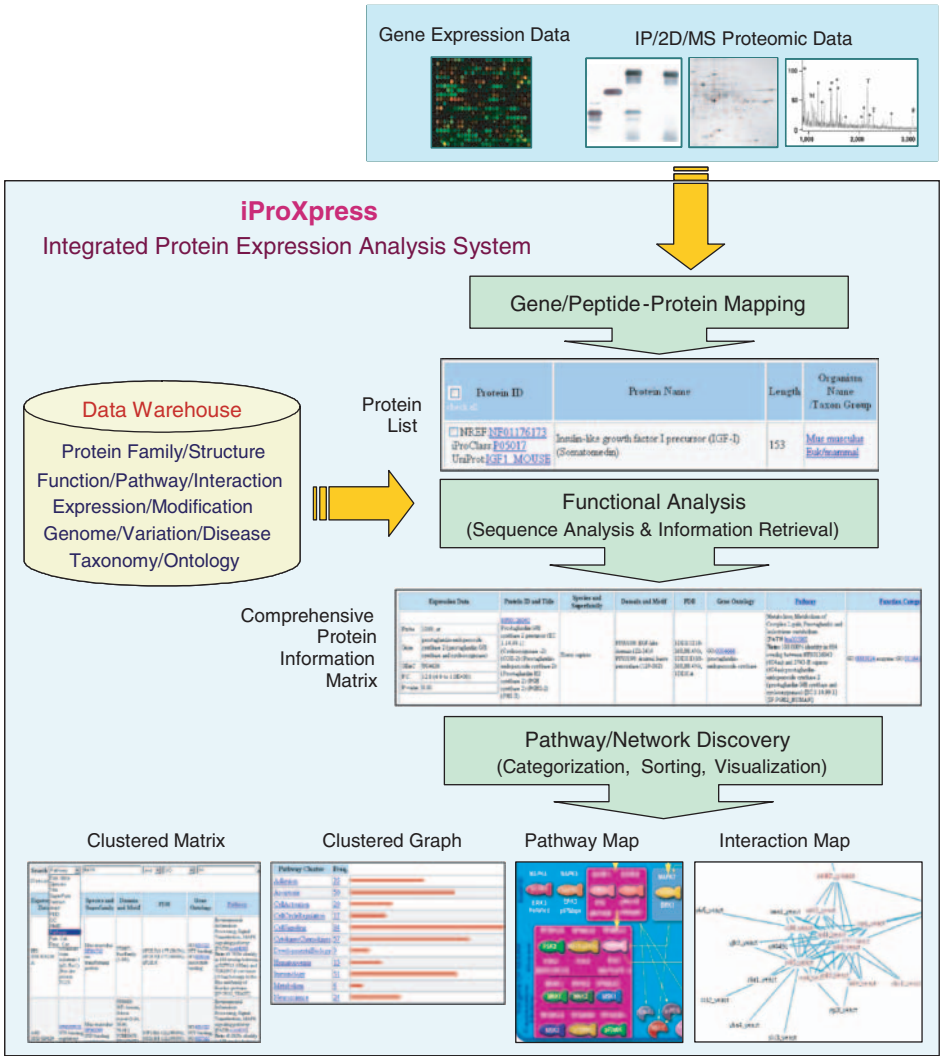


Fig. 6. iProXpress integrated knowledge system for gene expression and proteomic data analysis.

5.2.2. Functional analysis

The UniProt IDs assigned to corresponding genes and proteins link all mRNA and protein expression data. Protein family, domain, and functional site features for each protein are identified by BLAST, HMM, signal peptide, transmembrane helix predictions, and other automated searches. For direct human comparison of expressed genes/proteins, a comprehensive protein information matrix is generated, summarizing salient features retrieved from the underlying PIR data warehouse or inferred based on sequence similarity. Attributes in the protein matrix include:

protein name, family, domain, motif, site, post-translational modification, isoform, GO, function/functional category, structure/structural classification, pathway/pathway category, protein interaction, and complex.

5.2.3. Pathway and network discovery

Users can conduct iterative categorization and sorting of proteins in the information matrix and correlate expression and interaction patterns to salient protein properties for pathway and network discovery. Proteins are clustered based on functions, pathways, and/or other attributes in the information matrix to identify hidden relationships not apparent in the data on expression patterns and protein interactions, and to recognize candidate proteins of unknown identity that warrant further investigation. We detect new or different clusters based on combined attributes of the information matrix and the expression and/or interaction data. Unknown "hypothetical" proteins involved in critical pathways or networks can be manually curated based on phylogenetic analysis, structure homology modeling, genome context, and functional associations using an integrative approach that has led to novel functional inference for uncharacterized proteins [25]. This bioinformatics analysis thus provides a composite, global view of functional changes to identify critical nodes and hidden relationships in the response pathways and networks. These last iterative categorization steps in the process are currently done manually; however, many of them can be automated and rules developed to flag significant clusters.

A pilot iProXpress system has been applied to gene expression profile analysis for human chorionic gonadotropin (hCG)-induced changes in MA-10 mouse Leydig tumor cells [32]. The system has further been utilized to analyze proteomes of various stages of melanosomes from human melanoma cell lines [33] and for the comparative analysis of seven lysosome-related organelles (LROs) [34]. The organellar proteome analyses allow us to identify possible melanosome stage-specific proteins and organelle-specific proteins as well as proteins shared among different organelles, thereby facilitating a better understanding of melanosome biogenesis pathways and the dynamic process of LRO biogenesis.

6. Conclusions

Research in the areas of protein science and bioinformatics for over the last three decades has provided a solid foundation to automatically analyze and classify protein information. Here we presented some of the efforts and resources of PIR to annotate, classify, and integrate protein sequence information with other biomedical information. The ongoing challenges in data integration are as follows. (1) Scaling-up processes to deal with the ever-increasing high-throughput genomic/

proteomic analysis: PIR's development of the iProXpress system described above is one attempt to do this but there will be many others. Modern mass spectrometry methods can rapidly generate more data than humans or available software can quickly analyze for meaningful information. (2) The integration and interoperation of numerous related biological and medical information: The future here seems dependent on community efforts to define common data standards, object models, and ontologies to allow automated queries and analysis of disparate data sources. PIR is actively participating in many of these efforts including NCI's caBIG pilot project whose success or failure will have a large influence on how such efforts are conducted in the future. (3) Developing better ways to quickly and accurately present this information in a form that humans can easily use to develop new biomedical knowledge: This is an area that needs more attention. Bioinformatics professionals need to work more closely with potential users and human factors specialists to develop improved visualization techniques. To quote one colleague at a large mass spectrometry laboratory, "We routinely produce data from a single experiment whose summary of protein identifications contains more lines than Excel can handle and our biologist customers know what to do with."

Acknowledgments

The PIR protein bioinformatics activities are supported by the NIH grants for UniProt (U01 HG02712-01), NIAID proteomic resource (HHSN266200400061C), and grid enablement (NCI-caBIG-ICR-10-10-01), and National Science Foundation grants for iProClass (DBI-0138188) and BioThesaurus (ITR-0205470).

References

1. MacCoss, M.J., Computational analysis of shotgun proteomics data. *Curr. Opin. Chem. Biol.*, **9**, 88–94 (2005).
2. Johnson, R.S., Davis, M.T., Taylor, J.A. and Patterson, S.D., Informatics for protein identification by mass spectrometry. *Methods*, **35**, 223–236 (2005).
3. Galperin, M.Y., The molecular biology database collection: 2006 update. *Nucleic Acids Res.*, **34**, D3–D5 (2005).
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J. et al., GenBank. *Nucleic Acids Res.*, **34**, D16–D20 (2005).
5. Ball, C.A., Awad, I.A., Demeter, J. et al., The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.*, **33**, D580–D582 (2005).
6. Blake, J.A., Eppig, J.T., Bult, C.J. et al., The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res.*, **34**, D562–D567 (2006).
7. Sherry, S.T., Ward, M.H., Kholodov, M. et al., dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311 (2001).

8. Cole, J.R., Chai, B., Farris, R.J., Wang, Q. et al., The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.*, **33**, D294–D296 (2005).
9. Wu, C.H., Apweiler, R., Bairoch, A. et al., The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191 (2006).
10. Mulder, N.J., Apweiler, R., Attwood, T.K. et al., InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205 (2005).
11. Kouranov, A., Xie, L., de la Cruz, J. et al., The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302–D305 (2006).
12. Alfarano, C., Andrade, C.E., Anthony, K. et al., The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424 (2005).
13. Kanehisa, M., Goto, S., Hattori, M. et al., From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357 (2006).
14. Wheeler, D.L., Barrett, T., Benson, D.A. et al., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **34**, D173–D180 (2006).
15. Wu, C.H., Yeh, L.-S., Huang, H. et al., The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347 (2003).
16. Wu, C.H., Huang, H., Nikolskaya, A., Hu, Z. and Barker, W.C., The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem.*, **28**, 87–96 (2004).
17. Wu, C.H., Nikolskaya, A., Huang, H. et al., PIRSF family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**, D112–D114 (2004).
18. Natale, D.A., Vinayaka, C.R. and Wu, C.H., Large-scale, classification-driven, rule-based functional annotation of proteins. In: Subramaniam, S. (Ed.), *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Proteomics Volume*. 2005.
19. Hu, Z.Z., Mani, I., Hermoso, V., Liu, H. and Wu, C.H., iProLINK: an integrated protein resource for literature mining. *Comput. Biol. Chem.*, **28**, 409–416 (2004).
20. Andreeva, A., Howorth, D., Brenner, S.E. et al., SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229 (2004).
21. Gene Ontology Consortium, The gene ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326 (2006).
22. Huang, H., Barker, W.C., Chen, Y. and Wu, C.H., iProClass: an integrated database of protein family, function, and structure information. *Nucleic Acids Res.*, **31**, 390–392 (2003).
23. Pruitt, K.D., Tatusova, T. and Maglott, D.R., NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504 (2005).
24. Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T., Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58 (2005).
25. Nikolskaya, A.N., Arighi, C., Huang, H., Barker, W.C. and Wu, C.H., PIRSF family classification system for protein functional and evolutionary analysis. *Evol. Bioinformatics Online*, **2**, 209–221 (2006).
26. Mazumder, R., Natale, D., Murthy, S., Thiagarajan, R. and Wu, C.H., CUPID: computational identification of strain-, species- and genus-specific proteins. *BMC Bioinformatics*, **6**, 279 (2005).
27. Rubin, D.L., Lewis, S.E., Mungall, C.J. et al., National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS*, **10**, 185–198 (2006).
28. Mani, I., Hu, Z.Z., Jang, S.E. et al., Protein name tagging guidelines: lessons learned. *Comp. Funct. Genomics*, **6**, 72–76 (2005).
29. Hu, Z.Z., Narayanaswamy, M., Ravikumar, K.E., Vijay-Shanker, K. and Wu, C.H., Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **21**, 2759–2765 (2005).

30. Liu, H.F., Hu, Z.Z., Zhang, J. and Wu, C.H., BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103–105 (2006).
31. Yuan, X., Hu, Z.Z., Wu, H.T. et al., An online literature mining tool for protein phosphorylation. *Bioinformatics*, **22**, 1668–1669 (2006).
32. Li, W., Amri, H., Huang, H., Wu, C.H. and Papadopoulos, V., Gene and protein profiling of the acute response of MA-10 Leydig tumor cells to human chorionic gonadotropin. *J. Androl.*, **25**, 900–913 (2004).
33. Chi, A., Valencia, J.C., Hu, Z.Z. et al., Proteomic and bioinformatic characterization of the biogenesis and function of melanosomes. *J. Proteome Res.* (2006), in press.
34. Hu, Z.Z., Valencia, J.C., Huang, H. et al., Comparative bioinformatics analyses and profiling of lysosome-related organelle proteomes. *Int. J. Mass Spectrom.* (2006), in press.

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 11

Analysis of complex lipidomes

ANDREAS UPHOFF, MARTIN HERMANSSON, PERTTU HAIMI, and PENTTI SOMERHARJU*

Institute of Biomedicine, Department of Biochemistry, University of Helsinki, 00014 Helsinki, Finland

*Corresponding author. E-mail: somerhar@mappi.helsinki.fi (P. Somerharju).

1. Introduction

It is now firmly established that lipids, besides forming the "backbone" of all biological membranes, are also key players in a variety of other physiological phenomena including signal transduction, energy metabolism, intracellular sorting of membrane-bound molecules, morphogenesis, etc. A typical mammalian cell has been estimated to contain more than 1000 different lipid species. The meaning of such a great variety of lipid molecules is not well understood, even if one considers the numerous functions of lipids listed above. "Functional lipidomics" is a novel field of research probing the relationship between a (detailed) lipid composition of a cell or organism (the "lipidome") and a particular biological or medical problem.

Traditionally, lipids are analyzed by means of thin-layer chromatography (TLC), high-performance liquid chromatography (HPLC), and gas chromatography (GC) [1,2]. While still useful for many purposes, these methods are labor-intensive, time-consuming, and insensitive and thus detailed analyses of complex lipidomes have become feasible only recently due to the rapid development of novel mass-spectrometric methods, particularly electrospray ionization mass spectrometry (ESI–MS). Although even ESI–MS does not yet allow one to quantify all >1000 lipid molecules that form the "lipidome" of a cell or tissue, method development is progressing so fast that this will probably be feasible within a few years. Because of the great body of data produced even in a single ESI–MS analysis, computerized data analysis is a necessity. It is also essential to develop novel bioinformatics tools to correlate lipidomes (and changes therein) with the functions of the system. MS-lipidomics has a great, albeit yet largely unrealized, potential in diagnosing diseases or pathological conditions including atherosclerosis, metabolic syndrome, and other disorders.

Recent developments in mass spectrometry, particularly the introduction of the ESI method, have paved way to "functional lipidomics," i.e., the use of detailed lipid profile of cell or tissues to unravel biological phenomena and the mechanisms underlying various metabolic perturbations or diseases.

At present, ESI–MS allows quantitative analysis of hundreds of phospholipid species present in a sample. Two different approaches are commonly used. The

MS/MS method relies on selective detection of lipid classes directly from the crude lipid extract by precursor ion or neutral loss (NL) scanning. For example, phosphatidylethanolamines (PE) can be selectively detected by scanning for the constant NL of 141 Da, while phosphatidylinositols (PI) can be analyzed scanning for the precursor of *m*/*z* 241. On the other hand, triacylglycerols (TAGs) have been analyzed by constant NL scanning for different fatty acids and many sphingolipids by scanning for the precursors of the dehydrated sphingoid base. Notably, the MS/MS approach allows very convenient and detailed studies of lipid metabolism by using heavy isotope (^{2}H or ^{13}C)-labeled precursors. For instance, labeling of cells with D_9-labeled choline combined with scanning for the precursors of *m*/*z* 193 (deuterated choline) allows one to selectively detect only the labeled species without interference by the unlabeled ones.

The other, LC–MS, approach makes use of on-line preseparation of the lipids before MS analysis. This method allows analysis of species for which no specific scan modes exist. Also, many isobaric species, not readily resolved by MS/MS, can be analyzed. Finally, this method, particularly when employing multiple reaction monitoring (MRM), provides the highest sensitivity of detection of many minor lipid classes as the suppression effects are minimized.

Independent of the approach, a particular problem in quantitative analysis of lipid compositions with MS is the lack of standards. For accurate results, it is obligatory to include one or preferably several internal standards for each lipid class to be analyzed. This is because the instrument response can vary markedly depending on structural details such as the length and unsaturation of the alkyl chains. The mechanisms behind such structure-dependent variations in instrument response are not fully understood, but differences in ionization and fragmentation efficiencies are probably involved.

Triple quadrupole MS instruments have been the most common ones in studies involving lipid analysis, but novel hybrid (quadrupole time-of-flight, etc.) instruments are rapidly gaining popularity due to their ability for multiple precursor ion scans simultaneously. Besides ESI, atmospheric pressure chemical ionization (APCI), atmospheric pressure photoionization (APPI), and matrix-assisted laser desorption ionization (MALDI) have been employed in analysis of lipids. However, these methods seem to have an advantage over ESI only in special cases. For instance, APPI and APCI allow analysis of sterols without derivatization, which is needed for ESI.

Due to its remarkable resolving power and speed of analysis, the amount of data produced by MS analysis of lipidomes is often overwhelming, even when a limited number of samples are to be analyzed. Therefore, computerized methods are necessary and have indeed been published recently for both MS/MS and LC–MS data. Beyond the analysis of primary data, there is an urgent need for programs allowing one to correlate lipid compositions with other compositional and functional data. This is because the lipid profiles as such are often difficult to interpret in terms of

functions and mechanisms. To this end, lipid databases are needed and are presently under construction.

Due to its novelty, MS-based lipidomics is still evolving and relative few practical applications have emerged so far. However, it is highly likely that in the near future MS-lipidomics will play, in combination with other "omics," a crucial role in biology, biotechnology, and medicine.

In this review, we will first describe the general methodology of MS-lipidomics, then present the state-of-the-art of analysis of different phospholipids, neutral lipids, e.g., TAGs, cholesterol esters (CEs), sphingolipids, and sterols, and, finally, we will discuss the potential clinical applications of MS-lipidomics. Unfortunately, due to lack of space we cannot deal with all relevant publications. These can be found in recent reviews on MS-lipidomics [3–10].

2. Methodology

2.1. Lipid extraction

The analysis of lipids by MS (or any another method) usually requires that they are first separated from other molecules and ions present in the sample under study. By far the most common separation method is extraction of the lipids with organic solvents followed by "washing" of the organic phase with a polar one in order to remove the contaminants potentially interfering with the analysis [11,12]. However, some lipids are relatively polar and thus partially lost during such washing procedure. This is particularly true for complex gangliosides, free sphingoid bases, and sphingosine-1-phosphate and several lysophospholipids. For these lipid-modified liquid/liquid partition methods need to be employed [9,13,14]. Alternatively, solid-phase extraction could be exploited [15]. Independent of the method used, it is useful to include internal standards before the extraction to correct for any losses upon extraction.

2.2. Mass spectrometry

Recent developments in soft ionization techniques, such as ESI, APCI, and MALDI as well as instrumentation, have "revolutionized" the analysis of lipidomes due to their simplicity, sensitivity, and resolving power. Several comprehensive reviews of the present state-of-the-art are available [3–10]. Most published MS analyses of lipidomes have utilized ESI and triple (or hybrid) quadrupole instruments and, therefore, we will focus on this setup below. However, references to other types of instrumentation will be made when they seem to offer special advantages.

Since lipid extracts often contain hundreds of different species, many of which are structurally very similar (e.g., differ only by one double bond), even standard MS analysis is not capable of resolving them all, but special strategies need to be applied

to enhance the selectivity of detection and discriminate against chemical noise. One approach is to vary the polarity of the ion source. For instance, phosphatidylcholines (PCs) and sphingomyelins (SMs) provide higher signal in the positive vs. negative mode, while the opposite is true for acidic lipids like phosphatidylinositols and–serines. Additional selectivity can be achieved by a judicious choice of the solvent, pH, and added salts [4]. Nevertheless, such manipulations are not adequate to reliably quantify all lipid species present in a typical sample, but additional measures are necessary to enhance the specificity of detection. To this end, two alternative approaches have been adopted. One of them involves (partial) on-line chromatographic separation of the lipids before the MS analysis [16–20]. This LC–MS method (Fig. 1, upper panel) often allows quantification of isobaric (of equal mass) species due to their differential retention in the column. The other significant benefit is that suppression effects are alleviated, thus allowing more sensitive detection of low-abundance species [3]. LC–MS with MRM is probably the most sensitive method of lipid analysis, particularly when using capillary columns.

The second commonly used approach makes use of precursor ion (PI) or NL scans to selectively detect specific lipids in crude lipid extracts directly (i.e., without preseparation) infused to the mass spectrometer [21–23]. A common phenomenon among phospholipids is the loss of the head group as a charged or neutral fragment upon collisionally activated dissociation (CAD). Due to differences in the chemical structure of the lipid head group, CAD often gives rise to a fragment(s) which is characteristic to a phospholipid class (Fig. 2), and thus the members of this class can be selectively detected by PI or constant NL scanning. This approach (Fig. 1, lower panel) can also be used to selectively detect glycerolipids containing specific fatty acids [4]. Analogously, many sphingolipids can be selectively detected by scanning for the precursors of dehydrated sphingoid bases.

The choice between these (or other) strategies depends on, for example, the accessible instrumentation, complexity of the lipidome, and the amount of sample available. In general, LC–MS provides higher resolving power and sensitivity, while a notable advantage of the direct infusion MS/MS method is that isotope-labeled lipids can be detected selectively (without interference by the unlabeled ones) simply by changing a single scan parameter [24–26].

Full analysis of lipidomes requires that the structures and positions of fatty acyl moieties present in individual lipid species can be resolved. In case of glycerophospholipids the acyl moieties can be identified based on the formation of the corresponding fragments upon CAD in the negative ion mode [21]. Even the *sn*-1 and *sn*-2 positions (*cf.* Fig. 2) of the acyl moieties can be identified based on the relative intensities of the lysolipid fragments [19,27–30].

Recently, interest in hybrid instruments, such as Q-TOFs, has grown as these enable simultaneous recording of almost unlimited number of fragmentations [31–33]. This is particularly useful for elucidating the acyl residues present, since multiple precursor ion scans can be performed simultaneously [34].

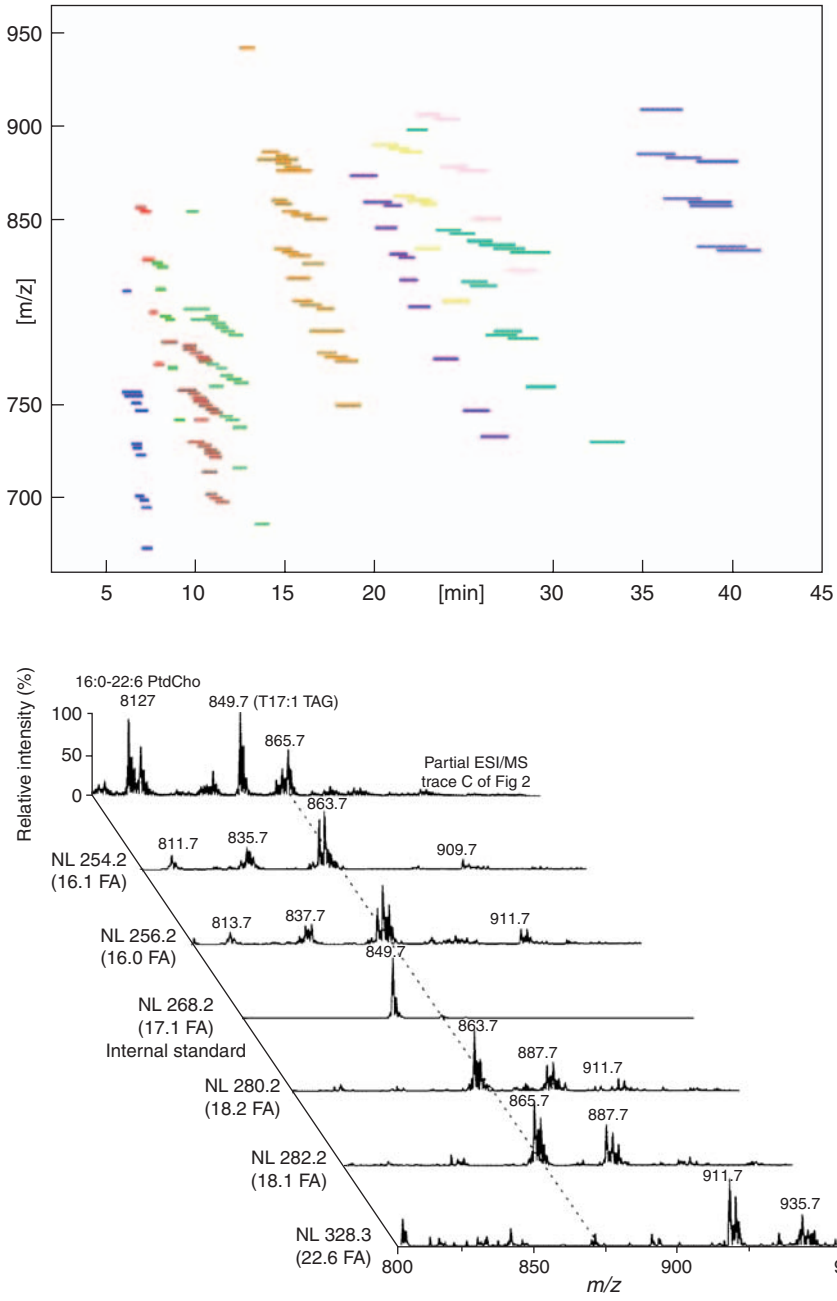


Fig. 1. *Upper panel:* 2D display of mouse brain polar lipids as analyzed by LC–MS. Color-coded data are shown for clarity; original can be found in ref. [19]. The lipid classes from left to right are as follows: blue at ~7 min, PA; black, GalCer; gray, α-OH-GalCer; red, plasmalogen-PE; green at ~12 min, diacyl-PE; orange, PC; violet, SM; yellow, sulfatides; pink, α-hydroxysulfatides; green at ~25 min, PS; and blue at ~38 min, PI (reprinted in part with permission from Hermasson et al., *Anal Chem.*, **77**, 2166–2175 (2005)). *Lower panel:* 2D MS/MS spectral analysis of triacylglycerol molecular species from mouse liver. Multiple neutral loss scans specific for different fatty acid moieties were combined to identify and quantify the triacylglycerol molecular species present (reprinted from Han et al., *Anal. Biochem.*, **330**, 317–331 (2004)) (reprinted in part with permission).

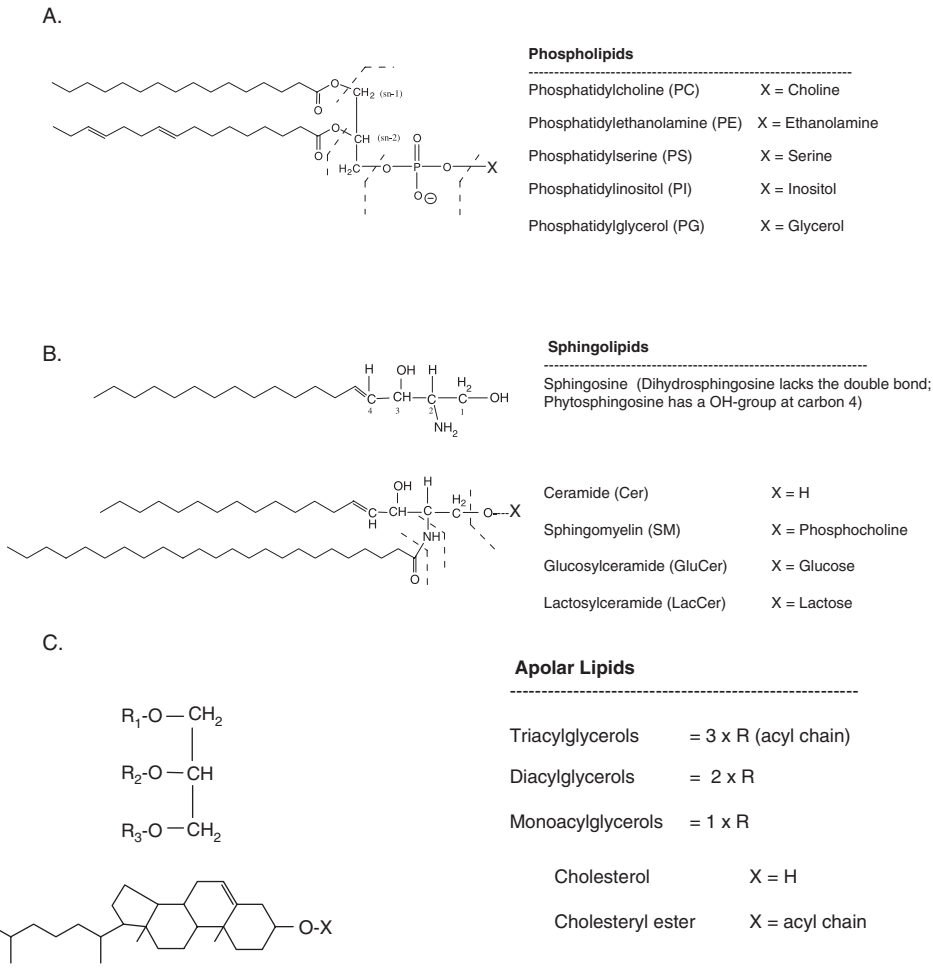


Fig. 2. *Panel A:* Structure of some common glycerophospholipids and the typical fragmentation sites. *Panel B:* Structure of some common sphingolipids and typical fragmentation sites. *Panel C:* Structure of acylglycerols, cholesterol, and cholesterylesters.

2.3. Data analysis

While successful data acquisition is an obvious prerequisite for the elucidation of a lipidome, the task does not end there. Data analysis can prove to be a difficult undertaking due to complexity of the samples and the various corrections needed for accurate quantification. The major tasks to be performed are: (i) identification of all relevant signals, (ii) assignment of signals to lipids, adducts, and fragments, (iii) isotopic correction, (iv) deconvolution of overlapping signals, and (v) quantification according to internal standards.

Since data analysis is a complicated and very time-consuming task, attempts have been made to automate this process. Kurvinen et al. have developed an algorithm that allows automated assignment and quantification of TAGs based on MS data [35,36]. A similar algorithm was implemented by Liebisch et al. for determination of PC and SM species from PI spectra [37]. Han and Gross have published a protocol for the analysis of phospholipid and triglyceride molecular species (with acyl chain information) based on a two-dimensional matrix of data produced by multiple MS/MS scans [4]. We have recently developed a method that allows accurate quantitative analysis of lipidomes based on two-dimensional LC–MS data [19].

Quantification of lipids based on MS data is significantly complicated by the fact that the ionization efficiency depends markedly on the structure of the polar head group and acyl chains [3,9,23]. Additional problems arise from suppression effects depending on the concentration of lipids, ions, or other impurities present in the sample. Different solutions to these problems have been suggested including (i) the use of multiple internal standards for each lipid class [22,23], (ii) working at low concentration range [4], (iii) determining the response–concentration curve for each individual compound [38], and (vi) spiking with natural species [37]. Notably, suppression effects can be very pronounced in MALDI, as shown, for instance, by the finding that PC can preclude the analysis of the other phospholipid classes present [39].

Beyond the analysis of the primary MS data, there is a need for tools allowing one to correlate lipid compositions with other compositional and functional data. This is because the lipid profiles alone are difficult to interpret in terms of mechanisms and functions. Lipid databases are also needed, but are yet under construction [40].

3. Phospholipids

Phospholipids are the most abundant lipids in nearly all biological membranes. Besides forming the lipid bilayer "backbone" of biological membranes, they participate in several other processes, such as signal transduction, endocytic sorting, activation of membrane enzymes, and alveolar function. The general structures of glycerophospholipids and the characteristic fragments formed in ESI are shown in Table 1.

3.1. Phosphatidylcholine and sphingomyelin

PC is the main component of mammalian membranes and lipoproteins. PC is also keenly involved in cell signaling [41]. SM is specifically enriched to the plasma membranes of cells and is also abundant in lipoproteins. Notably, SM and cholesterol are thought to form segregated, ordered domains within the cellular membranes [42]. Such domains, also referred as "rafts," are presently under intensive investigation due to their putative roles in intracellular lipid and protein sorting, cellular signaling molecules, and various diseases [43].

Table 1
Lipid class-specific fragments produced by collision-activated degradation

	Lipid class	Precursor ion	Fragment ion	Fragmentation type and fragment mass
Acyl glycerols	TAG, DAG, MAG	$[M+NH_4]^+$	$[M+NH_4\text{- }NH_3\text{-RCOO}]^+$	NL of fatty acid
Steryl esters	Cholesteryl esters	$[M+NH_4]^+$	$[C_{27}H_{45}]^+$	PI 369.35
Glycerophospholipids	PA, PI, PS, PG	$[M\text{-}H]^-$	$[C_3H_6O_5P]^+$	PI 153.00
	PC	$[M+H]^+$	$[C_5H_{15}NO_4P]^+$	PI 184.07
	PC	$[M+Li]^+$	$[M+Li\text{- }N(CH_3)_3]^+$	NL 59.07
	PC	$[M+HCOO]^-$	$[M+HCOO\text{-}HCOOCH_3]^-$	NL 60.02
	PE	$[M+H]^+$	$[M+H\text{-}C_2H_8NO_4P]^+$	NL 141.02
	PS	$[M+H]^+$	$[M+H\text{-}C_3H_8NO_6P]^+$	NL 185.01
	PS	$[M\text{-}H]^-$	$[M\text{-}H\text{- }C_3H_5NO_2]^-$	NL 87.03
	PI	$[M\text{-}H]^-$	$[C_6H_{10}O_8P]^-$	PI 241.01
	Lysophospholipids PG, PI, PA, PS	$[M\text{-}H]^-$	$[PO_3]^-$	PI 78.96
Sphingolipids	SM	$[M+H]^+$	$[C_5H_{15}NO_4P]^+$	PI 184.07
	NGSL	$[M+H]^+$	For example, $[C_{18}H_{34}N]$ (sphingosine)	PI 264
	Sulfatide	$[M\text{-}H]^-$	$[HSO_4]^-$	PI 96.95

Abbreviations: DAG, diacylglycerol; MAG, monoacylglycerol; TAG, triacylglycerol; NGSL, neutral glycosphingolipids; PA, phosphatidic acid; PC, phosphatidylcholine; PE, phosphatidylethanolamine; PG, phosphatidylglycerol; PI, phosphatidylinositol; PS, phosphatidylserine; SM, sphingomyelin. See the text for other details.

Both PC and SM contain a phosphocholine head group, which makes the molecules zwitterionic and largely dictates their ionization and fragmentation behavior. PC and SM readily form $[M+H]^+$ ions which upon CAD yield an abundant phosphocholine fragment with *m/z* 184 (*cf.* Table 1) and can thus be selectively detected by scanning for parents of this ion. In the presence of different salts, PC and SM form both cation and anion adducts, which may be utilized for the elucidation of the fatty acids present in the molecules [19,29,34,44,45]. In case of SM, also a fragment indicative of the long-chain base is found [45].

Quantification of SM species is often hampered by spectral overlap due to the much more abundant PC species. This problem can be solved either by removing the PCs with alkaline hydrolysis *in silico* using a spectral subtraction protocol [37]

or by employing LC–MS [19]. Also, NL scanning in the negative mode with properly adjusted instrument settings seems to allow selective detection of SM in the presence of PC [46]. Many tissues contain PC species with an ether-linked alkyl chain in the *sn*-1 position. At present, these molecules are probably most readily analyzed by LC–MS [19,47].

3.2. Phosphatidylethanolamine

PE is another major phospholipid in the membranes of eukaryotic cells. PE plasmalogens, which comprise ~50% of total PE lipids, contain a vinyl ether-linked hydrocarbon chain in the *sn*-1 position instead of an ester-linked chain. Plasmalogens represent a major source of arachidonic acid, an important second messenger, and are also thought to have an important protective role against oxidative damage [48].

PE, including the plasmalogens, can be detected either as the $[M+H]^+$ ion in the positive mode or as $[M\text{-}H]^-$ ion in the negative mode, particularly under alkaline conditions [3,4,21]. In the positive mode, PE exhibits a specific NL of 141 (phosphoethanolamine), whereas in the negative mode carboxylate anions from fatty acid fragmentation are the most abundant fragments.

The presence of alkenyl ether (plasmalogens), alkyl ether, and diacyl species complicates the analysis of PE, due to extensive spectral overlaps between these sub-classes. Also, ether PEs do not lose their head group as readily as the diacyl acyl species, and thus scanning for the neural loss of 141 underestimates their concentration [49]. This can be avoided by derivatization of the amino group [50]. The PE sub-classes can be separated and analyzed by using normal-phase LC–MS [19,47].

3.3. Phosphatidylserine

Phosphatidylserine (PS) usually comprises ~3–5 mol% of total phospholipids of mammalian cells, but its concentration is much higher (15–33 mol%) in the inner leaflet of the plasma membrane. PS is an activator of the protein kinase C, and PS exposure to the cell surface plays a key role in platelet aggregation as well as in elimination of apoptotic cells [51].

Due to its net negative charge, PS yields abundant $[M\text{-}H]^-$ ions in negative ion mode while $[M+H]^+$ or $[M+Na]^+$ ions are formed in positive mode. Upon CAD, the $[M\text{-}H]^-$ ion readily loses [serine-H_2O] as a neutral fragment as well as the fatty acids as carboxylate anions [21,22,30,52]. Thus, scanning for an NL of 87 allows specific detection of PS species in crude lipid extracts, while information on the fatty acyl substituents can be obtained by product or precursor ion scanning. Notably, PS can also be analyzed in the positive mode based on an NL of phosphoserine (185 Da) [21,22,30].

3.4. Phosphatidylglycerol, lysobisphosphatidic acid, and phosphatidic acid

Phosphatidylglycerol (PG) is a major lipid in many bacterial membranes, but comprises only 1–2 mol% of phospholipids in animal tissues. However, PG is much more abundant (~10 mol%) in the lung surfactant, thus pointing to an important role therein. Lysobisphosphatidic acid (LBPA) is another minor lipid and is present only in lysosomes and secondary endosomes where it seems to play an important role in controlling the formation of multivesicular bodies [53]. Phosphatidic acid (PA) is a key intermediate of glycerolipid synthesis and is also involved in cell signaling [41]. Lyso-PA (PA lacking one fatty acid moiety) is a highly active signaling molecule, which is involved in proliferation, migration, and survival of cells [54].

Ionization of PG can be achieved in positive mode as $[M+H]^+$ or $[M+Na]^{3+}$, but is usually analyzed in the negative mode as the $[M\text{-}H]^-$ ion. LBPA is isobaric with PG containing the same (or isobaric) fatty acids, which, together with the fact that abundant class-specific fragments are produced by neither lipid, complicates their analysis by the MS/MS method. Both PG and LBPA can be analyzed by scanning for precursors of *m/z* 153 $[\text{phosphoglycerol-}H_2O]^-$ [44,55], but the 153 ion is produced from several other phospholipids as well. However, PG and LBPA can be readily analyzed by using normal-phase LC–MS [18,56], since LBPA elutes well before PG (our unpublished data).

PA forms $[M\text{-}H]^-$ ions avidly and can be usually distinguished from other phospholipids due to its smaller mass [57]. The fragmentation of the $[M\text{-}H]^-$ is quite similar to that of PG and LBPA, the most significant fragments being fatty acid anions as well as $[\text{phosphoglycerol-}H_2O]^-$ (*m/z* 153) [58].

3.5. Cardiolipin

Cardiolipin (CL) is a complex lipid which, in essence, consists of PG molecules attached to a PA molecule. In mammals it is present only in mitochondria (mainly in the inner membranes), and C18 fatty acids, particularly linoleate (18:2), are predominant in CLs [59]. CL is an activator of many enzymes of the respiratory chain and its deficiency results in serious defects. In the Barth syndrome CL is diminished, probably due to impaired remodeling of its fatty acids [60].

CL forms both $[M\text{-}H]^-$ and $[M\text{-}2H]^{2-}$ ions and its fatty acid moieties can be deduced from the product ion spectrum [33,56,61]. Complete structural analysis of CL, i.e., identification of the positions of the fatty acids in the molecule, requires the use of an ion trap instrument and multiple fragmentation steps [62,63]. Quantification of the positional isomers is not straightforward, however, due to acyl chain structure-dependent fragmentation efficiency [64].

3.6. Phosphatidylinositol and polyphosphoinositides

Phosphatidylinositol (PI) and its phosphorylated derivatives PI-4-phosphate (PI-4-P), PI-4,5-bisphosphate (PI-4,5-P_2), etc., are widespread in nature, albeit they do not occur in bacteria. While the biological role(s) of PI are not clear yet, polyphosphoinositides are known to be crucial in various physiological phenomena, including intracellular signaling and vesicle traffic [65,66].

PI, as other acidic lipids, is best analyzed in the negative mode [21]. The fragment ion of *m/z* 241 (inositol phosphate-H_2O) is specific for PI and its phosphorylated derivatives, thus allowing specific detection by precursor ion scanning [22]. PI-P shows additional signals at *m/z* 321 (inositol diphosphate-H_2O) and *m/z* 303 (inositol diphosphate-2 $\times$ H_2O), while PI-3,4-P shows a fragment corresponding to inositol triphosphate-H_2O at *m/z* 401 [28]. These fragments allow specific detection of polyphosphoinositides by precursor ion scanning [65]. For each lipid, the fatty acid residues can be identified from the product ion spectra [28]. Analysis of PI and PIP with MALDI is also possible [67], but only with rather low sensitivity [39].

4. Acylglycerols

4.1. Triacylglycerols

TAGs in the adipose tissue serve as the main energy store of the body as well as "carriers" of fatty acids in lipoproteins. TAGs are also found in the cytoplasmic lipid droplets present in cells of many tissues.

TAGs can be detected as adducts of NH_4^+, Na^+, Li^+, or similar ions added to the solvent [4]. Only small amounts of H^+ adducts are found under acidic conditions [68]. The cation adducts of TAGs usually do not fragment spontaneously, which greatly assist the interpretation of their spectra. Diagnostic fragments can be obtained by either in-source fragmentation or CAD, or some other activation method [69]. The fragments are typically formed via a loss of a fatty acid. High-energy CAD produces charge-remote fragments allowing the determination of the positions of double bonds in the acyl moieties [70].

A special complication in the analysis of TAGs is the large number of isobars resulting from presence of different combinations of the three acyl moieties and their regioisomers. The "shotgun lipidomics" approach [4] developed by Han and Gross provides detailed structural information on TAGs as lithium adducts based on the molecular mass and NL scans for the expected fatty acids. However, complete structural analysis of TAGs requires the use of hyphenated methods like HPLC–MS/MS [71] or silver-ion chromatography [72] with MS [73]. A review of separation methods is available [74].

Besides ESI, chemical ionization (CI or APCI) has been used for TAG analysis. In CI in the negative ion mode [75], ionization takes place via abstraction of a proton to

form $[M\text{-}H]^-$ ions. Both $[RCOO]^-$ and $[M\text{-}H\text{-}RCOOH\text{-}100]^-$ ions are observed and can be used to assign the individual fatty acids as well as the regioisomers. In APCI (usually combined with HPLC) $[M+H]^+$ ions are detected [76] and fragmentation can be used to assign regioisomers [77]. With MALDI, $[M+Na]^+$ ions but no $[M+H]^+$ ions are observed [6], possibly due to prompt fragmentation of the latter to $[M\text{-}RCOO]^+$ ions [78]. The observed, usually not very abundant, fragments do not contain Na^+ and are again diagnostic of the fatty acid residues. However, the relative intensities of the fragment ions do not seem to reflect the position of the fatty acids [78].

4.2. Diacylglycerols and monoacylglycerols

These partially acylated glycerols are usually present only in trace amounts in fresh tissues. They can be analyzed analogously to TAGs using any of the techniques mentioned above [79].

5. Sphingolipids

The best characterized functions of sphingolipids are related to the structure of biological membranes, signal transduction, and biological recognition of these molecules [42,80–82]. Common to all lipids of this class is the sphingoid base, which comprises the backbone of the molecule. The sphingoid base is acylated, usually with a long-chain fatty acid, to the amino group at position 2 to produce ceramide, which serves as a precursor for the synthesis of more complex sphingolipids. Ceramide is then appended at the 1-position of the sphingoid base to give rise to a variety of different glycosphingolipids or sphingophospholipids [83,84].

5.1. Free sphingoid bases

Sphingosine (d18:1) or sphinganine (d18:0) comprises the backbone of all sphingolipids. They and their phosphorylated derivatives are also important second messengers involved in functions such as cell growth, differentiation, and apoptosis [81,85]. Sphingoid bases are readily protonated to form $[M+H]^+$ ions (e.g., *m/z* 300.3 for d18:1 and *m/z* 302.3 for d18:0). Upon collisional activation they lose H_2O to yield a carbocation and can thus be analyzed by scanning for the precursors of *m/z* 282/284 (or *m/z* 264/266 for doubly dehydrated molecules) [86]. The phosphorylated sphingoid bases can be analyzed in either positive or negative mode as $[M+H]^+$ or $[M\text{-}H]^-$ ion, respectively. Upon collisional activation the $[M+H]^+$ ion of sphingosyl-1-phosphate forms a carbocation product ion (*m/z* 264 for d18:1 base) [87], while the $[M\text{-}H]^-$ ion gives rise to abundant $[PO_3]^-$ ion (*m/z* 79). Scanning for the precursors of *m/z* 79 thus yields all free phosphorylated sphingoid bases [88].

5.2. Ceramides

Ceramides perform similar vital functions as the sphingoid bases [85,89,90] and are also key structural components of stratum corneum [91]. Ceramides produce abundant $[M+H]^+$ ions and upon CAD readily lose the fatty acid and one or two water molecules from the sphingoid base. Accordingly, ceramides can be analyzed by scanning for the precursors of different sphingoid base fragments, e.g., *m/z* 264 for the d18:1 [88,92,93]. Alternatively, CAD of deprotonated ceramides enables their analysis based on the sphingoid base-specific NL in the negative ion mode [94]. The benefit in the latter approach is the formation of a single fragmentation product, whereas during scanning for the precursors of the sphingoid bases in positive ion mode the analysis may be complicated by the occurrence of double peaks due to loss of water from the $[M+H]^+$ ion.

5.3. Neutral glycosphingolipids

The most abundant neutral glycosphingolipids in mammals are galactosyl-, glucosyl-, and lactosylceramides (GalCer, GlcCer, and LacCer, respectively). GalCer is abundant in tissues of the central nervous system, especially in the myelin sheath. It is also a precursor to sulfatides and the ganglioside GM4, while GlcCer can be converted to LacCer, which in turn serves as a precursor for a large number of different glycosphingolipid species [83,95]. GalCer and GlcCer and/or their derivatives are involved in the regulation of cell growth, protein trafficking and sorting, and modulation of cell adhesion and cell morphogenesis [96,97].

Neutral glycosphingolipids are best detected in the positive ionization mode as $[M+H]^+$ or $[M+Li]^+$ ions, although they also form $[M+HCOO]^-$ and $[M+Cl]^-$ adducts in the negative ionization mode [19,88,98,99]. Upon low-energy CAD, the [M+H] ions dissociate by a neural loss of the glycan, the charge remaining in the ceramide moiety. At higher collision energies, the sphingoid base (-1 or 2 H_2O) is the characteristic product ion and thus these lipids can be selectively detected by scanning for precursors of the different sphingoid bases. In some tissues, galactosyl- and glucosylceramides contain additional hydroxyl groups in the sphingoid base and/or in the fatty acyl moiety. LC–MS allows for accurate quantitation of these lipids, especially with MS/MS [9,19]. It must also be noted that GlcCer and GalCer are isobaric, and thus distinction of these molecules requires their prior separation by LC [87]. Alternatively, they can be distinguished by CAD of chlorinated adducts [99].

5.4. Sulfatides

Sulfatides are derived from GalCer via esterification of a sulfate group to 3-hydroxyl of the galactose moiety. Sulfatides are abundant in the brain, where they act as essential components of the myelin sheaths, and also appear to participate

in protein trafficking, signal transduction, and neuronal cell differentiation [100]. Sulfatide has been implicated as a potential cause in some autoimmune diseases (e.g., diabetes mellitus), and substantial loss of sulfatides has been detected in brains of Alzheimer patients [101].

Sulfatides are readily deprotonated to form $[M\text{-}H]^-$ ions and upon CAD they yield an abundant HSO_4^- ion (*m/z* 97), which allows convenient analysis of sulfatides by precursor ion scanning. Identification of the long-chain base and the fatty acyl constituents is also feasible, since a daughter ion diagnostic of the acyl moiety can be detected. The method allows identification of isobaric species and those containing α-hydroxylated fatty acid substituents as well [102,103]. Alternatively, the LC–MS method can be used [19,104]. Notably, LC–MS provides a clear distinction between the α-hydroxylated and nonhydroxylated species and thus their accurate quantitation.

5.5. Gangliosides

Gangliosides are negatively charged sialic acid containing glycosphingolipids with a greatly varying number of sugar moieties in their polar head group [95]. They are enriched by the outer leaflet of the plasma membrane and participate in vital functions including signal transduction, cell–cell interactions, as well as cell proliferation, differentiation, and apoptosis [82,97].

Due to their low abundance, characterization of gangliosides usually involves isolation from other lipid classes by TLC, LC, or capillary electrophoresis before the MS analysis [105–108]. The gangliosides form several types of ions. In the negative mode, they appear as deprotonated pseudomolecular ions, in which the number of charges usually corresponds to the number of sialic acids present [105,109, 110]. In the presence of Na^+ ions also single or multiple sodium adducts are detected [105,107]. The negative ions gives rise to a characteristic fragmentation pattern providing information on the structure of the oligosaccharide and ceramide moieties [111]. All gangliosides produce the fragment *m/z* 290 deriving from the sialic acid moiety, which has been utilized for direct analysis of gangliosides from lipid extracts by precursor ion scanning [112]. Information of the substitutions of the individual sugars can be obtained by MS/MS analysis of permethylated molecules in the positive mode.

6. Sterols

6.1. Cholesterol and other sterols

Cholesterol is an essential component of mammalian membranes and is a precursor for all steroid hormones. However, cholesterol can accumulate in certain tissues

and cause serious pathological consequences in the body, such as artherosclerosis. Recently, the analysis of oxidation products of cholesterol (and other lipids) has gained interest due to their putative pathophysiological role [113]. There is also growing interest toward various plant sterols (phytosterols) as their intake significantly reduces plasma cholesterol levels [114].

Sterols cannot be analyzed by ESI–MS without derivatization as they are not readily ionized [115]. Sandhoff et al. have used chemical sulfatation to achieve high-sensitivity detection of cholesterol [116]. Cholesterol has also been derivatized with dimethylglycine, MDMABS [117], or ferrocenecarbamate [118]. Notably, derivatization can be avoided by using APCI or APPI, which have been applied for the analysis of cholesterol and other sterols [119–122] or oxidized cholesterol [123].

6.2. Steryl esters and steroid hormones

Although cholesteryl esters (CEs) play an important role in cholesterol metabolism, there are few reports on CE analysis by MS. The behavior of cholesteryl esters in ESI is very similar to that of TAGs, and thus NH_4^+ [124,125] and Ag^+ adducts [126] have been used for their analysis. Fragmentation of the adduct ions produces a characteristic ion of *m*/*z* 369.35 corresponding to [cholesterol-H_2O]$^+$. Other steryl esters behave similarly (e.g., estrone esters [127]). With APPI, the [M+H-H_2O]$^+$ ions are the most prominent ones [121].

Steroids are more polar than sterols, and can be analyzed by ESI and APCI without derivatization [128], albeit derivatization increases the sensitivity of detection [115,129].

7. Medical applications of MS-lipidomics

MS-lipidomics appears to have a great potential in medicine including diagnosis and therapy, analysis of mechanisms underlying various diseases, or other pathophysiological conditions and nutrition. Below we will briefly discuss the results obtained so far in these fields.

7.1. Diagnostics and therapy

There are several studies indicating the potential of MS-lipidomics in the diagnosis of genetic and other diseases. Perhaps the most prominent example is the diagnosis of the Barth syndrome based on the analysis of the phospholipid species of the platelets. Platelets from patients with this disorder contained greatly reduced levels of the CL molecular species containing four linoleyl (18:2) residues as compared to unaffected individuals [61]. LC–MS analysis allows the detection of this diagnostic

parameter far more easily than traditional methods of lipid analysis, and potentially also allows one to monitor the therapeutic effects of fatty acid supplementation [60]. MS-lipidomics is also helpful in diagnosis and therapy of other hereditary diseases like Fabry disease based on increased levels of, e.g., trihexosyl and lactosylceramides found in the urine due to lack of the lysosomal α-galactosidase [130,131]. Also, some peroxisomal disorders like X-linked adrenoleykodystrophy and Zellweger syndrome could be diagnosed based on MS analysis of long-chain fatty acids [132], ceramides [133], or PE plasmalogens (our unpublished data). MS-lipidomics may also allow one to diagnose the Lowe syndrome based on altered polyphosphoinositol lipid composition [65]. MS analysis of ascitic fluid from patients with ovarian cancer has revealed increased levels of some lysophospholipids, which could thus provide useful biomarkers for this disease [134,135]. Yet, MS-lipidomics has significant potential in diagnosing and judging the predisposition for various multifactorial disorders, such as type 2 diabetes and artherosclerosis, particularly when combined with the analysis of other biomarkers [136].

7.2. Disease mechanisms

MS-lipidomics is also likely to be very helpful in resolving the metabolic defects underlying various common diseases, like type 2 diabetes, atherosclerosis, Alzheimer's disease, and different cancers. So far, MS-lipidomics has been helpful in understanding the mechanism of lipid accumulation in atherosclerotic plaques [137,138], optic nerve hypoplasia [139], cystic fibrosis [140], neuronal-ceroid lipofuscinosis [18], aggressive periodontal tissue damage [141], ulcerative colitis [142], glycosphingolipid disorders [143], and diabetic cardiomyopathy [144].

Besides compositional data, MS is also highly useful to study lipid metabolism in cells or whole animals. A variety of heavy isotope (^{2}H, ^{13}C)-labeled lipid precursors are commercially available and can be used to obtain highly detailed data regarding the biosynthetic routes and kinetics of lipid metabolism. So far this approach has been employed to study, e.g., the contribution of different pathways to the biosynthesis of PC in malignant cells [25], β-oxidation of fatty acids in peroxisomal disorders [145], surfactant PC synthesis [146], and lipid metabolism related to obesity [147].

7.3. Nutrition and other issues

MS-lipidomics also offers a powerful tool for nutritional studies. For instance, the effect of caloric restriction on lipid composition of murine myocardium has been investigated by this method [148]. In another study, the effect of structure of TAGs on the chylomicron TAG composition in humans was investigated [149]. MS-lipidomics also appears to be useful in resolving the mechanisms underlying drug addiction and related issues [150].

8. Future trends

Since MS-based lipidomics has evolved only very recently and the methodology is evolving rapidly, we can expect major advancements in understanding the physiological roles of the multitude of lipid species present in the human body within the next few years. However, this will require integration of the lipidomics data with those of the other "omics," which remains a major challenge at this time. It does not seem unrealistic to assume that MS-lipidomics will become a routine method in the clinic for screening for rare lipid disorders as well as the predisposition for a variety of common lipid-related pathological conditions such as atherosclerosis, type 2 diabetes, as well as the Alzheimer's disease.

9. Conclusions

Recent methodological developments in MS have made quantitative analysis of the complex lipidomes of mammalian cells and tissues feasible for the first time. Phospholipids can already be analyzed in a routine manner and the analysis of most other lipid classes should be possible as soon as some issues regarding quantification, particularly the availability of suitable standards, have been resolved. A detailed analysis of lipidomes is expected to provide a powerful tool for the diagnosis and understanding of the mechanism of various lipid-related diseases and disorders, including atherosclerosis and type 2 diabetes. This, however, will require integration of the lipidomics data with those of genomics, proteomics, and other "omics," which may not be trivial.

References

1. Christie, W.W., *Lipid Analysis Isolation, Separation, Identification and Structural Analysis of Lipids*. The Oily Press, Bridwater, UK, 2003.
2. Kates, M., *Techniques of Lipidology: Isolation, Analysis, and Identification of Lipids*. North-Holland Publishing Company, Amsterdam, 1972.
3. Pulfer, M. and Murphy, R.C., Electrospray mass spectrometry of phospholipids. *Mass Spectrom. Rev.*, **22**, 332–364 (2003).
4. Han, X. and Gross, R.W., Shotgun lipidomics: electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples. *Mass Spectrom. Rev.*, **24**, 367–412 (2005).
5. Forrester, J.S., Milne, S.B., Ivanova, P.T. and Brown, H.A., Computational lipidomics: a multiplexed analysis of dynamic changes in membrane lipid composition during signal transduction. *Mol. Pharmacol.*, **65**, 813–821 (2004).
6. Schiller, J., Suss, R., Arnhold, J., Fuchs, B., Lessig, J., Muller, M., Petkovic, M., Spalteholz, H., Zschornig, O. and Arnold, K., Matrix-assisted laser desorption and ionization time-of-flight (MALDI-TOF) mass spectrometry in lipid and phospholipid research. *Prog. Lipid Res.*, **43**, 449–488 (2004).

7. Welti, R. and Wang, X., Lipid species profiling: a high-throughput approach to identify lipid compositional changes and determine the function of genes involved in lipid metabolism and signaling. *Curr. Opin. Plant Biol.*, **7**, 337–344 (2004).
8. Johnson, D.W., Contemporary clinical usage of LC/MS: analysis of biologically important carboxylic acids. *Clin. Biochem.*, **38**, 351–361 (2005).
9. Merrill Jr., A.H., Sullards, M.C., Allegood, J.C., Kelly, S. and Wang, E., Sphingolipidomics: high-throughput, structure-specific, and quantitative analysis of sphingolipids by liquid chromatography tandem mass spectrometry. *Methods*, **36**, 207–224 (2005).
10. Murphy, R.C., Barkley, R.M., Zemski Berry, K., Hankin, J., Harrison, K., Johnson, C., Krank, J., McAnoy, A., Uhlson, C. and Zarini, S., Electrospray ionization and tandem mass spectrometry of eicosanoids. *Anal. Biochem.* (2005).
11. Folch, J., Lees, M. and Stanley, G.H.S., A simple method for the isolation and purification of total lipids from animal tissues. *J. Biol. Chem.*, **226**, 497–509 (1957).
12. Bligh, E.G. and Dyer, W.J., A rapid method of total lipid extraction and purification. *Can. J Biochem. Physiol.*, **37**, 911–917 (1959).
13. Kralik, S.F., Du, X.N., Patel, C. and Walsh, J.P., A method for quantitative extraction of sphingosine 1-phosphate into organic solvent. *Anal. Biochem.*, **294**, 190–193 (2001).
14. Bjerve, K.S., Daae, L.N.W. and Bremer, J., The selective loss of lysophospholipids in some commonly used lipid-extraction procedures. *Anal. Biochem.*, **58**, 238–245 (1974).
15. Ruiz-Gutierrez, V. and Perez-Camino, M.C., Update on solid-phase extraction for the analysis of lipid classes and related compounds. *J. Chromatogr. A*, **885**, 321–341 (2000).
16. Kim, H.Y., Wang, T.C. and Ma, Y.C., Liquid chromatography/mass spectrometry of phospholipids using electrospray ionization. *Anal. Chem.*, **66**, 3977–3982 (1994).
17. Taguchi, R., Hayakawa, J., Takeuchi, Y. and Ishida, M., Two-dimensional analysis of phospholipids by capillary liquid chromatography/electrospray ionization mass spectrometry. *J. Mass Spectrom.*, **35**, 953–966 (2000).
18. Kakela, R., Somerharju, P. and Tyynela, J., Analysis of phospholipid molecular species in brains from patients with infantile and juvenile neuronal-ceroid lipofuscinosis using liquid chromatography–electrospray ionization mass spectrometry. *J. Neurochem.*, **84**, 1051–1065 (2003).
19. Hermansson, M., Uphoff, A., Kakela, R. and Somerharju, P., Automated quantitative analysis of complex lipidomes by liquid chromatography/mass spectrometry. *Anal. Chem.*, **77**, 2166–2175 (2005).
20. Houjou, T., Yamatani, K., Imagawa, M., Shimizu, T. and Taguchi, R., A shotgun tandem mass spectrometric analysis of phospholipids with normal-phase and/or reverse-phase liquid chromatography/electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom.*, **19**, 654–666 (2005).
21. Kerwin, J.L., Tuininga, A.R. and Ericsson, L.H., Identification of molecular species of glycerophospholipids and sphingomyelin using electrospray mass spectrometry. *J. Lipid Res.*, **35**, 1102–1114 (1994).
22. Brugger, B., Erben, G., Sandhoff, R., Wieland, F.T. and Lehmann, W.D., Quantitative analysis of biological membrane lipids at the low picomole level by nano-electrospray ionization tandem mass spectrometry. *Proc. Nat. Acad. Sci. U.S.A.*, **94**, 2339–2344 (1997).
23. Koivusalo, M., Haimi, P., Heikinheimo, L., Kostiainen, R. and Somerharju, P., Quantitative determination of phospholipid compositions by ESI–MS: effects of acyl chain length, unsaturation, and lipid concentration on instrument response. *J. Lipid Res.*, **42**, 663–672 (2001).
24. Hunt, A.N., Clark, G.T., Attard, G.S. and Postle, A.D., Highly saturated endonuclear phosphatidylcholine is synthesized in situ and colocated with CDP–choline pathway enzymes. *J. Biol. Chem.*, **276**, 8492–8499 (2001).

25. DeLong, C.J., Hicks, A.M. and Cui, Z., Disruption of choline methyl group donation for phosphatidylethanolamine methylation in hepatocarcinoma cells. *J. Biol. Chem.*, **277**, 17217–17225 (2002).
26. Boumann, H.A., Chin, P.T., Heck, A.J., De Kruijff, B. and De Kroon, A.I., The yeast phospholipid *N*-methyltransferases catalyzing the synthesis of phosphatidylcholine preferentially convert di-C16:1 substrates both in vivo and in vitro. *J. Biol. Chem.*, **279**, 40314–40319 (2004).
27. Brouwers, J., Vernooij, E., Tielens, A.G.M. and van Golde, L.M.G., Rapid separation and identification of phosphatidylethanolamine molecular species. *J. Lipid Res.*, **40**, 164–169 (1999).
28. Hsu, F.F. and Turk, J., Characterization of phosphatidylinositol, phosphatidylinositol-4-phosphate, and phosphatidylinositol-4, 5-bisphosphate by electrospray ionization tandem mass spectrometry: a mechanistic study. *J. Am. Soc. Mass Spectrom.*, **11**, 986–999 (2000).
29. Hsu, F.F. and Turk, J., Electrospray ionization/tandem quadrupole mass spectrometric studies on phosphatidylcholines: the fragmentation processes. *J. Am. Soc. Mass Spectrom.*, **14**, 352–363 (2003).
30. Hsu, F.F. and Turk, J., Studies on phosphatidylserine by tandem quadrupole and multiple stage quadrupole ion-trap mass spectrometry with electrospray ionization: structural characterization and the fragmentation processes. *J. Am. Soc. Mass Spectrom.*, **16**, 1510–1522 (2005).
31. Ekroos, K., Chernushevich, I.V., Simons, K. and Shevchenko, A., Quantitative profiling of phospholipids by multiple precursor ion scanning on a hybrid quadrupole time-of-flight mass spectrometer. *Anal. Chem.*, **74**, 941–949 (2002).
32. Ekroos, K. and Shevchenko, A., Simple two-point calibration of hybrid quadrupole time-of-flight instruments using a synthetic lipid standard. *Rapid Commun. Mass Spectrom.*, **16**, 1254–1255 (2002).
33. Sparagna, G.C., Johnson, C.A., McCune, S.A., Moore, R.L. and Murphy, R.C., Quantitation of cardiolipin molecular species in spontaneously hypertensive heart failure rats using electrospray ionization mass spectrometry. *J. Lipid Res.*, **46**, 1196–1204 (2005).
34. Ekroos, K., Ejsing, C.S., Bahr, U., Karas, M., Simons, K. and Shevchenko, A., Charting molecular composition of phosphatidylcholines by fatty acid scanning and ion trap MS3 fragmentation. *J. Lipid Res.*, **44**, 2181–2192 (2003).
35. Kurvinen, J.P., Aaltonen, J., Kuksis, A. and Kallio, H., Software algorithm for automatic interpretation of mass spectra of glycerolipids. *Rapid Commun. Mass Spectrom.*, **16**, 1812–1820 (2002).
36. Kurvinen, J.P., Rua, P., Sjovall, O. and Kallio, H., Software (MSPECTRA) for automatic interpretation of triacylglycerol molecular mass distribution spectra and collision induced dissociation product ion spectra obtained by ammonia negative ion chemical ionization mass spectrometry. *Rapid Commun. Mass Spectrom.*, **15**, 1084–1091 (2001).
37. Liebisch, G., Lieser, B., Rathenberg, J., Drobnik, W. and Schmitz, G., High-throughput quantification of phosphatidylcholine and sphingomyelin by electrospray ionization tandem mass spectrometry coupled with isotope correction algorithm. *Biochim. Biophys. Acta-Mol. Cell Biol. Lipids*, **1686**, 108–117 (2004).
38. Zacarias, A., Bolanowski, D. and Bhatnagar, A., Comparative measurements of multicomponent phospholipid mixtures by electrospray mass spectroscopy: relating ion intensity to concentration. *Anal. Biochem.*, **308**, 152–159 (2002).
39. Muller, M., Schiller, J., Petkovic, M., Oehrl, W., Heinze, R., Wetzker, R., Arnold, K. and Arnhold, J., Limits for the detection of (poly-)phosphoinositides by matrix-assisted laser desorption and ionization time-of-flight mass spectrometry (MALDI–TOF MS). *Chem. Phys. Lipids*, **110**, 151–164 (2001).
40. Fahy, E., Subramaniam, S., Brown, H.A., Glass, C.K., Merrill Jr., A.H., Murphy, R.C., Raetz, C.R., Russell, D.W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., van Meer, G.,

VanNieuwenhze, M.S., White, S.H., Witztum, J.L. and Dennis, E.A., A comprehensive classification system for lipids. *J. Lipid Res.*, **46**, 839–862 (2005).
41. Foster, D.A. and Xu, L., Phospholipase D in cell proliferation and cancer. *Mol. Cancer Res.*, **1**, 789–800 (2003).
42. Simons, K. and Ikonen, E., Functional rafts in cell membranes. *Nature,* **387**, 569–572 (1997).
43. Rajendran, L. and Simons, K., Lipid rafts and membrane dynamics. *J. Cell Sci.*, **118**, 1099–1102 (2005).
44. Han, X.L. and Gross, R.W., Structural determination of picomole amounts of phospholipids via electrospray ionization tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **6**, 1202–1210 (1995).
45. Hsu, F.F. and Turk, J., Structural determination of sphingomyelin by tandem mass spectrometry with electrospray ionization. *J. Am. Soc. Mass Spectrom.*, **11**, 437–449 (2000).
46. Houjou, T., Yamatani, K., Nakanishi, H., Imagawa, M., Shimizu, T. and Taguchi, R., Rapid and selective identification of molecular species in phosphatidylcholine and sphingomyelin by conditional neutral loss scanning and MS3. *Rapid Commun. Mass Spectrom.*, **18**, 3123–3130 (2004).
47. Uran, S., Larsen, A., Jacobsen, P.B. and Skotland, T., Analysis of phospholipid species in human blood using normal-phase liquid chromatography coupled with electrospray ionization ion-trap tandem mass spectrometry. *J. Chromatogr. B*, **758**, 265–275 (2001).
48. Brites, P., Waterham, H.R. and Wanders, R.J., Functions and biosynthesis of plasmalogens in health and disease. *Biochim. Biophys. Acta*, **1636**, 219–231 (2004).
49. Berry, K.A.Z. and Murphy, R.C., Electrospray ionization tandem mass spectrometry of glycerophosphoethanolamine plasmalogen phospholipids. *J. Am. Soc. Mass Spectrom.*, **15**, 1499–1508 (2004).
50. Berry, K.A.Z. and Murphy, R.C., Analysis of cell membrane aminophospholipids as isotope-tagged derivatives. *J. Lipid Res.*, **46**, 1038–1046 (2005).
51. Vance, J.E. and Steenbergen, R., Metabolism and functions of phosphatidylserine. *Prog. Lipid Res.* (2005).
52. Hsu, F.F. and Turk, J., Studies on phosphatidylserine by tandem quadrupole and multiple stage quadrupole ion-trap mass spectrometry with electrospray ionization: structural characterization and the fragmentation processes (vol. 16, p. 1510, 2005). *J. Am. Soc. Mass Spectrom.*, **17**, 640–640 (2006).
53. Matsuo, H., Chevallier, J., Mayran, N., Le Blanc, I., Ferguson, C., Faure, J., Blanc, N.S., Matile, S., Dubochet, J., Sadoul, R., Parton, R.G., Vilbois, F. and Gruenberg, J., Role of LBPA and Alix in multivesicular liposome formation and endosome organization. *Science*, **303**, 531–534 (2004).
54. Moolenaar, W.H., van Meeteren, L.A. and Giepmans, B.N., The ins and outs of lysophosphatidic acid signaling. *Bioessays*, **26**, 870–881 (2004).
55. Larsen, A., Uran, S., Jacobsen, P.B. and Skotland, T., Collision-induced dissociation of glycero phospholipids using electrospray ion-trap mass spectrometry. *Rapid Commun. Mass Spectrom.*, **15**, 2393–2398 (2001).
56. Laurinavicius, S., Kakela, R., Bamford, D.H. and Somerharju, P., The origin of phospholipids of the enveloped bacteriophage phi6. *Virology*, **326**, 182–190 (2004).
57. Ivanova, P.T., Cerda, B.A., Horn, D.M., Cohen, J.S., McLafferty, F.W. and Brown, H.A., Electrospray ionization mass spectrometry analysis of changes in phospholipids in RBL-2H3 mastocytoma cells during degranulation. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 7152–7157 (2001).
58. Hsu, F.F. and Turk, J., Charge-driven fragmentation processes in diacyl glycerophosphatidic acids upon low-energy collisional activation. A mechanistic proposal. *J. Am. Soc. Mass Spectrom.*, **11**, 797–803 (2000).

59. Schlame, M., Brody, S. and Hostetler, K.Y., Mitochondrial cardiolipin in diverse eukaryotes—comparison of biosynthetic reactions and molecular acyl species. *Eur. J. Biochem.*, **212**, 727–735 (1993).
60. Valianpour, F., Wanders, R.J., Overmars, H., Vaz, F.M., Barth, P.G. and van Gennip, A.H., Linoleic acid supplementation of Barth syndrome fibroblasts restores cardiolipin levels: implications for treatment. *J. Lipid Res.*, **44**, 560–566 (2003).
61. Valianpour, F., Wanders, R.J.A., Barth, P.G., Overmars, H. and van Gennip, A.H., Quantitative and compositional study of cardiolipin in platelets by electrospray ionization mass spectrometry: application for the identification of Barth syndrome patients. *Clin. Chem.*, **48**, 1390–1397 (2002).
62. Hsu, F.F., Turk, J., Rhoades, E.R., Russell, D.G., Shi, Y.X. and Groisman, E.A., Structural characterization of cardiolipin by tandem quadrupole and multiple-stage quadrupole ion-trap mass spectrometry with electrospray ionization. *J. Am. Soc. Mass Spectrom.*, **16**, 491–504 (2005).
63. Hsu, F.F. and Turk, J., Characterization of cardiolipin from *Escherichia coli* by electrospray ionization with multiple stage quadrupole ion-trap mass spectrometric analysis of [M − 2H + Na](−) ions. *J. Am. Soc. Mass Spectrom.*, **17**, 420–429 (2006).
64. Hvattum, E., Hagelin, G. and Larsen, A., Study of mechanisms involved in the collision-induced dissociation of carboxylate anions from glycerophospholipids using negative ion electrospray tandem quadrupole mass spectrometry. *Rapid Commun. Mass Spectrom.*, **12**, 1405–1409 (1998).
65. Wenk, M.R., Lucast, L., Di Paolo, G., Romanelli, A.J., Suchy, S.F., Nussbaum, R.L., Cline, G.W., Shulman, G.I., McMurray, W. and De Camilli, P., Phosphoinositide profiling in complex lipid mixtures using electrospray ionization mass spectrometry. *Nat. Biotechnol.*, **21**, 813–817 (2003).
66. van Meer, G. and Sprong, H., Membrane lipids and vesicular traffic. *Curr. Opin. Cell Biol.*, **16**, 373–378 (2004).
67. Schiller, J., Arnhold, J., Benard, S., Muller, M., Reichl, S. and Arnold, K., Lipid analysis by matrix-assisted laser desorption and ionization mass spectrometry: a methodological approach. *Anal. Biochem.*, **267**, 46–56 (1999).
68. Duffin, K.L., Henion, J.D. and Shieh, J.J., Electrospray and tandem mass spectrometric characterization of acylglycerol mixtures that are dissolved in nonpolar solvents. *Anal. Chem.*, **63**, 1781–1788 (1991).
69. Hsu, F.F. and Turk, J., Structural characterization of triacylglycerols as lithiated adducts by electrospray ionization mass spectrometry using low-energy collisionally activated dissociation on a triple stage quadrupole instrument. *J. Am. Soc. Mass Spectrom.*, **10**, 587–599 (1999).
70. Cheng, C.F. and Gross, M.L., Complete structural elucidation of triacylglycerols by tandem sector mass spectrometry. *Anal. Chem.*, **70**, 4417–4426 (1998).
71. Segall, S.D., Artz, W.E., Raslan, D.S., Ferraz, V.P. and Takahashi, J.A., Analysis of triacylglycerol isomers in Malaysian cocoa butter using HPLC–mass spectrometry. *Food Res. Int.*, **38**, 167–174 (2005).
72. Christie, W.W., A stable silver-loaded column for the separation of lipids by high-performance liquid-chromatography. *J. High Resolut. Chromatogr. Chromatogr. Commun.*, **10**, 148–150 (1987).
73. Sandra, P., Medvedovici, A., Zhao, Y. and David, F., Characterization of triglycerides in vegetable oils by silver-ion packed-column supercritical fluid chromatography coupled to mass spectroscopy with atmospheric pressure chemical ionization and coordination ion spray. *J. Chromatogr. A*, **974**, 231–241 (2002).
74. Buchgraber, M., Ullberth, F., Emons, H. and Anklam, E., Triacylglycerol profiling by using chromatographic techniques. *Eur. J. Lipid Sci. Technol.*, **106**, 621–648 (2004).

75. Kallio, H., Tuomasjukka, S., Johansson, A., Tahvonen, R., Nieminen, N., Sjovall, O., Kurvinen, J.P. and Kivini, H., Analysis of regioisomers of triacylglycerols of northern currant seed oil by tandem mass spectrometry. *Eur. J. Lipid Sci. Technol.*, **107**, 101–106 (2005).
76. Nagy, K., Bongiorno, D., Avellone, G., Agozzino, P., Ceraulo, L. and Vekey, K., High performance liquid chromatography–mass spectrometry based chemometric characterization of olive oils. *J. Chromatogr. A*, **1078**, 90–97 (2005).
77. Byrdwell, W.C., The bottom-up solution to the triacylglycerol lipidome using atmospheric pressure chemical ionization mass spectrometry. *Lipids*, **40**, 383–417 (2005).
78. Al-Saad, K.A., Zabrouskov, V., Siems, W.F., Knowles, N.R., Hannan, R.M. and Hill, H.H., Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of lipids: ionization and prompt fragmentation patterns. *Rapid Commun. Mass Spectrom.*, **17**, 87–96 (2003).
79. Murphy, R.C., Fiedler, J. and Hevko, J., Analysis of nonvolatile lipids by mass spectrometry. *Chem. Rev.*, **101**, 479–526 (2001).
80. Brown, D.A. and London, E., Structure and function of sphingolipid- and cholesterol-rich membrane rafts. *J. Biol. Chem.*, **275**, 17221–17224 (2000).
81. Spiegel, S. and Merrill Jr., A.H., Sphingolipid metabolism and cell growth regulation. *FASEB J.*, **10**, 1388–1397 (1996).
82. Hakomori, S., Traveling for the glycosphingolipid path. *Glycoconjugate J.*, **17**, 627–647 (2000).
83. Sandhoff, K. and Kolter, T., Biosynthesis and degradation of mammalian glycosphingolipids. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **358**, 847–861 (2003).
84. Degroote, S., Wolthoorn, J. and van Meer, G., The cell biology of glycosphingolipids. *Semin. Cell Dev. Biol.*, **15**, 375–387 (2004).
85. Futerman, A.H. and Hannun, Y.A., The complex life of simple sphingolipids. *EMBO Rep.*, **5**, 777–782 (2004).
86. Lieser, B., Liebisch, G., Drobnik, W. and Schmitz, G., Quantification of sphingosine and sphinganine from crude lipid extracts by HPLC electrospray ionization tandem mass spectrometry. *J. Lipid Res.*, **44**, 2209–2216 (2003).
87. Sullards, M.C. and Merrill Jr., A.H., Analysis of sphingosine 1-phosphate, ceramides, and other bioactive sphingolipids by high-performance liquid chromatography–tandem mass spectrometry. *Sci. STKE*, **2001**, PL1 (2001).
88. Sullards, M.C., Analysis of sphingomyelin, glucosylceramide, ceramide, sphingosine, and sphingosine 1-phosphate by tandem mass spectrometry. *Methods Enzymol.*, **312**, 32–45 (2000).
89. Ruvolo, P.P., Ceramide regulates cellular homeostasis via diverse stress signaling pathways. *Leukemia*, **15**, 1153–1160 (2001).
90. Pettus, B.J., Chalfant, C.E. and Hannun, Y.A., Ceramide in apoptosis: an overview and current perspectives. *Biochim. Biophys. Acta*, **1585**, 114–125 (2002).
91. Goldstein, A.M. and Abramovits, W., Ceramides and the stratum corneum: structure, function, and new methods to promote repair. *Int. J. Dermatol.*, **42**, 256–259 (2003).
92. Gu, M., Kerwin, J.L., Watts, J.D. and Aebersold, R., Ceramide profiling of complex lipid mixtures by electrospray ionization mass spectrometry. *Anal. Biochem.*, **244**, 347–356 (1997).
93. Liebisch, G., Drobnik, W., Reil, M., Trumbach, B., Arnecke, R., Olgemoller, B., Roscher, A. and Schmitz, G., Quantitative measurement of different ceramide species from crude cellular extracts by electrospray ionization tandem mass spectrometry (ESI–MS/MS). *J. Lipid Res.*, **40**, 1539–1546 (1999).
94. Han, X., Characterization and direct quantitation of ceramide molecular species from lipid extracts of biological samples by electrospray ionization tandem mass spectrometry. *Anal. Biochem.*, **302**, 199–212 (2002).
95. Kolter, T., Proia, R.L. and Sandhoff, K., Combinatorial ganglioside biosynthesis. *J. Biol. Chem.*, **277**, 25859–25862 (2002).

96. Hakomori, S., Structure, organization, and function of glycosphingolipids in membrane. *Curr. Opin. Hematol.*, **10**, 16–24 (2003).
97. Bektas, M. and Spiegel, S., Glycosphingolipids and cell death. *Glycoconjugate J.*, **20**, 39–47 (2004).
98. Hsu, F.F. and Turk, J., Structural determination of glycosphingolipids as lithiated adducts by electrospray ionization mass spectrometry using low-energy collisional-activated dissociation on a triple stage quadrupole instrument. *J. Am. Soc. Mass Spectrom.*, **12**, 61–79 (2001).
99. Han, X. and Cheng H., Characterization and direct quantitation of cerebroside molecular species from lipid extracts by shotgun lipidomics. *J. Lipid Res.*, **46**, 163–175 (2005).
100. Merrill Jr., A.H., Schmelz, E.M., Dillehay, D.L., Spiegel, S., Shayman, J.A., Schroeder, J.J., Riley, R.T., Voss, K.A. and Wang, E., Sphingolipids—the enigmatic lipid class: biochemistry, physiology, and pathophysiology. *Toxicol. Appl. Pharmacol.*, **142**, 208–225 (1997).
101. Han, X., Holtzman, D.M., McKeel Jr., D.W., Kelley, J. and Morris, J.C., Substantial sulfatide deficiency and ceramide elevation in very early Alzheimer's disease: potential role in disease pathogenesis. *J. Neurochem.*, **82**, 809–818 (2002).
102. Hsu, F.F., Bohrer, A. and Turk, J., Electrospray ionization tandem mass spectrometric analysis of sulfatide. Determination of fragmentation patterns and characterization of molecular species expressed in brain and in pancreatic islets. *Biochim. Biophys. Acta*, **1392**, 202–216 (1998).
103. Hsu, F.F. and Turk, J., Studies on sulfatides by quadrupole ion-trap mass spectrometry with electrospray ionization: structural characterization and the fragmentation processes that include an unusual internal galactose residue loss and the classical charge-remote fragmentation. *J. Am. Soc. Mass Spectrom.*, **15**, 536–546 (2004).
104. Beausoleil, H.E., Lepine, F. and Dubreuil, J.D., LC–MS analysis of pig intestine sulfatides: interaction with *Escherichia coli* STb enterotoxin and characterization of molecular species present. *FEMS Microbiol. Lett.*, **209**, 183–188 (2002).
105. Metelmann, W., Muthing, J. and Peter-Katalinic, J., Nano-electrospray ionization quadrupole time-of-flight tandem mass spectrometric analysis of a ganglioside mixture from human granulocytes. *Rapid Commun. Mass Spectrom.*, **14**, 543–550 (2000).
106. Zamfir, A., Vukelic, Z. and Peter-Katalinic, J., A capillary electrophoresis and off-line capillary electrophoresis/electrospray ionization-quadrupole time of flight-tandem mass spectrometry approach for ganglioside analysis. *Electrophoresis*, **23**, 2894–2903 (2002).
107. Zamfir, A., Vukelic, Z., Bindila, L., Peter-Katalinic, J., Almeida, R., Sterling, A. and Allen, M., Fully-automated chip-based nanoelectrospray tandem mass spectrometry of gangliosides from human cerebellum. *J. Am. Soc. Mass Spectrom.*, **15**, 1649–1657 (2004).
108. Meisen, I., Peter-Katalinic, J. and Muthing, J., Direct analysis of silica gel extracts from immunostained glycosphingolipids by nanoelectrospray ionization quadrupole time-of-flight mass spectrometry. *Anal. Chem.*, **76**, 2248–2255 (2004).
109. Ju, D.D., Lai, C.C. and Her, G.R., Analysis of gangliosides by capillary zone electrophoresis and capillary zone electrophoresis–electrospray mass spectrometry. *J. Chromatogr. A*, **779**, 195–203 (1997).
110. Ghardashkhani, S., Gustavsson, M.L., Breimer, M.E. and Samuelsson, B.E., Negative electrospray ionization mass spectrometry of gangliosides, sulphatides and cholesterol 3-sulphate. *Rapid Commun. Mass Spectrom.*, **9**, 491–494 (1995).
111. Metelmann, W., Peter-Katalinic, J. and Muthing, J., Gangliosides from human granulocytes: a nano-ESI QTOF mass spectrometry fucosylation study of low abundance species in complex mixtures. *J. Am. Soc. Mass Spectrom.*, **12**, 964–973 (2001).
112. Tsui, Z.C., Chen, Q.R., Thomas, M.J., Samuel, M. and Cui, Z., A method for profiling gangliosides in animal tissues using electrospray ionization–tandem mass spectrometry. *Anal. Biochem.*, **341**, 251–258 (2005).

113. Olkkonen, V.M. and Lehto, M., Oxysterols and oxysterol binding proteins: role in lipid metabolism and atherosclerosis. *Ann. Med.*, **36**, 562–572 (2004).
114. Miettinen, T.A. and Gylling, H., Effect of statins on noncholesterol sterol levels: implications for use of plant stanols and sterols. *Am. J. Cardiol.*, **96**, 40–46 (2005).
115. Higashi, T. and Shimada, K., Derivatization of neutral steroids to enhance their detection characteristics in liquid chromatography–mass spectrometry. *Anal. Bioanal. Chem.*, **378**, 875–882 (2004).
116. Sandhoff, R., Brugger, B., Jeckel, D., Lehmann, W.D. and Wieland, F.T., Determination of cholesterol at the low picomole level by nano-electrospray ionization tandem mass spectrometry. *J. Lipid Res.*, **40**, 126–132 (1999).
117. Johnson, D.W., ten Brink, H.J. and Jakobs, C., A rapid screening procedure for cholesterol and dehydrocholesterol by electrospray ionization tandem mass spectrometry. *J. Lipid Res.*, **42**, 1699–1705 (2001).
118. Van Berkel, G.J., Quirke, J.M., Tigani, R.A., Dilley, A.S. and Covey, T.R., Derivatization for electrospray ionization mass spectrometry. 3. Electrochemically ionizable derivatives. *Anal. Chem.*, **70**, 1544–1554 (1998).
119. Toh, T.H., Prior, B.A. and van der Merwe, M.J., Quantitation of plasma membrane ergosterol of *Saccharomyces cerevisiae* by direct-injection atmospheric pressure chemical ionization/tandem mass spectrometry. *Anal. Biochem.*, **288**, 44–51 (2001).
120. Ruibal-Mendieta, N.L., Rozenberg, R., Delacroix, D.L., Petitjean, G., Dekeyser, A., Baccelli, C., Marques, C., Delzenne, N.M., Meurens, M., Habib-Jiwan, J.L. and Quetin-Leclercq, J., Spelt (*Triticum spelta* L.) and winter wheat (*Triticum aestivum* L.) wholemeals have similar sterol profiles, as determined by quantitative liquid chromatography and mass spectrometry analysis. *J. Agric. Food Chem.*, **52**, 4802–4807 (2004).
121. Lembcke, J., Ceglarek, U., Fiedler, G.M., Baumann, S., Leichtle, A. and Thiery, J., Rapid quantification of free and esterified phytosterols in human serum using APPI–LC–MS/MS. *J. Lipid Res.*, **46**, 21–26 (2005).
122. Nagy, K., Jakab, A., Pollreisz, F., Bongiorno, D., Ceraulo, L., Averna, M.R., Noto, D. and Vékey, K., Analysis of sterols by high-performance liquid chromatography/mass spectrometry combined with chemometrics. *Rapid Commun. Mass Spectrom.*, **20**, 2433–2440 (2006).
123. Razzazi-Fazeli, E., Kleineisen, S. and Luf, W., Determination of cholesterol oxides in processed food using high-performance liquid chromatography–mass spectrometry with atmospheric pressure chemical ionisation. *J. Chromatogr. A*, **896**, 321–334 (2000).
124. Kalo, P. and Kuuranne, T., Analysis of free and esterified sterols in fats and oils by flash chromatography, gas chromatography and electrospray tandem mass spectrometry. *J. Chromatogr. A*, **935**, 237–248 (2001).
125. Duffin, K., Obukowicz, M., Raz, A. and Shieh, J.J., Electrospray/tandem mass spectrometry for quantitative analysis of lipid remodeling in essential fatty acid deficient mice. *Anal. Biochem.*, **279**, 179–188 (2000).
126. Seal, J.R., Havrilla, C.M., Porter, N.A. and Hachey, D.L., Analysis of unsaturated compounds by Ag^+ coordination ionspray mass spectrometry: studies of the formation of the Ag^+/lipid complex. *J. Am. Soc. Mass Spectrom.*, **14**, 872–880 (2003).
127. Miilunpohja, M., Uphoff, A., Somerharju, P., Tiitinen, A., Wähälä, K. and Tikkanen, M.J., Fatty acid esterification of lipoprotein-associated estrone in human plasma and follicular fluid. *J. Steroid Biochem. Mol. Biol.*, **100**, 59–66 (2006).
128. Ma, Y.-C. and Kim, H.-Y., Determination of steroids by liquid chromatography/mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **8**, 1010–1020 (1997).
129. Chatman, K., Hollenbeck, T., Hagey, L., Vallee, M., Purdy, R., Weiss, F. and Siuzdak, G., Nanoelectrospray mass spectrometry and precursor ion monitoring for quantitative steroid analysis and attomole sensitivity. *Anal. Chem.*, **71**, 2358–2363 (1999).

130. Fuller, M., Sharp, P.C., Rozaklis, T., Whitfield, P.D., Blacklock, D., Hopwood, J.J. and Meikle, P.J., Urinary lipid profiling for the identification of Fabry hemizygotes and heterozygotes. *Clin. Chem.*, **51**, 688–694 (2005).
131. Whitfield, P.D., Calvin, J., Hogg, S., O'Driscoll, E., Halsall, D., Burling, K., Maguire, G., Wright, N., Cox, T.M., Meikle, P.J. and Deegan, P.B., Monitoring enzyme replacement therapy in Fabry disease—role of urine globotriaosylceramide. *J. Inherit. Metab. Dis.*, **28**, 21–33 (2005).
132. Valianpour, F., Selhorst, J.J., van Lint, L.E., van Gennip, A.H., Wanders, R.J. and Kemp, S., Analysis of very long-chain fatty acids using electrospray ionization mass spectrometry. *Mol. Genet. Metab.*, **79**, 189–196 (2003).
133. Pettus, B.J., Baes, M., Busman, M., Hannun, Y.A. and Van Veldhoven, P.P., Mass spectrometric analysis of ceramide perturbations in brain and fibroblasts of mice and human patients with peroxisomal disorders. *Rapid Commun. Mass Spectrom.*, **18**, 1569–1574 (2004).
134. Xiao, Y.J., Schwartz, B., Washington, M., Kennedy, A., Webster, K., Belinson, J. and Xu, Y., Electrospray ionization mass spectrometry analysis of lysophospholipids in human ascitic fluids: comparison of the lysophospholipid contents in malignant vs nonmalignant ascitic fluids. *Anal. Biochem.*, **290**, 302–313 (2001).
135. Sutphen, R., Xu, Y., Wilbanks, G.D., Fiorica, J., Grendys Jr., E.C., LaPolla, J.P., Arango, H., Hoffman, M.S., Martino, M., Wakeley, K., Griffin, D., Blanco, R.W., Cantor, A.B., Xiao, Y.J. and Krischer, J.P., Lysophospholipids are potential biomarkers of ovarian cancer. *Cancer Epidemiol. Biomarkers Prev.*, **13**, 1185–1191 (2004).
136. Clish, C.B., Davidov, E., Oresic, M., Plasterer, T.N., Lavine, G., Londo, T., Meys, M., Snell, P., Stochaj, W., Adourian, A., Zhang, X., Morel, N., Neumann, E., Verheij, E., Vogels, J.T., Havekes, L.M., Afeyan, N., Regnier, F., van der Greef, J. and Naylor, S., Integrative biological analysis of the APOE*3-leiden transgenic mouse. *OMICS*, **8**, 3–13 (2004).
137. Thukkani, A.K., McHowat, J., Hsu, F.F., Brennan, M.L., Hazen, S.L. and Ford, D.A., Identification of alpha-chloro fatty aldehydes and unsaturated lysophosphatidylcholine molecular species in human atherosclerotic lesions. *Circulation*, **108**, 3128–3133 (2003).
138. Ravandi, A., Babaei, S., Leung, R., Monge, J.C., Hoppe, G., Hoff, H., Kamido, H. and Kuksis, A., Phospholipids and oxophospholipids in atherosclerotic plaques at different stages of plaque development. *Lipids*, **39**, 97–109 (2004).
139. Rodemer, C., Thai, T.P., Brugger, B., Kaercher, T., Werner, H., Nave, K.A., Wieland, F., Gorgas, K. and Just, W.W., Inactivation of ether lipid biosynthesis causes male infertility, defects in eye development and optic nerve hypoplasia in mice. *Hum. Mol. Genet.*, **12**, 1881–1895 (2003).
140. Dombrowsky, H., Clark, G.T., Rau, G.A., Bernhard, W. and Postle, A.D., Molecular species compositions of lung and pancreas phospholipids in the cftr(tm1HGU/tm1HGU) cystic fibrosis mouse. *Pediatr. Res.*, **53**, 447–454 (2003).
141. Gronert, K., Kantarci, A., Levy, B.D., Clish, C.B., Odparlik, S., Hasturk, H., Badwey, J.A., Colgan, S.P., Van Dyke, T.E. and Serhan, C.N., A molecular defect in intracellular lipid signaling in human neutrophils in localized aggressive periodontal tissue damage. *J. Immunol.*, **172**, 1856–1861 (2004).
142. Ehehalt, R., Wagenblast, J., Erben, G., Lehmann, W.D., Hinz, U., Merle, U. and Stremmel, W., Phosphatidylcholine and lysophosphatidylcholine in intestinal mucus of ulcerative colitis patients. A quantitative approach by nanoelectrospray–tandem mass spectrometry. *Scand. J. Gastroenterol.*, **39**, 737–742 (2004).
143. Sandhoff, R., Hepbildikler, S.T., Jennemann, R., Geyer, R., Gieselmann, V., Proia, R.L., Wiegandt, H. and Grone, H.J., Kidney sulfatides in mouse models of inherited glycosphingolipid disorders: determination by nano-electrospray ionization tandem mass spectrometry. *J. Biol. Chem.*, **277**, 20386–20398 (2002).

144. Su, X., Han, X., Mancuso, D.J., Abendschein, D.R. and Gross, R.W., Accumulation of long-chain acylcarnitine and 3-hydroxy acylcarnitine molecular species in diabetic myocardium: identification of alterations in mitochondrial fatty acid processing in diabetic myocardium by shotgun lipidomics. *Biochemistry*, **44**, 5234–5245 (2005).
145. Kemp, S., Valianpour, F., Mooyer, P.A., Kulik, W. and Wanders, R.J., Method for measurement of peroxisomal very-long-chain fatty acid beta-oxidation in human skin fibroblasts using stable-isotope-labeled tetracosanoic acid. *Clin. Chem.*, **50**, 1824–1826 (2004).
146. Bernhard, W., Pynn, C.J., Jaworski, A., Rau, G.A., Hohlfeld, J.M., Freihorst, J., Poets, C.F., Stoll, D. and Postle, A.D., Mass spectrometric analysis of surfactant metabolism in human volunteers using deuterated choline. *Am. J. Respir. Crit. Care Med.*, **170**, 54–58 (2004).
147. Dolnikowski, G.G., Marsh, J.B., Das, S.K. and Welty, F.K., Stable isotopes in obesity research. *Mass Spectrom. Rev.*, **24**, 311–327 (2005).
148. Han, X., Cheng, H., Mancuso, D.J. and Gross, R.W., Caloric restriction results in phospholipid depletion, membrane remodeling, and triacylglycerol accumulation in murine myocardium. *Biochemistry*, **43**, 15584–15594 (2004).
149. Yli-Jokipii, K.M., Schwab, U.S., Tahvonen, R.L., Kurvinen, J.P., Mykkanen, H.M. and Kallio, H.P., Triacylglycerol molecular weight and to a lesser extent, fatty acid positional distribution, affect chylomicron triacylglycerol composition in women. *J. Nutr.*, **132**, 924–929 (2002).
150. Hillard, C.J., Lipids and drugs of abuse. *Life Sci.*, **77**, 1531–1542 (2005).

This page intentionally left blank

Part IV
Selected Medical Applications

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 12

Medical laboratory practice—possibilities for mass spectrometry

OLAF BODAMER*

Division of Biochemical and Paediatric Genetics, University Children's Hospital, Vienna, Austria

1. Introduction

Diagnosis of disorders that demonstrate (bio)chemical abnormalities frequently requires (semi-)quantitative analysis of small molecules, metabolites, peptides, proteins, or hormones in plasma, urine, or other body fluids [1–3]. Biomarkers may also be used to judge efficacy of treatment [1]. Modern medical laboratories employ, depending on their analytical repertoire, an array of analytical techniques including gas chromatography–mass spectrometry (GC–MS), high pressure liquid chromatography–mass spectrometry (HPLC–MS), liquid chromatography–mass spectrometry (LC–MS), or electrospray tandem mass spectrometry (ESI–MS/MS). In this context MS is typically used for the analysis of small molecules for the diagnosis of inborn errors of metabolism (IEM), endocrine disorders, in

*Tel.: +43-140-400 3232; Fax: +43-1406-3484. E-mail: olaf.bodamer@meduniwien.ac.at (O. Bodamer).

toxicology and pharmacology. In addition, ESI–MS/MS is used for neonatal screening of IEM [4–6, see relevant chapter for more detailed information] and together with GC–MS is used for analyzing isotopic enrichment in samples from *in vivo* studies employing stable isotope tracer [7, and below]. Additional analytical techniques that are frequently used in a medical laboratory include chromatography (LC, HPLC), enzyme-linked immunosorbent assay (ELISA), radio-immuno assay (RIA), and enzymatic and molecular techniques (PCR, sequencing).

2. Quality management

A well-organized medical laboratory should adapt quality standards not only for the analytical process itself but also for pre- and post-analytical processes [8,9]. Ideally, a quality management system according to International Organization for Standardization (ISO) should be implemented [9]. System requirements include definition of laboratory functions and responsibilities (i.e., laboratory manager, technician, secretary), standardization of analytical and operational procedures (SOP) and processes, and definition of quality policy and aims [9]. Particular emphasis should be placed on continued improvement using a system of regular reviews and internal and external audits [9]. Implementation of quality management systems is frequently required for laboratory certification and accreditation with health care organizations.

Core processes that are unique to the laboratory have to be identified and documented preferably using flowcharts (Fig. 1). This stepwise approach allows characterization of each single step, for example, in sample analysis in laboratories (shipment of samples, pre-analytical processing, analysis and post-analytical processing, reporting of results), determination of responsibilities, and documentation of analysis-specific SOP's. All documents are summarized in identical copies of the Quality Handbook that serve as important source of information for the laboratory staff (Fig. 1).

3. Inborn errors of metabolism

IEM are individually rare but may be collectively as frequent as 1 affected individual in 500 newborn infants. They encompass many different single-gene disorders affecting many aspects of cellular metabolism [10]. A significant portion relates to disorders of fatty acid, protein, and carbohydrate metabolism [10]. Analysis of IEM relies on semiquantitative or quantitative measurement of characteristic small molecules in different body fluids including blood, plasma, cerebral spinal fluid, and urine [11,12]. Diagnosis must then be typically

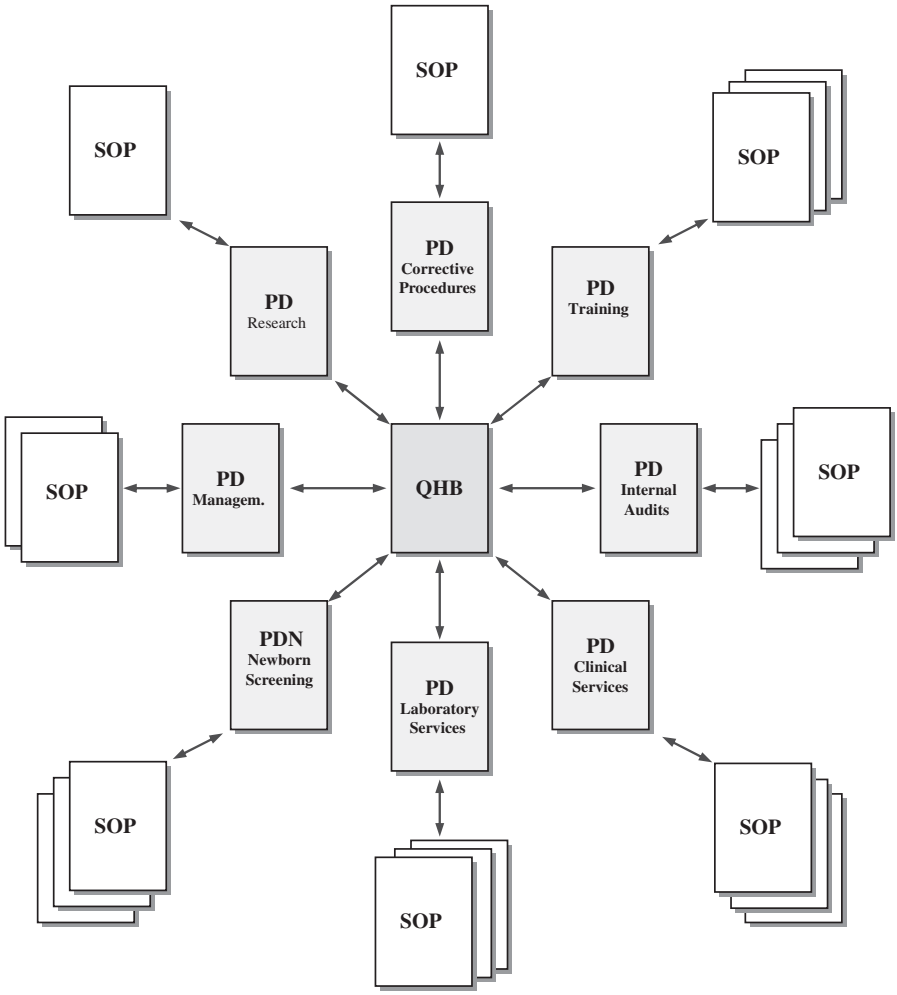


Fig. 1. Structure of a quality management system at the Division of Biochemical Genetics at the University Children's Hospital, Vienna. SOP: standard operation procedure; PD: process description; QHB: Quality Handbook.

confirmed using enzyme analysis in appropriate tissues and by genotyping. Some selected applications of MS for IEM are listed in Table 1. The complexity of a selected pathway is shown in Fig. 2 using the methionine–homocysteine cycle as example.

3.1. Analysis of homocysteine

Homocysteine is markedly elevated in different inborn errors of homocysteine metabolism such as cystathionine β-synthase, methionine synthase deficiencies,

Table 1
Selected applications of mass spectrometry in IEM

Analyte(s)	Disorder(s)	Material	Method	References
Oligosaccharides	Glycoproteinoses[a]	Urine	LC–MS/MS	[17]
Ceramides, glycosylceramides	LSD[b]	Urine	LC–MS/MS	[23]
Acylcarnitine ester	Fatty acid oxidation defects, organic acidopathies	Dried blood plasma	ESI–MS/MS	[11]
Amino acids	Amino acidopathies (PKU, tyrosinemia type I)	Dried blood plasma, urine	ESI–MS/MS	[12]
Guanidinoacetate	GAMT deficiency[c]	Dried blood plasma, urine	ESI–MS/MS	[24]
Homocysteine	Homocystinuria	Dried blood plasma	ESI–MS/MS	[13]
Cholesterol and metabolites	SLO[d] and other defects of cholesterol biosynthesis	Plasma	GC–MS, LC–MS/MS	[25]
Bile acid intermediates	Disorders of bile acid biosynthesis, peroxisomal disorders	Plasma, urine	GC–MS, LC–MS/MS	[26]

ESI–MS/MS: electrospray tandem mass spectrometry; GC–MS: gas chromatography–mass spectrometry; LC–MS/MS: liquid chromatography–tandem mass spectrometry.

[a] G_{M1} gangliosidosis, G_{M2} gangliosidosis, sialic acid storage disorder, sialidase/neuraminidase deficiency, galactosialidosis, I-cell disease, fucosidosis, and Pompe and Gaucher diseases.

[b] Lysosomal storage disorder including Gaucher, Fabry, Niemann–Pick A/B, Krabbe, and Pompe diseases.

[c] Guanidinoacetate methyltransferase deficiency.

[d] Smith–Lemli Opitz syndrome (defect of cholesterol biosynthesis).

and disorders of vitamin B12 metabolism affecting the conversion of Cbl-I to Cbl-II (Fig. 2). Although betaine-homocysteine methyltransferase deficiency in mice is known to cause hyperhomocysteinemia, this defect has not been reported in humans. In contrast, *S*-adenosyl homocysteine hydrolase deficiency may actually lead to low homocysteine levels. All compounds in this pathway can be analyzed although only few have clinical significance (homocysteine, methionine, arginine, ornithine, glycine, guanidinoacetate). Among the latter homocysteine may serve as an important biomarker to evaluate treatment efficacy and future risk for premature artherosclerosis. Analysis of homocysteine is readily made by LC–MS/MS or ESI–MS/MS [13]. This approach allows fast sample turnover and consequently screening of at-risk populations.

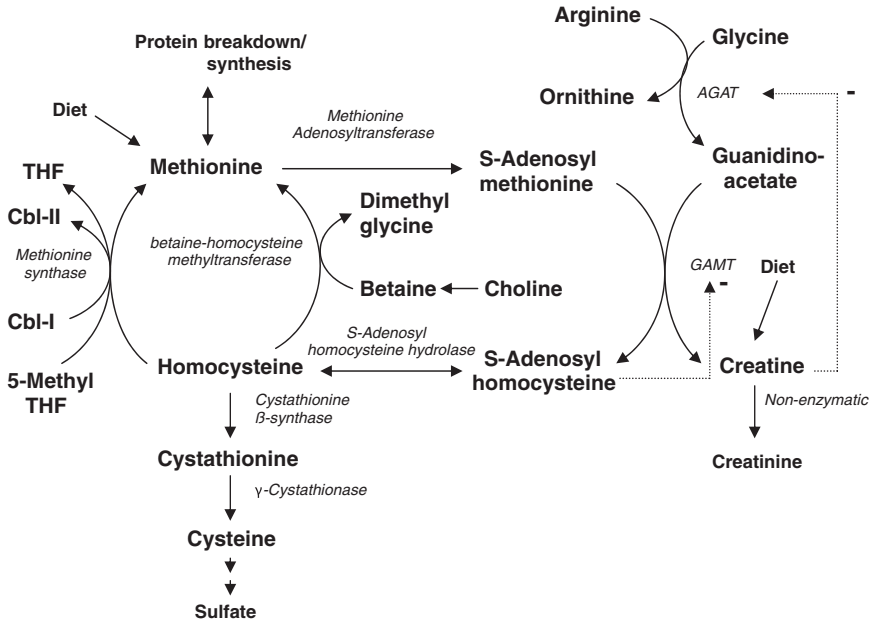


Fig. 2. Methionine–homocysteine metabolism as an example for the complexity of intermediary metabolism. Most metabolites in the depicted pathways can be quantified by either GC–MS or ESI–MS/MS (AGAT: arginine–guanidinoacetate amidinotransferase; GAMT: guanidionacetate methyltransferase).

3.2. Analysis of organic acids including orotic acid

Organic acidemias, also known as organic acidurias, are a group of disorders characterized by increased excretion of organic acids in urine. They result primarily from deficiencies of specific enzymes in the breakdown pathways of amino acids or from enzyme deficiencies in β-oxidation of fatty acids or carbohydrate metabolism. Organic acidemias can be classified into five categories including branched-chain organic acidemias, multiple carboxylase deficiency, including holocarboxylase synthetase deficiency and biotinidase deficiency, glutaric aciduria type I and related organic acidemias, fatty acid oxidation defects, and disorders of energy metabolism. For example, the diagnosis of methylmalonic aciduria (MMA) is made by measurement of organic acids in the urine using GC–MS. In MMA large amounts of methylmalonic acid, as well as methylcitrate, propionic acid, and 3-OH propionic acid, are present [14,15].

3.3. Analysis of oligosaccharides

The application of ESI–MS/MS allows the identification and quantification of individual oligosaccharides for the diagnosis of glycoproteinoses (oligosaccharidurias)

such as G_{M1} gangliosidosis, G_{M2} gangliosidosis, sialic acid storage disorder, sialidase/neuraminidase deficiency, galactosialidosis, I-cell disease, fucosidosis, and Pompe and Gaucher diseases [16]. Recent work demonstrated the feasibility of this approach using 1-phenyl-3-methyl-5-pyrazolone derivatization and MS/MS precursor scan of *m/z* 175 in positive ion mode [17]. This method has been adapted to high-throughput use allowing the application to management follow-up and eventually newborn screening for this group of disorders [17].

3.4. Analysis of lysosomal enzyme activities

Similarly, a direct multiplex assay of lysosomal enzymes in dried blood spots has been developed for newborn screening [18]. This approach is based on the incubation of dried blood spots at 37°C overnight with the appropriate substrates and stable isotopically labeled internal standards. If the enzyme was fully active, substrate was converted completely to the corresponding product which was quantified based on its relationship to the known concentration of the internal standard. Importantly, samples without dried blood spots ("blank") have to be used to adjust for background noise. Corresponding enzyme activities were calculated based on the assumption that 10 μl of extraction solution contained 0.98 μl of blood [18].

4. Assessment of *in vivo* metabolism using stable isotope techniques

Stable isotopes are non-radioactive atoms of the same chemical element, which differ only in their number of neutrons [19]. Many elements also have radioactive (non-stable) isotopes. The most commonly used stable isotopes in studies of macronutrient metabolism are ^{2}H (D or deuterium), ^{13}C, ^{15}N, and ^{18}O, while ^{25}Mg, ^{26}Mg, ^{42}Ca, ^{46}Ca, ^{48}Ca, ^{57}Fe, ^{58}Fe, ^{67}Zn, and ^{70}Zn are the most commonly used stable isotopes for studies of mineral metabolism. The most commonly used radioactive isotopes are ^{14}C and ^{3}H (tritium) [19]. More than 6000 stable isotope-labeled compounds (tracers) are commercially available for use in metabolic studies. Examples for some of these tracers are [1-^{13}C] leucine, [1-^{13}C, ^{15}N] leucine, [ring-$^{2}H_5$] phenylalanine, and [6,6]-D_2 glucose. It is currently accepted that these compounds have negligible biological side-effects, which renders them ethically acceptable for use in children [20].

Following intravascular or oral application, the tracer is metabolically indistinguishable from the equivalent unlabeled compound of interest (tracee). The metabolic fate of the compound can be assessed qualitatively and quantitatively by measuring the relative abundance of tracer and tracee and/or their respective breakdown products with time. The detectable mass difference of tracer and tracee allows the analysis of compounds extracted from plasma by either GC–MS or LC–MS [21]. Both require nanogram or picogram sample size (analytical range is 0.1–100 mol%, precision ±0.2%). The detection limit is considerably less than

0.1 mol%, when tracers with multiple stable isotope labels (for example, ring-D_5 phenylalanine) are used [22]. Stable isotopes in breath (i.e., $^{12}CO_2$ and $^{13}CO_2$) are analyzed using an isotope-ratio mass spectrometer (IRMS, microgram sample size, analytical range 0.001–10 atom% excess, precision ±0.00005 atom% (5 ppm)) [19]. Combustion-IRMS essentially has the same analytical capabilities as IRMS but allows the combustion of tissue samples with subsequent analysis of gaseous isotope enrichment [19]. Stable isotopes of minerals are typically analyzed by thermal ionization mass spectrometry (TIMS) or inductively coupled plasma mass spectrometry (ICP-MS) with high precision and sensitivity [19].

The advantages of stable isotope-labeled compounds compared with their radioactive counterparts are manifold. Most importantly, several different stable isotope tracers can be safely administered simultaneously to the same subject without limiting future studies. The plasma volume which is needed for one study to analyze isotope enrichment is small, allowing even pre-term infants to be studied. On average 0.5 ml of plasma is needed for one study. The intramolecular location of one or more label(s) is determined easily, which allows the assessment of metabolic pathways [19].

Stable isotopes are ideal "tools" for the dynamic assessment of *in vivo* metabolism in the pediatric population. Not only are these tracers safe and therefore ethically justifiable, in addition these "tools" may be particularly important for the validation of new treatment modalities, such as novel drug treatment or gene therapy. Many more exciting studies are currently under way to enhance our knowledge of pediatric metabolism and (patho-)physiology, an important factor for the continued reduction of pediatric morbidity and mortality [7].

5. Conclusions

MS is an essential analytical technique in the medical laboratory that provides a unique array of diagnostic opportunities. However, for many applications MS has to be combined with pre-analytical chromatography to allow optimal separation of metabolites and requires an in-depth knowledge of basic chemistry. Day-to-day operation should therefore be the responsibility of a dedicated chemist that works in close collaboration with laboratory physician and clinicians in a multidisciplinary team.

References

Selected Web sites:

www.cdc.gov: Centers of Disease Control, Atlanta, USA;
www.erndimqa.nl: Quantitative External Quality Assessment Schemes European Research Network;
www.eurogentest.org: EuroGenTest—Genetic Testing in Europe;
www.orpha.net: OrphaNet—rare disease database.

1. Leeming, D.J., Alexandersen, P., Karsdal, M.A., Qvist, P., Schaller, S. and Tanko, L.B., An update on biomarkers of bone turnover and their utility in biomedical research and clinical practice. *Eur. J. Clin. Pharmacol*, **62**, (**August**) 781–792 (2006).
2. Trof, R.J., Di Maggio, F., Leemreis, J. and Groeneveld, A.B., Biomarkers of acute renal injury and renal failure. *Shock*, **26**, 245–253 (2006).
3. Yuan, G., Wang, J. and Hegele, R.A., Heterozygous familial hypercholesterolemia: an under-recognized cause of early cardiovascular disease. *CMAJ*, **174**, 1124–1229 (2006).
4. Marsden, D., Larson, C. and Levy, H.L., Newborn screening for metabolic disorders. *J. Pediatr.*, **148**, 577–584 (2006).
5. Schulze, A., Lindner, M., Kohlmüller, D., Olgemöller, K., Mayatepek, E. and Hoffmann, G.F., Expanded newborn screening for inborn errors of metabolism by electrospray ionization–tandem mass spectrometry: results, outcome, and implications. *Pediatrics*, **111**, 1399–1406 (2003).
6. Wilcken, B., Wiley, V., Hammond, J. and Carpenter, K., Screening newborns for inborn errors of metabolism by tandem mass spectrometry. *N Engl. J. Med.*, **348**, 2304–2312 (2003).
7. Bodamer, O.A. and Halliday, D., Uses of stable isotopes in clinical diagnosis and research in the pediatric population. *Arch. Dis. Child.*, **84**, 444–448 (2001).
8. Plebani, M., Errors in clinical laboratories or errors in laboratory medicine? *Clin. Chem. Lab. Med.*, **44**, 750–759 (2006).
9. Burnett, D., ISO 15189:2003—quality management, evaluation and continual improvement. *Clin. Chem. Lab. Med.*, **44**, 733–739 (2006).
10. Childs, B., Valle, D. and Jiminez-Sanchez, G., The inborn error and biochemical individuality. In: Scriver, C., Beaudet, A., Valle, D. and Sly, W.S. (Eds.), *The Metabolic and Molecular Bases of Inherited Disease* (8th edition). McGraw-Hill, New York. 2001, p. 155 (Chapter 3).
11. Bodamer, O.A. and Mühl, A., Analysis of acylcarnitine ester for the diagnosis of inborn errors of metabolism using tandem mass-spectrometry. *Chem. Month.*, **136**, 1293–1297 (2005).
12. Roschinger, W., Olgemoller, B., Fingerhut, R., Liebl, B. and Roscher, A.A., Advances in analytical mass spectrometry to improve screening for inherited metabolic diseases. *Eur. J. Pediatr.*, **162**, S67–S76 (2003).
13. Tuschl, K., Bodamer, O.A., Erwa, W. and Muhl, A., Novel method for rapid analysis of total plasma homocysteine by tandem mass spectrometry. *Clin. Chim. Acta*, **351**, 139–141 (2005).
14. Fenton, W.A., Gravel, R.A. and Rosenblatt, D.S., Disorders of propionate and methylmalonate metabolism. In: Scriver, C.R., Beaudet, A.L., Sly, W.S. and Valle, D. (Eds.), *The Metabolic and Molecular Bases of Inherited Disease* (8th edition). McGraw-Hill, New York, 2001, p. 2165.
15. Deodato, F., Boenzi, S., Santorelli, F.M. and Dionisi-Vici, C., Methylmalonic and propionic aciduria. *Am. J. Med. Genet. C Semin. Med. Genet.*, **142**, 104–112 (2006).
16. Young, S.P., Stevens, R.D., An, Y., Chen, Y.T. and Millington, D.S., Analysis of a glucose tetrasaccharide elevated in Pompe disease by stable isotope dilution-electrospray ionization tandem mass spectrometry. *Anal. Biochem.*, **316**, 175–180 (2003).
17. Ramsay, S.L., Meikle, P.J., Hopwood, J.J. and Clements, P.R., Profiling oligo-saccharidurias by electrospray tandem mass spectrometry: quantifying reducing oligosaccharides. *Anal. Biochem.*, **345**, 30–46 (2005).
18. Li, Y., Scott, C.R., Chamoles, N.A., Pinto, B.M., Turecek, F. and Gelb, M.H., Direct multiplex assay of lysosomal enzymes in dried blood spots for newborn screening. *Clin. Chem.*, **50**, 1785–1796 (2004).
19. Wolfe, R.R., Basic characteristics of isotopic tracers. In: Wolfe, R.R. (Ed.), *Radioactive and Stable Isotope Tracers in Biomedicine. Principles and Practice of Kinetic Analysis*. Wiley-Liss, New York, 1992, pp. 1–17.

20. Jones, P.J.H. and Leatherdale, S.T., Stable isotope in clinical research: safety reaffirmed. *Clin. Sci.*, **80**, 277–280 (1991).
21. Patterson, B.W., Zhao, G. and Klein, S., Improved accuracy and precision of gas chromatograph/mass spectrometry measurements for metabolic tracers. *Metabolism*, **47**, 706–712 (1998).
22. Bier, D., Stable isotopes in biosciences, their measurement and models for amino acid metabolism. *Eur. J. Pediatr.*, **156**(**Suppl. 1**), S2–S8 (1997).
23. Fauler, G., Rechberger, G.N., Devrnja, D., Erwa, W., Plecko, B., Kotanko, P., Breunig, F. and Paschke, E., Rapid determination of urinary globotriaosylceramide isoform profiles by electrospray ionization mass spectrometry using stearoyl-d35-globotriaosylceramide as internal standard. *Rapid Commun. Mass Spectrom.*, **19**, 1499–1506 (2005).
24. Bodamer, O.A., Bloesch, S.M., Gregg, A., Stockler-Ipsiroglu, S. and O'Brien, W.E., Analysis of guanidinoacetate and creatine by isotope dilution electrospray tandem mass-spectrometry. *Clin. Chim. Acta*, **308**, 173–178 (2001).
25. Moebius, F.F., Fitzky, B.U. and Glossmann, H., Genetic defects in postsqualene cholesterol biosynthesis. *Trends Endocrinol. Metab.*, **11**, 106–114 (2000).
26. Johnson, D.W., ten Brink, H.J., Schuit, C. and Jakobs, C., Rapid and quantitative analysis of unconjugated C(27) bile acids in plasma and blood samples by tandem mass spectrometry. *J Lipid Res.*, **42**, 9–16 (2001).

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 13

Therapeutic drug monitoring and measurement of drug concentrations using mass spectrometry

ANDRÁS TELEKES[a,*], MÁRTA HEGEDŰS[a,1], and ISTVÁN KISS[b,2]

[a]*National Institute of Oncology, Budapest, Hungary*
[b]*St. Imre Government Teaching Hospital, Budapest, Hungary*

1. Introduction

Measurement of drug concentration is an inherent part of preclinical and clinical investigation of new therapeutic agents since no pharmacokinetic studies can be carried out without it. It is also necessary for investigating drug–effect or drug–toxicity relationship. Moreover, measurement of drug concentrations may help in understanding the mechanism of action as well. Most drugs in clinical use have well-defined pharmacodynamic profiles, i.e., there is a direct relationship between serum concentration and pharmacological response. These aspects are well known and are widely used in drug development.

*Corresponding author. Tel.: +36-204-119080; Fax: +36-139-52835.
E-mail: andras.telekes@gmail.com (A. Telekes).
[1]E-mail: hegedus.marta@gmail.com (M. Hegedűs).
[2]E-mail: ikiss@enternet.hu (I. Kiss).

To exert any biological/therapeutic efficacy, molecules should reach appropriate concentration at the target organ(s). Thus, they should travel from their application site to their receptors, which involves absorption through the skin or from the gastrointestinal system and distribution via the vascular system and passing through biological membranes. The journey of a drug continues even after it express its effect, as it must still leave the body. Moreover, during their movement drugs are metabolized. Pharmacokinetics is what the organism does to the drug; pharmacodynamics is what the drug does to the body. Thus, pharmacokinetics deals with the principles of absorption, distribution, metabolism, and elimination of drugs. Pharmacodynamics deals with the mechanism of action and biological activity of drugs and drug-induced clinical outcomes. Pharmacodynamics and pharmacokinetics applied in clinical settings, including patients or healthy volunteers, are called clinical pharmacology, which attempts to explain and predict the reasons of variability of drug action. Clinical pharmacological investigation of a drug cannot be carried out without measuring drug concentrations in different compartments of the body (e.g., blood, urine, feces, saliva, mother's milk, etc.). Measuring interpatient variability in drug kinetics can lead to implementation of strategies to decrease variability and thus achieve more consistent clinical outcomes.

Proper dosing of drugs is complicated by various factors. One of the main sources of interpatient pharmacokinetic variability is that different persons could metabolize a particular drug differently. Note, some drugs have high variability, which is due to not only several factors such as altered absorption, genetic polymorphism, pharmacological interactions, poor aqueous solubility but also a high metabolism mediated by the CYP450 system or presystemic first-pass effect with the involvement of transporters, such as P-glycoprotein.

Therapeutic drug monitoring (TDM) is the measurement of drug concentration, usually in plasma or in serum, for individual patients, to help develop and control proper dosage. It is usually expensive and complicated, so used only in a small fraction of clinical treatments. Most significant indications of TDM are shown in Table 1. TDM is often performed by chromatographic analysis (HPLC), but mass spectrometry (MS) is increasingly used and is becoming the prime analytical tool in this field. It may be used either alone (usually tandem mass spectrometry, MS–MS) or in combination with chromatography (HPLC–MS). The use of MS is becoming widespread, as it has much higher sensitivity than conventional HPLC and typically requires simpler sample preparation. Moreover, MS technique provides the possibility of simultaneous measurement of parent compounds and metabolites.

The therapeutic and toxic blood concentrations of several hundred drugs and xenobiotics have recently been published [1]. This may help in determining critical cases of narrow therapeutic index. The rationale of TDM was proven first for phenytoin treatment of epileptic patients. In other words, it was shown that the side effects were reduced and the seizure control was improved if instead of applying standard doses (based on the body weight of the patients) the dosages were

Table 1
Indications for therapeutic drug monitoring

The indication of therapeutic drug monitoring
Narrow therapeutic index: when the ratio between therapeutic and toxic doses is small
Organ deficiency: in the case of reduced renal excretion, decreased hepatic metabolism, heart failure leading to decreased clearance
Extremes of age: in childhood due to the lability of metabolism and variability of extra- and intracellular fluid spaces; in elderly because of age-related pharmacodynamic and pharmacokinetic changes
High interpatient pharmacokinetic variability
Polypharmacy: concurrent use of many medications increase the chance for pharmacokinetic interactions
Suspected noncompliance: for example, in case of inefficacy, acute overdose, and chronic abuse

adjusted according to blood concentrations [2]. This underlies the principle that concentration–response relationships are usually less variable than dose–response relationships for any drug. TDM traditionally used to control the management of epilepsy, asthma, depression, cardiac arrhythmias, or antibiotic treatment. Recently, it was proven that appropriate plasma concentrations of antiretroviral drugs are necessary to achieve and maintain the suppression of HIV replication [3]. MS-based methodology for therapeutic plasma monitoring of antiretroviral drug concentrations was also developed [4].

Drug monitoring is used not only to check therapeutic and/or toxic blood levels but also to determine absorption, rate of metabolism, excretion, or interaction with concomitantly applied drug(s). The success of TDM depends not only on the appropriate clinical indication but also on the timing and collection of the sample, the quality of the analysis, and proper interpretation. All of these lead to appropriate interventions, thus improving the clinical outcome such as enhanced efficacy, reduced adverse effects, or decreased time for resolution [5]. In fact, by applying TDM one can achieve individualized drug therapy (e.g., adjustment of dosage based on individual metabolism). TDM is also suitable to detect noncompliance of the patients. During early drug development, prospective (preplanned) concentration–clinical response measurements seem to provide a better background for TDM than retrospective concentration–effect analysis, in terms of timing blood withdrawal, or using other compartments for sampling (e.g., urine) [6]. Moreover, proper TDM may reduce the overall cost of patient care [7].

TDM requires appropriate and bias-free analytical methodology for the measurement of parent drugs or metabolites. Sophisticated technology, however, does not automatically guarantee accuracy; moreover, sample preparation can still have significant effect on the results, regardless of the technique applied [8]. It is important to emphasize that the method of measurement of the sample could significantly influence the value of pharmacokinetic parameters [9]. Thus, a statistical procedure for assessing

concordance between two methods of clinical measurement (e.g., compare the results of different techniques and sample preparation used) has been developed [10].

For reasons discussed above, measurement of drug concentrations and/or monitoring of prescribed and nonprescribed drug use provide a useful tool for optimally managing patients. In the following text, clinical applications of MS-based drug monitoring and kinetic measurements are discussed. These partly illustrate the range of MS applications in daily clinical practice, as well as provide some examples for clinicians about the sensitivity of MS when measuring drug concentrations. Note that data obtained from kinetic studies as well as the methodology itself might be used for reasoning and applying TDM. In fact, drug prescriptions contain information about the necessity of TDM if pharmacokinetic or pharmacodynamic data provide indications to do so.

In this chapter, various applications of measuring drug concentrations in a clinical environment are discussed, where MS is the prime analytical tool. Most examples relate to recently introduced drugs. Some are clearly TDM, some are clinical applications that may be used for TDM in the future, while some others are selected applications illustrating the range in which MS is used for measuring

Table 2
Alphabetical list of drugs discussed in the present chapter

Drugs studied by TDM	Potential candidates for TDM	Selected studies of drug concentration measurements
Actinomycin-D, amitriptyline, amoxapine, amprenavir, apomorphine atazanavir, citalopram, clomipramine, desipramine, dothiepin, doxepin, efavirenz, fluoxetine, imipramine, indinavir, lopinavir, maprotiline, meropenem, mianserin, nelfinavir, nevirapine, norfluoxetine, nortriptyline, pergolide, paroxetine, procarbazine, ritonavir, sertraline, saquinavir, trimipramine, vincristine, zidovudine	Adefovir, amlodipine, busulfan, cefaclor, cefdinir, cefixime, citalopram, cyclosporine, fenofibric acid, hydrocodone, hydromorphone, lonafarnib, nalmefene, rosuvastatin voriconazole	Ajulemic acid, aliskiren, amantadine, azacitidine, carbovir, carvedilol, dextran, diclofenac, isosorbide, lamivudine, 2′-fluoro-5-methyl-beta-L-arabinofuranosyl uracil, flunixin, fulvestrant, gefitinib, ibuprofen, imatinib, indomethacin, ketoprofen, lumiracoxib, madol, mefenamic acid, metronidazole, midazolam, naproxen, oritavancin, oxyphenbutazone, phenazopyridine, phenylbutazone, piroxicam, pravastatin, salicylic acid, spiramycin, synthetic insulins, tenofovir, tolmetin, trimetazidine vitamins B5, D, and K

drug levels in clinical samples. The drugs discussed are listed in alphabetical order in Table 2, while a more detailed discussion is presented in the subsequent sections, where drugs are classified according to their applications.

Modern analytical techniques, and especially those based on chromatography and MS, make it possible to monitor drug concentrations accurately, fast, using only a very small amount of biological sample, and in high throughput. These advances made TDM possible, which can significantly improve therapy and prevent toxicity. This trend is likely to continue; more and more drugs will be monitored routinely in everyday clinical practice.

2. Antiinfection drugs

Antibacterial agents can be used empirically, either knowing the pathogen or as a prophylactic treatment. Duration of the treatment can be short (e.g., less than 24 h as perioperative prophylaxis) or could take several months (e.g., in the case of endocarditis or tuberculosis). Since there are several similarly effective antibiotics against many pathogens, selection of the treatment depends on various factors, such as pharmacokinetics, side effects, resistance profile, and cost of the drug. For most antimicrobial drugs, TDM is not necessary. In case of drugs associated with toxicity (e.g., nephrotoxicity, ototoxicity), usually there is a relationship established between drug concentration and severity of adverse events. TDM is used not only to prevent toxicity but also to guide patient-tailored dosing regimens and to assess tissue penetration [11]. Several such applications are summarized in Table 3.

Fungal infections are usually local (such as skin, nails, mouth cavity, urogenital tract), but systemic, life-threatening infections may also occur. Several antifungal

Table 3
Applications of MS in the measurement of drugs used against infections

Ref.	Drug	Comment
[13]	Cefdinir	An LC–MS–MS method to measure cefdinir in human plasma. Linear calibration curve in the concentration range 5–2000 ng/ml, quantification limit is 5 ng/ml. Intra- and interday standard deviations: $<4.3\%$. Accuracy over the whole concentration range vary between 97 and 107%. Test time: <3 min. Suitable for pharmacokinetic testing.
[14]	Cefixime	An LC–MS–MS method to determine cefixime in human plasma. Linear calibration curve was found between 0.05 and 8.0 μg/ml. Quantitative measurements were possible down to 0.05 μg/ml. Intra- and interday standard deviations are $<12.7\%$. Accuracy is better than 2% (relative error) for the whole linear calibration range. Test time: 3.2 min. The method was successfully used in pharmacokinetic tests.

(*continues*)

Table 3
Continued

Ref.	Drug	Comment
[15]	Cefaclor	An LC–MS–MS method to determine cefaclor in human plasma. Plasma samples are treated by precipitation (PPT) or solid-phase extraction (SPE). The LC column is a C18 phase; the detector is a triple quadrupole tandem mass spectrometer in positive electrospray ionization (ESI) mode. Quantitative measurements or the PPT method were possible down to 100 ng/ml. Intra- and interday standard deviations: $<12\%$. Accuracy: $>3\%$ (relative error) for the whole linear calibration range. Test time: 3.2 min. In the case of the SPE method, the quantitation was as low as 2 ng/ml. Precision and accuracy were 7 and 3%, respectively. The method was successfully used in pharmacokinetic tests of a cefaclor sustained-release formulation.
[16]	Meropenem	An LC–MS–MS and a HPLC–UV method to measure meropenem, a broad-spectrum carbapenem antibacterial agent in human plasma and urine, respectively. The aim is to optimize doses in terms of plasma levels and pharmacokinetic behavior. The results were interpreted using a two-compartment open model. Two groups with intermittent and continuous infusion were compared, but no significant differences were found in total clearance and renal clearance. In case of certain infections, the intermittent therapy proved to be acceptable, but other, more tenacious bacteria needed high-dosage therapy.
[17]	Metronidazole, spiramycin I	An LC–MS–MS method for simultaneous determination of metronidazole and spiramycin I concentrations in human plasma, saliva, and gingival crevicular fluid (GCF), preceded by liquid–liquid extraction (LLE). Ornidazole is used as an internal standard. A C18 column is used with an eluent of acetonitrile, water, and formic acid. Intra- and interbatch precision: 7, 12, and 9% in plasma, saliva, and GCF, respectively. Accuracy was lowest in saliva (15.4%) and better in plasma (8.7%) or in GCF (10.7%). Linearity, specificity, recovery, matrix effect, dilution process, stability in human plasma and saliva after three freeze-thaw cycles, stability in human plasma and saliva at ambient temperature, and stability of the extracts in the automatic injector of the HPLC system have been studied. Useful for pharmacokinetic evaluations.
[18]	Oritavancin	An LC–MS–MS method has been utilized in a pharmacokinetic study aimed at oritavancin, a novel glycopeptide currently being developed for the treatment of complicated skin and skin structure infections (cSSSI), including those caused by multidrug-resistant gram-positive pathogens. The drug concentration was monitored in a cantharide-induced

Table 3
Continued

Ref.	Drug	Comment
		blister fluid model. Although the oritavancin level in the blister fluid was much (8–11 times) lower than in the plasma, it was still high enough to inhibit the proliferation of 90% of strains of *Staphylococcus aureus,* so it has a therapeutic potential.
[19]	Voriconazole	An LC–LC–MS–MS method for fully automated and direct analysis of voriconazole (a novel broad-spectrum antifungal agent) in raw human serum. The raw serum sample is first fractionated using a size-selective extraction column, followed by LC (C18 column) and ESI–MS–MS detection. Using parallel extraction and chromatographic separation, analysis time is 13 min. Lower quantification limit: 0.05 μg/ml. Eliminates the need for complicated sample pretreatment, and requires only 5 μl serum.
[20]	Protease inhibitor drugs as amprenavir, nelfinavir, indinavir, lopinavir, saquinavir, ritonavir, and atazanavir, and nonnucleoside reverse transcriptase inhibitors drugs nevirapine and efavirenz	A novel XLC–MS–MS (extraction liquid chromatographic + tandem mass spectrometric) technique for the simultaneous measurement of two samples from diluted human plasma samples for the monitoring of HIV/AIDS patient samples. Analysis time: 3.3 min; detection limit: 2–70 ng/ml; lower limit of quantification: 78–156 ng/ml. Good linearity is achieved in a wide concentration range (from the lower limit of quantification to 10,000 ng/ml). Intra- and interday precision values: 7.5–13.5% (depending on the concentration): accuracy and recovery: 86–113 and 60–110%, respectively. The method is useful in routine monitoring.
[21]	Lopinavir	An LC–MS method to measure the concentration of lopinavir (LPV) in cerebrospinal fluid (CSF) samples of HIV patients. As LPV binds strongly to plasma proteins, it was not sure whether the concentration of the drug is enough to inhibit HIV replication. The method developed had a lower limit of quantification of 3.7 μg/l. In patients with typical plasma levels of LPV, the drug is detectable in the CSF at concentrations that exceed those needed to inhibit HIV replication. Despite being >98% bound to plasma proteins, LPV penetrates into the central nervous system and may contribute to the control of HIV in this potential reservoir.
[22]	Tenofovir diphosphate, carbovir triphosphate, lamivudine triphosphate	Nucleotide concentrations can be measured directly using LC–MS in evaluating the intracellular concentrations and pharmacokinetics of tenofovir diphosphate (TFV-DP), carbovir triphosphate (CBV-TP), and lamivudine triphosphate (3TC-TP). An intracellular drug interaction does not explain the suboptimal viral response in patients treated with the nucleoside-only regimen of TDF, ABC, and 3TC.
[23]	Zidovudine triphosphate (ZDV-TP)	An LC–MS–MS method determines molar ZDV directly, corresponding to the intra-hPBMC molar ZDV-TP

(*continues*)

Table 3
Continued

Ref.	Drug	Comment
		concentration. ZDV-TP concentrations were determined in femtomoles per million hPBMCs (fmol/10^6 human peripheral blood mononuclear cells). The method is accurate and precise within the 5–640 fmol/10^6 cells range with 10 million cells per sample analyzed. Inter- and intraday accuracy and precision values fell within 15% of nominal. The assay correlates well with previous ELISA results. The method has been applied successfully in therapeutic monitoring.
[24]	Amantadine	An LC–MS–MS method measures directly the concentration of amantadine (1-adamantylamine, used for treatment of influenza, hepatitis C, parkinsonism, and multiple sclerosis) without protein precipitation, centrifugation, extraction, and derivatization steps. Only 50 μl sample is needed. Internal standard is 1-(1-adamantyl)pyridinium bromide. The serum sample is diluted by water in a 96-well plate. The chromatographic separation is performed using an eluent of isocratic water/acetonitrile (60/40, v/v) with 5 g/l formic acid on a C8 column. Run time is 3 min. Electrospray atmospheric pressure ionization, positive ion, and selective reaction monitoring mode were used. Detection limit: 20 mg/l, linearity: 20–5000 mg/l, intraassay/interassay coefficient of variation: $<6\%/<8\%$; recovery: 99–101%.
[25]	Adefovir	An LC–MS–MS method to study the pharmacokinetic behavior of adefovir, an antihepatitis B virus drug. Following protein precipitation the sample is analyzed on a C18 column, using a triple-quadrupole tandem mass spectrometer as detector in the positive electrospray ionization mode and PMPA as the internal standard. The method is linear in the concentration range 0.25–100 ng/ml, with the lower limit of quantification 0.25 ng/ml. The intra- and interday relative standard deviation over the entire concentration range is $\leq 5.7\%$. The accuracy determined at three concentrations is within $\pm 2.5\%$ relative error. The method was successfully used in pharmacokinetic studies.

drugs have long elimination times; for example, amphotericin B can be detected in the body even 6 weeks after stopping the treatment [12]. Absorption of oral antifungal agents is also variable. Thus, measurement of blood concentration might help to individualize the dose and/or treatment schedule.

The viruses are cell parasites; they use the cell machinery for replication. Targeting the virus-specific enzymes is an attractive option, and may lead to significant improvement for the treatment of specific viruses. Since obtaining accurate viral diagnosis is difficult and time-consuming, the start of a specific

antiviral drug treatment is frequently delayed. Maintaining appropriate concentrations of antiviral agents may enhance therapeutic activity and reduce development of resistance.

3. Drugs acting on the central nervous system (CNS)

Epilepsy can be explained as communication disorder among nerve cells. The disintegration of the balance of the excitatory and inhibitory stimuli leads to the predominance of excitatory impulses; thus, epileptic fit could occur. In the brain the most important excitatory neurotransmitter is glutamate, while the major inhibitory neurotransmitter is GABA. Thus, increasing GABA or decreasing glutamate can suppress the incidence of fits. TDM is indicated in the treatment of epilepsy since the symptoms of inefficacy (uncontrolled disease) and toxicity can be similar. This is further complicated by the fact that compliance of epileptic patients is not always appropriate [26]. Moreover, epilepsy is one of the most frequently occurring neurological disorder affecting millions of patients worldwide.

Depression is beyond any doubt the major psychiatric disorder that could affect every fifth individual at least once during their lifetime. There are several theories to explain development of depression including the role of noradrenaline, serotonine, acetylcholine, and dysregulation of neurotransmission. Therefore, antidepressants have many different and well-defined mechanisms of action such as enhancement of neurotransmitter synthesis, inhibition of neurotransmitter reuptake, monoamino oxidase (MAO) inhibition, antagonism of the activity of presynaptic inhibitory receptors, or increase in the activity of postsynaptic receptors. The reason for TDM in this class of drugs is that the metabolism and elimination show wide interindividual variability; thus, when standard doses are applied the serum concentration is often out of the therapeutic range [27].

The most frequent movement disorders are Parkinson's disease and Huntington chorea. The characteristic features of the former are hypokinetic movements and rigor of the muscles, while of the latter are hyperkinetic movements and hypotension of the muscles. In case of Parkinson's disease, the dopaminerg tracts are damaged in nigrostriatal system, while in Huntington chorea GABAerg neurons are insufficiently functioning, acetylcholine synthesis decreased, dopamine level increased, and the activity of NMDA receptors are enhanced. Thus, influencing dopamine synthesis and/or metabolism is beneficial in Parkinson's disease, while substituting acetylcholine, increasing GABA, antagonizing dopamine, and blocking the activity of NMDA receptors are all therapeutic targets in Huntington chorea. Measuring the correlation of drug concentration and efficacy may improve the benefit from all of these therapies or help to individualize dosing (e.g., in elderly).

MS-based methodology used in the measurement of CNS-acting drugs are shown in Table 4.

Table 4
Applications of MS in the measurement of drugs acting on CNS

Ref.	Drug	Comment
[28]	Amoxapine, amitriptyline, citalopram, clomipramine, dothiepin, doxepin, fluoxetine, imipramine, maprotiline, mianserin, paroxetine, sertraline, trimipramine (and some of their respective active metabolites: nortriptyline, monodesmethyl citalopram, desmethylclomipramine, desipramine, norfluoxetine, desmethyl mianserin, *N*-desmethyl sertraline)	A special turbulent-flow liquid chromatographic (TFC) technology, coupled with MS–MS to monitor 13 antidepressants and some of their active metabolites in human serum. Such tests are necessary if the drug either is toxic in high concentration or appears ineffective in therapy. Owing to their different chromatographic behavior, the antidepressants are divided into two separate groups (two parallels should be injected to cover the whole range of compounds). Calibration curves have been established for the concentration range of 10–500 ng/ml. No memory effect was observed even after the highest concentration samples. Intraassay and interassay precisions: 0.4–12 and 1–16%, respectively.
[29]	Citalopram (CIT), desmethylcitalo-pram (DCIT)	A GC–MS technique to elucidate the effect of aging on the steady-state plasma concentrations of citalopram (CIT) and desmethylcitalopram (DCIT). One hundred and twenty-eight depressive patients were treated with 10–80 mg/ day CIT. Patients were divided into three age groups ($<$64 years, 65–79 years, and $>$80 years). Despite comparable body mass indices (BMI) and renal and hepatic functions, plasma levels of CIT and DCIT exhibited large variations (16-fold and 12-fold, respectively). When compared to adults, mean plasma concentration of CIT and DCIT was 48% in the oldest age group and 33% higher in the elderly group, which has to be taken into account in their treatment, the dose should be reduced.
[30]	Midazolam	An HPLC–ESI–MS method simultaneously quantifies midazolam (MDZ) and its major metabolite 1′-hydroxymidazolam (1′-OHM) in a small volume (200 μl) of human plasma. Midazolam, 1′-OHM, and 1′-chlordiazepoxide (internal standard) are extracted from plasma samples using liquid–liquid extraction with 1-chlorobutane. The chromatographic separation is performed on a C18 column using as mobile phase water–acetonitrile, 75:25% (v/v), containing formic acid (0.1%, v/v). Protonated molecular ions were detected in the positive-ion mode. Calibration curves are linear ($r^2 \geq 0.99$) from 15 to 600 ng/ml (MDZ) and 5 to 200 ng/ml (1′-OHM). Limits of detection and quantification: 2 and 5 ng/ml, respectively, for both MDZ and 1′-OHM. Mean relative recoveries at 40 and 600 ng/ml (MDZ): 79.4 and 84.2, respectively; for 1′-OHM at 30 and 200 ng/ml the values were 89.9 and 86.9, respectively. The intraassay and interassay coefficients of variation (CVs) for MDZ were less than 8%, and for 1′-OHM less than 13%.

Table 4
Continued

Ref.	Drug	Comment
		There was no interference from other commonly used antimalarials, antipyretic drugs, and antibiotics. The method was successfully applied in a pharmacokinetic study.
[31]	Pergolide	An LC–MS–MS technique for pharmacokinetic purposes to monitor drug levels in patients with mild-to-moderate Parkinson's disease treated orally by pergolide. Plasma levels were correlated with the efficacy of the treatment. Steady-state pharmacokinetic profiles and motor score were determined on 14 patients in this dose-escalating study. Typical absorption times: 2–3 h, elimination half-life: ±21 h. The fast absorption and slow elimination presumably help in reducing motor problems in patients with Parkinson's disease.
[32]	Clozapine	Chromatographic (LC–MS) and solid-phase extraction (SPE) conditions have been optimized for Clozapine, with cycle times of 2.2 min. Depending on the ionization modes detection limits varied between 0.15 and 0.3 mg/ml. A quadratic calibration curve was found for clozapine and its N-oxide and a linear one for the desmethyl metabolite ($R > 0.99$ in all cases). Accuracy is better than 10% in the whole therapeutic concentration range. Interassay precision: 5–20% of the standard deviation from the highest to the lowest therapeutic concentrations. Quantitative measurements are possible down to 350 ng/ml.

4. Cardiovascular drugs

Cardiovascular drugs are the common name of compounds used to treat different heart disorders (such as congestive heart failure, angina, or arrhythmia) or diseases of the vascular system (e.g., hypertension). Heart failure can be acute (sudden left-ventricular insufficiency and as a consequence lung failure without hypertrophy of heart muscle), compensatory (no lung failure but hypertrophy of heart muscle), or exhaustive (no more compensation of the heart muscle even though there is hypertrophy). Heart failure can be influenced by different classes of drugs, including nitrates, Ca^{2+} antagonists, β-blockers, digitalis, ACE inhibitors, phosphodiesterase inhibitors, etc.

There are several ion channels that could participate in the pathomechanism of arrhythmia such as Na^{+}, Ca^{2+}, K^{+}, and Cl^{-} channels. Thus, these channels represent therapeutic targets of the therapy. Hypertension can be treated by drugs with distinctly different mechanism of actions. There are first-line (e.g., diuretics,

β-receptor antagonists, calcium antagonists, ACE inhibitors, angiotensin II receptor antagonists, α_1-receptor antagonists) and second-line drugs (e.g., α_2-receptor antagonists, angiotensin I receptor antagonists, potassium channel activators, direct vasodilators). The therapeutic indices for many of the drugs applied are narrow, leading to complications [33]. Since all heart disease could lead to acute cardiac death, maintaining the drug concentrations in proper therapeutic range might save the patient's life, especially if the life-threatening danger of significant overdosing is also considered with some of these drugs. MS also gained role in measurement of cardiac drugs as indicated in Table 5.

Table 5
Applications of MS in the measurement of cardiovascular drugs

Ref.	Drug	Comment
[34]	Amlodipine	An HPLC–MS–MS method to determine plasma levels of amlodipine. The results were utilized in bioequivalence tests of two tablets, wherein sex differences and tolerability were also investigated. The pharmacokinetic curves of all patients were within the ranges prescribed by the authorities, and both tablets were well tolerated by the patients. Bioavailability and pharmacodynamic differences between the sexes could be explained by body weight differences, and no significant differences appeared between the sexes in drug clearance.
[35]	Aliskiren	An LC–MS method to monitor the pharmacokinetic behavior of aliskiren (an orally effective renin inhibitor for the treatment of hypertension) and its interactions with lovastatin, atenolol, celecoxib, or cimetidine. Single doses of aliskiren showed no evidence of clinically important pharmacokinetic interactions with lovastatin, atenolol, celecoxib, or cimetidine.
[36]	Carvedilol	A GC–MS method to detect carvedilol and its metabolites in human urine. Before the liquid–liquid extraction of the analytes, urine samples are exposed to hydrolytic treatment by beta-glucuronidase/arylsulfatase. Trimethylsilyl derivatives are produced using *N*-methyl-*N*-trimethylsilyltrifluoroacetamide (MSTFA). Linear calibration curves are obtained between 3.0 and 75 ng/ml; the recovery rates of the various compounds from urine are between 80 and 98%. Detection limits: 0.75–3.0 ng/ml; intraday reproducibilities: 1.86–11.5%; interday values: 0.70–1.71%. The method can be used routinely.
[37]	Isosorbide 5-mononitrate (5-ISMN)	An LC–MS–MS technique coupled with photospray ionization and preceded by liquid–liquid extraction to determine plasma levels of isosorbide 5-mononitrate (5-ISMN), an organic nitrate vasodilatant used to alleviate the pains in angina pectoris. The analyte is extracted from 0.5 ml human plasma, followed by a chromatographic

Table 5
Continued

Ref.	Drug	Comment
		separation on a C8 column, and a typical test time is 2 min. A linear calibration curve with $R^2 > 0.995$ is obtained in the concentration range 20–2000 ng/ml. Interrun precision was 5–7% of the standard deviation, while interrun accuracy was better than 90%. The test has been utilized in bioequivalence studies.
[38]	Pravastatin	An HPLC–MS–MS method used in a comparative bioavailability test on two formulations of pravastatin. The drug was detected from human plasma with a lowest detection limit of 0.40 ng/ml. In addition to a general linear model, gender-related effects were also investigated. Bioequivalence was established by both models and gender differences could be explained by body weight differences.
[39]	Rosuvastatin, fenofibric acid	An LC–MS method to simultaneously determine rosuvastatin (RST) and fenofibric acid (FFA) in human plasma, using carbamazepine internal standard. The analytes are first extracted (LLE) into ethyl acetate. After evaporating the solvent, the residue is dissolved in a mobile phase consisting of 0.05 M formic acid:acetonitrile (45:55, v/v) and injected onto a C18 column. The MS–MS system is operated under the multiple reaction-monitoring mode (MRM) using EI and positive ion detection mode. Absolute recovery of RST, FFA, and IS was 74, 61, and 69%, respectively. The lower limits of quantification (LLOQ) of RST and FFA were 1.00 and 0.50 μg/ml, respectively. Response function was established for the range of concentrations 1.00–50.0 and 0.50–20.0 μg/ml for RST and FFA, respectively, with r^2 of 0.999 for both compounds. The inter- and intraday precision values for RST were in the range 8.93–9.37% relative standard deviation (RSD) and 1.74–16.1% RSD, respectively. Similarly, the inter- and intraday precision values in the measurement of FFA were in the range 9.78–11.6% RSD and 0.22–17.4% RSD, respectively. Accuracy values for RST and FFA: 88.1–108 and 87–115%, respectively. RST and FFA proved to be stable in the standard tests. Has been applied in a clinical study.
[40]	Trimetazidine	An LC–MS method to determine plasma levels of trimetazidine using an internal standard [1-(2,4,5-trimethoxybenzyl) piperazine]. Proteins are precipitated with trifluoroacetic acid; the neutralized supernatant is separated on a C(8) column with methanol aqueous 0.11% triethylamine adjusted to pH 3.3 with formic acid (1:4, v/v). Test time is 8 min. An ion trap analyzer with an AP-CI interface is used for detection, in the selected reaction-monitoring mode. Lowest quantification limit: 1.5 ng/ml. Used in bioequivalence studies.

5. Anticancer agents

TDM in oncology was first applied for the measurement of methotrexate plasma concentrations following high-dose treatment, then it was gradually extended to the dose modifications in case of liver or kidney failure [41]. There are arguments in favor of and against TDM in antineoplastic treatment. Anticancer agents are usually applied in maximal tolerable doses even though their pharmacokinetics are variable; thus, TDM could improve safety (by preventing overdose) and efficacy (providing maximum treatment intensity and preventing underdose). The limitation is, however, that blood concentration cannot reliably indicate the concentration achieved at the site of therapeutic target since the blood supply of tumors highly fluctuates, even within the same tumor tissue [42]. In case of anticancer treatments, biological features of the tumor (such as growing rate, specific growth factor receptors on tumor cell membranes, etc.) should also be considered. The difference between tumor and host tissue is sometimes slight; therefore, toxicity of chemotherapy could be severe. The mechanism of actions of chemotherapeutic drugs is different, including alkylating agents, antimetabolites, antimicrotubule agents, antitumor antibiotics, topoisomerase targeting drugs, anthracyclins, etc. Endocrine therapy of hormone-sensitive tumors such as in the case of breast cancer, prostate cancer, etc. should be implemented as standard part of care. Some new biological therapies such as inhibitors of tumor angiogenesis, proteosome inhibitors, growth factor receptor inhibitors, etc. might also be applied in monotherapy or combinations with chemotherapeutic agents. In Table 6, examples of MS-based measurements of anticancer agents are presented.

Table 6
Applications of MS in the measurement of anticancer drugs

Ref.	Drug	Comment
[43]	Actinomycin-D (Act-D), vincristine (VCR)	An LC–MS–MS method for the simultaneous quantitative determination of actinomycin-D (Act-D) and vincristine (VCR), which are cytotoxic agents commonly used in the treatment of pediatric cancers. Following solid-phase extraction, plasma samples are separated and analyzed using electrospray ionization (ESI). Lower limit of quantitation (LLOQ) for both Act-D and VCR: 0.5 ng/ml. Analytical accuracy for detection of both Act-D and VCR: ≤90%. Analytical precision, as estimated by the coefficient of variation: ≤6% for Act-D and ≤11% for VCR. Useful in clinical monitoring.
[44]	5-Azacitidine	LC–MS–MS was used to monitor the pharmacokinetic behavior of 5-azacitidine (5-AC), a cytidine nucleoside analog, when given with phenylbutyrate, a histone deacetylase inhibitor. Pharmacokinetic data were obtained from trials

Table 6
Continued

Ref.	Drug	Comment
		involving patients with solid tumor and hematologic malignancies. 5-AC at doses ranging from 10 to 75 mg/m^2 day was administered once daily as a subcutaneous injection for 5–21 days in combination with phenylbutyrate administered as a continuous intravenous infusion for varying dose and duration every 28 or 35 days. Despite a short terminal half-life of 1.5 ± 2.3 h, inhibition of DNA methyl transferase activity in tumors of patients receiving 5-AC has been documented. It can be concluded that 5-AC is rapidly absorbed and eliminated when administered subcutaneously. Sufficient 5-AC exposure is achieved to produce pharmacodynamic effects in tumors.
[45]	2′-Fluoro-5-methyl-beta-l-arabinofuranosyl uracil triphosphate (L-FMAU-TP)	Ion-pairing, reverse-phase, liquid chromatography/electrospray tandem mass spectrometry to determine the level of 2′-fluoro-5-methyl-beta-l-arabinofuranosyl uracil triphosphate (L-FMAU-TP) from human peripheral blood mononuclear cells of hepatitis B virus-infected patients treated with L-FMAU. Limit of detection: 1.6 pmol/10^6 human peripheral blood mononuclear cells. The calibration curve for L-FMAU-TP is linear over the concentration range 1.6–80 pmol/10^6 cells. Intra- and interday precision: <11.2%; accuracy: 97.1–106.9%. When applied to the determination of L-FMAU-TP in PBMCs isolated from HBV-infected patients undergoing L-FMAU treatment, the levels reached a steady state concentration 4 weeks after daily single oral administration of 20 mg L-FMAU, and these levels were maintained for up to 12 weeks, but then decreased 12 weeks after drug cessation. The terminal half-life of L-FMAU-TP in PBMCs after drug cessation was estimated to be 15.6 days.
[46]	Busulfan	Busulfan was determined quantitatively by LC–MS–MS in saliva and plasma in children after hematopoietic stem cell transplantation. Lowest limit of detection: 2 μg/l; lower limit of quantification: 10 μg/l. Only 100 μl of plasma/saliva was needed. The mean recoveries (SD) of busulfan: 97.2% (2.7) in plasma and 100.4% (1.3) in saliva. Intra- and interassay imprecision: 2–3 and 2–4% for plasma, and 1–2 and 2–4% for saliva (concentration range 30–1500 μg/l). The bias was <4% for both plasma and saliva. The correlation between the busulfan concentration in plasma and saliva was highly significant ($r = 0.958$; $p < 0.0001$; saliva/plasma ratio = 1.09 ± 0.04; $n = 69$ sample pairs). The apparent plasma clearance was slightly higher than the apparent saliva clearance (202 ± 31 ml/h/kg vs. 189 ± 28 ml/h/kg; $p = 0.001$). The mean elimination half-life is 2.31 ± 0.46 h for plasma

(*continues*)

Table 6
Continued

Ref.	Drug	Comment
		and 2.30 ± 0.36 h for saliva; these were not significantly different ($p = 0.83$). Analysis of busulfan in saliva could be a valuable and reliable alternative to plasma analysis.
[47]	Fulvestrant	HPLC–MS–MS was used in a pharmacokinetic study of a long-acting formulation of fulvestrant within two global phase III efficacy studies comparing intramuscular fulvestrant with oral anastrozole. Preliminary pharmacokinetic analysis suggests that observed single- and multiple-dose plasma profiles can be adequately described with a two-compartment kinetic model. The intramuscular formulation of fulvestrant displays predictable kinetics and approximately twofold accumulation on administration once monthly. At the proposed therapeutic dosage (250 mg once monthly), plasma fulvestrant concentrations are maintained within a narrow range throughout the administration interval, thus ensuring stable systemic drug exposure during long-term treatment.
[48]	Gefitinib	An LC–MS–MS method to measure the concentration of gefitinib in human plasma, mouse plasma, and tissue. The chromatographic separation was preceded by protein precipitation with acetonitrile. A deuterated analogue was used as internal standard. The sample was analyzed on a C18 column using isocratic flow and acetonitrile–water (70:30, v/v) mobile phase containing 0.1% formic acid. EI and MS–MS were used for detection. Linear calibration curves were found in a wide concentration range from 1–5 ng/ml (depending on the matrix) up to 1000 ng/ml with $R^2 > 0.99$. Intra- and interday precision and accuracy values were better than 15%. The method was successfully applied in animal and human pharmacokinetic studies using oral or intraperitoneal administration.
[49]	Imatinib	An LC–MS–MS method to monitor plasma levels of imatinib, a selective tyrosine kinase inhibitor used for the treatment of chronic myeloid leukemia (CML) and other malignant diseases. The analyte is extracted from plasma and a deuterated internal standard is added, prior to analysis on a C18 column using gradient elution with acetonitrile–ammonium formate buffer 4 mmol/l, pH 3.2. EI–MS in multiple reaction-monitoring mode is used for detection. Linear calibration curves are in the concentration range 10–5000 ng/ml. The limit of quantification was set at 10 ng/ml. Intra- and interday precisions: <8%. Extraction recovery: >90%. The method can be routinely used in pharmacokinetic and drug interaction studies.
[50]		An LC–MS–MS method used in a pharmacokinetic study where imatinib mesylate (Glivec) and its main metabolite (CGP74588) were measured in blood samples of a patient with end-stage renal disease on hemodialysis and compared with

Table 6
Continued

Ref.	Drug	Comment
		published data from subjects with normal renal function. Maximum concentrations, absorption rates, and half-life values determined for the drug and its metabolite were comparable with those obtained on patients with normal renal function. Thus, the standard dose of imatinib can be safely administered to patients on hemodialysis, and probably with renal failure, at any stage.
[51]	Lonafarnib	An LC–MS–MS method to monitor lonafarnib (a novel anticancer drug that inhibits farnesyl transferase) in human plasma. Deuterated internal standard is used; proteins are precipitated by acetonitrile. Reverse-phase chromatographic separation is performed using acetonitrile/water/formic acid (50:50:0.05, v/v/v) mobile phase. Time of analysis: 8 min. A triple quadrupole tandem mass spectrometer in the positive-ion mode with multiple reaction monitoring is used for detection. The calibration curve has been established in the 2.5–2500 ng/ml concentration range. The validated method was successfully used in phase I trials of the drug.
[52]	Procarbazine	A reverse-phase HPLC method with ESI–MS detection to characterize the pharmacokinetic behavior of procarbazine, a cytotoxic chemotherapeutic agent used in the treatment of lymphomas and brain tumors. The data are used in a phase I trial; concentrations are measured in human plasma. The calibration curve is linear in the 0.5–50 ng/ml concentration range. Average recovery rate: 102.9%. Lower limit of quantitation: 0.5 ng/ml; accuracy: 105.2%; interday precision: 3.6% RSD; sample volume: 150 μl. Interday precisions at widely different concentrations: 97–98%. The stability of the drug under storage and sample preparation conditions have also been thoroughly tested. Sensitivity is sufficient for monitoring plasma levels after oral administration.
[53]	Tamoxifen	An LC–MS–MS method to determine tamoxifen (tam) and its metabolites 4-hydroxytamoxifen (4OHtam), *N*-demethyltamoxifen (NDtam), *N*-dedimetvhyltamoxifen (NDDtam), tamoxifen-*N*-oxide (tamNox), and 4-hydroxy-*N*-demethyltamoxifen (4OHNDtam) human serum. Proteins are precipitated with acetonitrile. Deuterated tamoxifen (D5 tam) is added as internal standard. Sample supernatant is injected into an online reverse-phase extraction column coupled with a C18 analytical column and analytes are detected by tandem mass spectrometry. Lower limits of quantification: 0.25 ng/ml for 4OHtam, Ndtam, and tam, and 1.0 ng/ml for NDDtam and tamNox. Within- and between-day variation: 2.9–15.4 and 4.4–12.9%, respectively.

6. Analgesics

The most consumed drugs are analgesics. Basically there are two main categories of analgesics, namely nonsteroidal antiinflammatory drugs (NSAIDs) and opioids. Many active ingredients are in the market in several hundreds of brand name. NSAIDs are used to treat not only pain but also other medical problems as well, including fever, inflammation, and prevention of thrombosis. According to their chemical structure, NSAIDs can be classified into different classes such as anthranilic acid, indoleacetic acids, naphthylalkanone, oxicam, phenylacetic acid, propionic acid, pyrazolone, pyrrole acetic acid, salicylic acid, etc. Even within the same class there could be wide variations between the analgesic activities of different compounds. It is important to note that antiinflammatory doses of NSAIDs are usually higher than the doses required to achieve analgesia.

Opioids can be classified as antagonists, full agonists, and partial agonists. Agonists can be further divided into weak and strong opioids. The structure–activity relationships of opioids are well established; thus, synthetic strong opioids are available, including fentanyl and methadone. The indications for TDM are suspected dose-related toxicity, suspected noncompliance, acute overdose, chronic abuse, reduced kidney or liver function, potential interaction with other drugs, evaluation of absorption, and optimalization of treatment (in patients who are frail, elderly, obese, etc.) [54]. In Table 7, some examples of measuring analgesics by MS are listed.

Table 7
Applications of MS in the measurement of analgesics

Ref.	Drug	Comment
[55]	Ajulemic acid	A GC–MS method combined with solid-phase extraction to detect ajulemic acid (AJA), a nonpsychoactive synthetic cannabinoid in human plasma. The calibration curve exhibits two linear portions between 10 and 750 ng/ml, and 750 and 3000 ng/ml, respectively. Intra- and interday precision values (expressed as the percentage of the RSD value) for the two segments of the calibration curves: 1.5–7.0 and 3.6–7.9, respectively. Detection limit: 10 ng/ml. The amount of the glucoronide derivative could be estimated by comparing the free AJA levels with those obtained after enzymatic hydrolysis. The method was tested on 21 patients suffering from neuropathic pain with hyperalgesia and allodynia.
[56]	Apomorphine	The drug could be detected above concentrations of 0.010 ng/ml and quantification was possible above 0.025 ng/ml. Accuracy and precision tests were made by analyzing 54 quality-control samples for 3 days. The concentration range studied was between 0.075 and 15 ng/ml (logarithmic scale, 3 points). Intraday precision: 10.1–3.8%; interday

Table 7
Continued

Ref.	Drug	Comment
		precision: 4.8–6.6%. Accuracy values: 99.5–104.2%. Simple and convenient test method for therapeutic drug monitoring and pharmacokinetic studies, requiring only 0.5 ml sample volume.
[57]	Hydrocodone, hydromorphone	A rapid hyphenated LC assay method based on MS–MS detection for hydrocodone (HYC) and its metabolite hydromorphone (HYM) in human plasma. Sample handling, internal standard addition, and analyte extraction are performed on a 96-channel automatic workstation, analyte and internal standard extraction on a 96-well solid-phase extraction cartridge system. The LC column is based on silica; the mobile phase contains acetonitrile and water and is acidified by trifluoroacetic acid (TFA). The total time of analysis is 2.5 min. The singly charged precursor ions and the product ions were $300 \rightarrow 199$ (m/z) for HYC and $28 \rightarrow 185$ (m/z) for HYM. A validated calibration curve was established between 0.100 and 100 ng/ml, using 0.3 ml plasma. Correlation coefficients for both analytes: >0.999. Quantitative determination was possible from 0.100 ng/ml for both HYC and HYM, the signal-to-noise ratio was >50 for HYC and 10 for HYM. Interday precision: $>5\%$ standard deviation; interday precision $\pm 2\%$ (relative error) for both analytes. Intraday precision was about 2.5% (standard deviation) and intraday accuracy was better than 3% (relative error) for both compounds. Five times dilution of the QC sample did not result in a significant deviation of the nominal value. Samples remained stable after 24 h room temperature storage of three freeze-thaw cycles. Extraction yields: 86 and 78% for HYC and HYM, respectively. No carryover was detected using a blank after a highly concentrated test solution.
[58]	Lumiracoxib	This validated HPLC–MS technique has been utilized in a pharmacokinetic study of lumiracoxib, a cyclo-oxygenase-2 (COX-2) selective inhibitor in development for the treatment of rheumatoid arthritis, osteoarthritis, and acute pain. Levels of the drug and its metabolites (4′-hydroxy-lumiracoxib and 5-carboxy-4′-hydroxy-lumiracoxib) were determined in plasma and synovial fluid. The pharmacokinetic curves were interpreted by two independent methods. Both absorption (peak plasma concentration after 2 h) and decay (plasma half-life 6 h) are relatively fast. Lumiracoxib concentrations were first higher in the plasma, later in the synovial fluid (peak drug concentrations in the latter were about three times higher). Concentrations of 4′-hydroxy-lumiracoxib, the active COX-2 selective metabolite, remained low in comparison with parent drug in both plasma and synovial

(*continues*)

Table 7
Continued

Ref.	Drug	Comment
		fluid. The extended presence of the antiinflammatory drug in the synovial fluid extends its therapeutic action beyond that expected from plasma pharmacokinetics.
[59]	Nalmefene	A sensitive test method to detect nalmefene, an opioid antagonist in human and rabbit serum, using nalbuphine internal standard. The analyte is extracted using *n*-butyl chloride/acetonitrile (4:1). HPLC combined with electrospray and MS–MS detection is used for quantitative tests. First validated for human plasma, and cross-checked by rabbit serum and plasma. Specificities are 3.21, 5.55, and 3.62% for human plasma, rabbit plasma, and rabbit serum, respectively. Extraction yield: 80% from human plasma. A calibration curve has been established for the 0.1–100 ng/ml concentration range. Within one run, at the lower limit the accuracy was 18%; the precision was 13.6%. In the concentration range 0.3–75 ng/l the accuracy was better than 128%; the precision was 6.6% for all matrix types. Interrun accuracy and precision: 8.0 and 6.6%, respectively. The analyte remained stable after 24 h room temperature storage of three freeze-thaw cycles.
[60]	Phenazopyridine	An LC–MS method for phenazopyridine, a strong analgesic used in the treatment of urinary tract infections to support *in vivo* pharmacokinetic studies. In spite of its widespread use, no assay method previously available for measuring plasma concentrations in human after oral administration. The analyte is extracted using liquid–liquid extraction, followed by LC–MS using a C18 column and soft ionization mode. An unexpected peak-doubling was observed and a two-site absorption compartment model was developed to explain the observed phenomena. Concentration profiles could be well explained for various dosage groups using the model.
[61]	Phenylbutazone, indomethacin, flunixin, piroxicam, diclofenac, ketoprofen, mefenamic acid, oxyphenbutazone, ibuprofen, salicylic acid, tolmetin, naproxen	An LC–MS method with normal pressure CI using negative-ion mode to detect nonsteroidal antiinflammatories (NSAIDs) and acetaminophen (ACE). Ion chromatograms for each species have been identified using full-scan fragmentation spectrograms. Linear quantitative calibration curves were obtained for the concentration range of 0.05–25.0 μg/ml. Detection limits: 0.05–1.0 μg/ml. Matrix detection limits: 0.05 μg/ml for phenylbutazone (*m*/*z* 307); 0.1 μg/ml for indomethacin (*m*/*z* 312), flunixin (*m*/*z* 295), and piroxicam (*m*/*z* 330); 0.5 μg/ml for ACE (*m*/*z* 150), diclofenac (*m*/*z* 250), ketoprofen (*m*/*z* 209), and mefenamic acid (*m*/*z* 240); 1.0 μg/ml for oxyphenbutazone (*m*/*z* 323); 5.0 μg/ml for ibuprofen (*m*/*z* 205), salicylic acid (*m*/*z* 137), and tolmetin (*m*/*z* 212); and 10 μg/ml for naproxen (*m*/*z* 185).

Table 7
Continued

Ref.	Drug	Comment
[62]	Salicylic acid, acetaminophen, theophylline, barbiturates, bromvalerylurea	An electrospray LC–MS methodology for monitoring the serum levels of patients poisoned by salicylic acid, acetaminophen, theophylline, barbiturates, and bromvalerylurea. It is preceded by solid-state extraction, using *o*-acetamidophenol as internal standard. As these drugs cause acute poisoning relatively frequently, the test needs to be rapid with a high precision. For acetaminophen the positive-ion mode is utilized, for all others the negative-ion mode, the base ions are used for quantitative determination. Quantitative determination is possible from 100 μg/ml, the upper limit varies between 0.5 and 5 μg/ml. Lowest detection limit: 0.1–1 μg/ml using full-scan MS for identifying acute poisoning. Using Oasis HLB 1-cc solid-phase extraction cartridges the recovery rates are 89–96%. Intraday reproducibility values: 3.55–6.05%; interday values: 3.68–6.38%.

7. Miscellaneous drug classes

There are many categories of drugs that are used in clinical practice. In these cases, TDM is usually not required. However, measuring drug concentrations in blood or other body fluids (e.g., urine) might contribute to developing optimal therapeutic (dosing) strategies, predicting accumulation, and measuring elimination. MS can also be used to differentiate between natural and synthetic molecules within the body, such as steroids or insulin, to detect doping. In Table 8, results on miscellaneous drug classes are summarized, including immunosuppressives and vitamins.

Table 8
Applications of MS in the measurement of various drugs

Ref.	Drug	Comment
[63]	Cyclosporine	An HPLC–MS method used in a pharmacokinetic monitoring study of cyclosporine (CsA, an immunosuppressive drug), which has a narrow therapeutic window, immunosuppressive, and/or toxic metabolites and a wide range of metabolic rates between individuals, thus requiring great care in establishing individual doses. As the drug and its metabolites (AM1, AM1c, DihydroAM1, AM19, and AM4N) tend to bind to lipoproteins, protein precipitation and solid-phase extraction are necessary prior to reverse-phase chromatographic analysis. The drug and its metabolites (which exhibit patient-specific patterns) are detected by MS in the form of sodium adducts after EI. Hepatotoxic potential has been confirmed and strong correlation between AM19 and CRP and IL6 observed.

(*continues*)

Table 8
Continued

Ref.	Drug	Comment
[64]	Dextran	A method based on LC–MS–MS for doping control, allowing quantitative determination of the plasma volume expander dextran. The dextran polymer is enzymatically converted into disaccharides, such as lactose, saccharose, and isomaltose; the analyte is detected in human urine. After the basal concentration of isomaltose was established, the concentration of dextran was measured as isomaltose in urine specimens obtained from patients treated with dextran. Linear and reproducible calibration curves for dextran were obtained. Inter- and intraassay coefficients of variation: 4.9–7.3% between 53 and 1186 μg/ml concentration levels. Recovery scattered between 97 and 112%. Lower limit of detection: 3.8 μg/ml; lower limit of quantification: 12.5 μg/ml. The highest concentrations measured in control samples were more than 100-fold lower than those found in urine samples of patients after treatment with dextran.
[65]	Madol	A method for rapid screening of urine by GC–MS to measure the concentration of trimethylsilylated madol (17alpha-methyl-5alpha-androst-2-en-17beta-ol, an alleged anabolic steroid not covered by routine doping tests) by monitoring peaks at *m*/*z* 143, 270, and 345.
[66]	Synthetic insulins (Humalog Lispro, Novolog Aspart, Lantus Glargine)	Synthetic insulins such as Humalog Lispro, Novolog Aspart, or Lantus Glargine should be analyzed in doping control as they are sometimes misused for nontherapeutic purposes. Plasma specimens of 2 ml fortified with three synthetic insulin analogues are purified by immunoaffinity chromatography, and extracts analyzed by microbore LC and MS–MS. Product ion scan experiments of intact proteins enable the differentiation between endogenously produced insulin and its synthetic analogues by collisionally activated dissociation of multiply charged precursor ions. This allows the assignment of individual fragment ions, particularly those comprising modifications that originate from C-termini of B-chains. Recoveries of synthetic insulins from plasma aliquots: 91–98%; detection limit: 0.5 ng/ml for all target analytes.
[67]	Vitamin B5	An LC–MS method to measure the concentration of vitamin B5 in human urine. Hopantenic acid (HOPA) is used as internal standard. Quantitative MS detection is performed in the single-ion monitoring mode. A linear calibration curve was obtained with $R^2 = 0.999$ in the concentration range of 0.25–10 μg/ml. The lower limit of detection is 0.1 μg/ml, with an intraassay coefficient of variation $<5\%$ and recoveries between 96 and 108%.

Table 8
Continued

Ref.	Drug	Comment
[68]	Vitamin D (25-hydroxy)	An LC–MS–MS method with AP-CI to determine the concentration of 25-hydroxyvitamin D (25-OH-D(2)/-D(3)) in human plasma. A deuterated standard is used and the tandem spectrometer is in the multiple-reaction-monitoring mode. Intra- and interassay variations: 2–6%; recoveries: 104–99%. Potential applications are the evaluation of the vitamin D status in postmenopausal women and elderly subjects, the diagnosis of vitamin D insufficiency/deficiency, as well as for the treatment and prevention of osteoporosis.
[69]	Vitamin K	An LC–MS–MS method with AP-CI to determine the concentration of vitamin K and related compounds (phylloquinone (PK), menaquinone-4 (MK-4), and menaquinone-7 (MK-7)) in human plasma. The internal standard is an isotope-labeled compound (O18); detection is by MS–MS using multiple reaction monitoring. Intra and interassay variations: $<10\%$; recoveries: 98–102%. Potential applications are the evaluation of vitamin K status in postmenopausal women and elderly subjects and in the treatment and prevention of osteoporosis.

References

1. Schulz, M. and Schmoldt, A., Therapeutic and toxic blood concentrations of more than 800 drugs and other xenobiotics. *Pharmazie*, **58**(**7**), 447–474 (2003).
2. Feldman, R.G. and Pippenger, C.E., The relation of anticonvulsant drug levels to complete seizure control. *J. Clin. Pharmacol.*, **16**, 51–59 (1976).
3. Rendón, A.L., Núñez, M., Jiménez-Nácher, I., González de Requena, D., González-Lahoz, J. and Soriano, V., Clinical benefit of interventions driven by therapeutic drug monitoring. *HIV Med.*, **6**(**5**), 360–365 (2005).
4. Alexander, C.S., Montaner, J.S., Asselin, J.J., Ting, L., McNabb, K., Harris, M., Guillemi, S. and Harrigan, P.R., Simplification of therapeutic drug monitoring for twice-daily regimens of lopinavir/ritonavir for HIV infection. *Ther. Drug Monit.*, **26**(**5**), 516–523 (2004).
5. Schumacher, G.E. and Barr, J.T., Total testing process applied to therapeutic drug monitoring: impact on patients' outcomes and economics. *Clin. Chem.*, **44**(**2**), 370–374 (1998).
6. Shaw, L.M., Kaplan, B. and Brayman, K.L., Prospective investigations of concentration-clinical response for immunosuppressive drugs provide the scientific bases for therapeutic drug monitoring. *Clin. Chem.*, **44**(**2**), 381–387 (1998).
7. Bates, D.W., Soldin, S.J., Rainey, P.M. and Micelli, J.N., Strategies for physician education in therapeutic drug monitoring. *Clin. Chem.*, **44**(**2**), 401–407 (1998).
8. Bowers, L.D., Analytical goals in therapeutic drug monitoring. *Clin. Chem.*, **44**(**2**), 375–380 (1998).
9. Aspeslet, L.J., Legatt, D.F., Murphy, G. and Yatscoff, R.W., Effect of assay methodology on pharmacokinetic differences between cyclosporine Neoral® and Sandoimmune® formulations. *Clin. Chem.*, **43**, 104–108 (1997).

10. Bland, J.M. and Altman, D.G., Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307–310 (1986).
11. Hammett-Stabler, C.A. and Johns, T., Laboratory guidelines for monitoring of antimicrobial drugs. *Clin. Chem.*, **44**(**5**), 1129–1140 (1998).
12. Gallis, H.A., Drew, R.H. and Packard, W.W., Amphotericin B: 30 years of clinical experience. *Rev. Inf. Dis.*, **12**, 308–329 (1990).
13. Chen, Z.J., Zhang, J., Yu, J.C., Cao, G.Y., Wu, X.J. and Shi, Y.G., Selective method for the determination of cefdinir in human plasma using liquid chromatography electrospray ionization tandem mass spectrometry. *J. Chromatogr. B*, **834**(**1–2**), 163–169 (2006).
14. Meng, F., Chen, X., Zeng, Y. and Zhong, D., Sensitive liquid chromatography-tandem mass spectrometry method for the determination of cefixime in human plasma: application to a pharmacokinetic study. *J. Chromatogr. B*, **819**(**2**), 277–282 (2005).
15. Chen, X., Zhong, D., Huang, B. and Cui, J., Determination of cefaclor in human plasma by a sensitive and specific liquid chromatographic-tandem mass spectrometric method. *J. Chromatogr. B*, **784**(**1**), 17–24 (2003).
16. Krueger, W.A., Bulitta, J., Kinzig-Schippers, M., Landersdorfer, C., Holzgrabe, U., Naber, K.G., Drusano, G.R. and Sorgel, F., Evaluation by monte carlo simulation of the pharmacokinetics of two doses of meropenem administered intermittently or as a continuous infusion in healthy volunteers. *Antimicrob. Agents Chemother*., **49**(**5**), 1881–1889 (2005).
17. Sagan, C., Salvador, A., Dubreuil, D., Poulet, P.P., Duffaut. D. and Brumpt, I., Simultaneous determination of metronidazole and spiramycin I in human plasma, saliva and gingival crevicular fluid by LC-MS/MS. *J. Pharm. Biomed. Anal.*, **38**(**2**), 298–306 (2005).
18. Fetterly, G.J., Ong, C.M., Bhavnani, S.M., Loutit, J.S., Porter, S.B., Morello, L.G., Ambrose, P.G. and Nicolau, D.P., Pharmacokinetics of oritavancin in plasma and skin blister fluid following administration of a 200-milligram dose for 3 days or a single 800-milligram dose. *Antimicrob. Agents Chemother.*, **49**(**1**), 148–152 (2005).
19. Egle, H., Trittler, R., Konig, A. and Kummerer, K., Fast, fully automated analysis of voriconazole from serum by LC-LC-ESI-MS-MS with parallel column-switching technique. *Chromatogr. B*, **814**(**2**), 361–367 (2005).
20. Koal, T., Sibum, M., Koster, E., Resch, K. and Kaever, V., Direct and fast determination of antiretroviral drugs by automated online solid-phase extraction-liquid chromatography-tandem mass spectrometry in human plasma. *Clin. Chem. Lab. Med.*, **44**(**3**), 299–305 (2006).
21. Capparelli, E.V., Holland, D., Okamoto, C., Gragg, B., Durelle, J., Marquie-Beck, J., van den Brande, G., Ellis, R., Letendre, S. and HNRC Group, Lopinavir concentrations in cerebrospinal fluid exceed the 50% inhibitory concentration for HIV. *AIDS*, **19**(**9**), 949–952 (2005).
22. Hawkins, T., Veikley, W., St Claire, L.R. 3rd., Guyer, B., Clark, N. and Kearney, B.P., Intracellular pharmacokinetics of tenofovir diphosphate, carbovir triphosphate, and lamivudine triphosphate in patients receiving triple-nucleoside regimens. *J. Acquir. Immune Defic. Syndr.*, **39**(**4**), 406–411 (2005).
23. King, T., Bushman, L., Anderson, P.L., Delahunty, T., Ray, M. and Fletcher, C.V., Quantitation of zidovudine triphosphate concentrations from human peripheral blood mononuclear cells by anion exchange solid phase extraction and liquid chromatography-tandem mass spectroscopy; an indirect quantitation methodology. *J. Chromatogr. B*, **831**(**1–2**), 248–257 (2006).
24. Arndt, T., Guessregen, B., Hohl, A. and Reis, J., Determination of serum amantadine by liquid chromatography-tandem mass spectrometry. *Clin. Chim. Acta*, **359**(**1–2**), 125–131 (2005).
25. Chen, X., Liu, D., Zhu, L. and Zhong, D., Development and validation of a liquid chromatography/tandem mass spectrometry procedure for the quantification of adefovir in human plasma. *Rapid Commun. Mass Spectrom*., **19**(**13**), 1893–1898 (2005).
26. Dowse, R. and Futter, W.T., Outpatient complience with thephylline and phenytoin therapy. *S Afr. Med*., **80**, 550–553 (1991).

27. Linder, M.W. and Keck, P.E. Jr., Standards of laboratory practice: antidepressant drug monitoring. *Clin. Chem.*, **44**(**5**), 1073–1084 (1998).
28. Sauvage, F.L., Gaulier, J.M., Lachatre, G. and Marquet, P., A fully automated turbulent-flow liquid chromatography-tandem mass spectrometry technique for monitoring antidepressants in human serum. *Ther. Drug Monit.*, **28**(**1**), 123–130 (2006).
29. de Mendonca Lima, C.A., Baumann, P., Brawand-Amey, M., Brogli, C., Jacquet, S., Cochard, N., Powell-Golay, K. and Eap, C.B., Effect of age and gender on citalopram and desmethylcitalopram steady-state plasma concentrations in adults and elderly depressed patients. *Prog. Neuropsycopharmacol. Biol. Psychiatry*, **29**(**6**), 952–956 (2005).
30. Muchohi, S.N., Ward, S.A., Preston, L., Newton, C.R., Edwards, G. and Kokwaro, G.O., Determination of midazolam and its major metabolite 1′-hydroxymidazolam by high-performance liquid chromatography-electrospray mass spectrometry in plasma from children. *J. Chromatogr. B*, **821**(**1**), 1–7 (2005).
31. Thalamas, C., Rajman, I., Kulisevsky, J., Lledo, A., Mackie, A.E., Blin, O., Gillespie, T.A., Seger, M. and Rascol, O., Pergolide: multiple-dose pharmacokinetics in patients with mild to moderate Parkinson disease. *Clin. Neuropharmacol.*, **28**(**3**), 120–125 (2005).
32. Niederlander, H.A., Koster, E.H., Hilhorst, M.J., Metting, H.J., Eilders, M., Ooms. B. and de Jong, G.J., High throughput therapeutic drug monitoring of clozapine and metabolites in serum by on-line coupling of solid phase extraction with liquid chromatography-mass spectrometry. *J. Chromatogr. B*, **834**(**1–2**), 98–107 (2006).
33. Valdes, R. Jr., Jortani, A.S. and Gheorghiade, M., Standards of laboratory practice: cardiac drug monitoring. *Clin. Chem.*, **44**(**5**), 1096–1109 (1998).
34. Abad-Santos, F., Novalbos, J., Galvez-Mugica, M.A., Gallego-Sandin, S., Almeida, S., Vallee, F. and Garcia, A. G., Assessment of sex differences in pharmacokinetics and pharmacodynamics of amlodipine in a bioequivalence study. *Pharmacol. Res.*, **51**(**5**), 445–452 (2005).
35. Dieterle, W., Corynen, S., Vaidyanathan, S. and Mann, J., Pharmacokinetic interactions of the oral renin inhibitor aliskiren with lovastatin, atenolol, celecoxib and cimetidine. *Int. J. Clin. Pharmacol. Ther.*, **43**(**11**), 527–535 (2005).
36. Myung, S.W. and Jo, C.H., Gas chromatograph-mass spectrometric method for the determination of carvedilol and its metabolites in human urine. *J. Chromatogr. B*, **822**(**1–2**), 70–77 (2005).
37. Silva, L.C., Oliveira, L.S., Mendes, G.D., Gracia, G., Pereira Ados, S. and De Nucci, G., Quantification of isosorbide 5-mononitrate in human plasma by liquid chromatography-tandem mass spectrometry using atmospheric pressure photoionization. *J. Chromatogr. B*, **832**(**2**), 302–306 (2006).
38. Almeida, S., Filipe, A., Almeida, A., Gich, T., Antonijoan, R., Puntes, M., Barbanoj, M. and Cruz Caturla, M., Comparative study on the bioequivalence of two formulations of pravastatin. Data from a crossover, randomised, open-label bioequivalence study in healthy volunteers. *Arzneimittelforschung*, **56**(**2**), 70–75 (2006).
39. Trivedi, R.K., Kallem, R.R., Mullangi, R. and Srinivas, N.R., Simultaneous determination of rosuvastatin and fenofibric acid in human plasma by LC-MS/MS with electrospray ionization: assay development, validation and application to a clinical study. *J. Pharm. Biomed. Anal.*, **39**(**3–4**), 661–669 (2005).
40. Medvedovici, A., Albu, F., Georgita, C. and David, V., Non-extractive procedure followed by LC/APCI MS/MS analysis of trimetazidine in plasma samples for assessing bioequivalence of immediate/modified release formulations. *Biomed. Chromatogr.*, **19**(**7**), 549–555 (2005).
41. Moore, M.J. and Erlichman, C., Therapeutic drug monitoring in oncology. Problems and potential in antineoplastic therapy. *Clin. Pharmacokinet.*, **13**(**4**), 205–227 (1987).
42. Hon, Y.Y. and Evans, W.E., Making TDM work to optimize cancer chemotherapy: a multidisciplinary team approach. *Clin. Chem.*, **44**(**2**), 388–400 (1998).

43. Skolnik, J.M., Barrett, J.S., Shi, H. and Adamson, P.C., A liquid chromatography-tandem mass spectrometry method for the simultaneous quantification of actinomycin-D and vincristine in children with cancer. *Cancer Chemother. Pharmacol.*, **57**(**4**), 458–464 (2006).
44. Rudek, M.A., Zhao, M., He, P., Hartke, C., Gilbert, J., Gore, S.D., Carducci, M.A. and Baker, S.D., Pharmacokinetics of 5-azacitidine administered with phenylbutyrate in patients with refractory solid tumors or hematologic malignancies. *J. Clin. Oncol.*, **23**(**17**), 3906–3911 (2005).
45. Lee, J., Yoo, B.C., Lee, H.S., Yoo, H.W., Yoo, H.H., Kang, M.J. and Kim, D.H., Rapid quantitative determination of L-FMAU-TP from human peripheral-blood mononuclear cells of hepatitis B virus-infected patients treated with L-FMAU by ion-pairing, reverse-phase, liquid chromatography/electrospray tandem mass spectrometry. *Ther. Drug Monit.*, **28**(**1**), 131–137 (2006).
46. Rauh, M., Stachel, D., Kuhlen, M., Groschl, M., Holter, W. and Rascher, W., Quantification of busulfan in saliva and plasma in haematopoietic stem cell transplantation in children: validation of liquid chromatography tandem mass spectrometry method. *Clin. Pharmacokinet.*, **45**(**3**), 305–316 (2006).
47. Robertson, J.F., Erikstein, B., Osborne, K.C., Pippen, J., Come, S.E., Parker, L.M., Gertler, S., Harrison, M.P. and Clarke, D.A., Pharmacokinetic profile of intramuscular fulvestrant in advanced breast cancer. *Clin. Pharmacokinet.*, **43**(**8**), 529–538 (2004).
48. Zhao, M., Hartke, C., Jimeno, A., Li, J., He, P., Zabelina, Y., Hidalgo, M. and Baker, S.D., Specific method for determination of gefitinib in human plasma, mouse plasma and tissues using high performance liquid chromatography coupled to tandem mass spectrometry. *J. Chromatogr. B*, **819**(**1**), 73–80 (2005).
49. Titier, K., Picard, S., Ducint, D., Teilhet, E., Moore, N., Berthaud, P., Mahon, F.X. and Molimard, M., Quantification of imatinib in human plasma by high-performance liquid chromatography-tandem mass spectrometry. *Ther. Drug Monit.*, **27**(**5**), 634–640 (2005).
50. Pappas, P., Karavasilis, V., Briasoulis, E., Pavlidis, N. and Marselos, M., Pharmacokinetics of imatinib mesylate in end stage renal disease. A case study. *Cancer Chemother. Pharmacol.*, **65**(**4**), 358–360 (2005).
51. Appels, N.M., van Maanen, M.J., Rosing, H., Schellens, J.H. and Beijnen, J.H., Quantitative analysis of the farnesyl transferase inhibitor lonafarnib (Sarasartrade mark, SCH66336) in human plasma using high-performance liquid chromatography coupled with tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **19**(**15**), 2187–2192 (2005).
52. He, X., Batchelor, T.T., Grossman, S. and Supko, J.G., Determination of procarbazine in human plasma by liquid chromatography with electrospray ionization mass spectrometry. *J. Chromatogr. B*, **799**(**2**), 281–291 (2004).
53. Gjerde, J., Kisanga, E.R., Hauglid, M., Holm, P.I., Mellgren, G. and Lien, E.A., Identification and quantification of tamoxifen and four metabolites in serum by liquid chromatography-tandem mass spectrometry. *J. Chromatogr. A*, **1082**(**1**), 6–14 (2005).
54. White, S. and Wong, S.H.Y., Standards of laboratory practice: analgesic drug monitoring. *Clin. Chem.*, **44**(**5**), 1110–1123 (1988).
55. Batista, C., Berisha, M., Karst, M., Salim, K., Schneider, U. and Brenneisen, R., Determination of ajulemic acid and its glucuronide in human plasma by gas chromatography-mass spectrometry. *J. Chromatogr. B*, **820**(**1**), 77–82 (2005).
56. Abe, R. and Alvarez, J.C., Sensitive quantification of apomorphine in human plasma using a LC-ESI-MS-MS method. *Ther. Drug Monit.*, **28**(**3**), 407–412 (2006).
57. Chen, Y.L., Hanson, G.D., Jiang, X. and Naidong, W., Simultaneous determination of hydrocodone and hydromorphone in human plasma by liquid chromatography with tandem mass spectrometric detection. *J. Chromatogr. B*, **769**(**1**), 55–64 (2002).

58. Scott, G., Rordorf, C., Reynolds, C., Kalbag, J., Looby, M., Milosavljev, S., Weaver, M., Huff, J.P. and Ruff, D.A., Pharmacokinetics of lumiracoxib in plasma and synovial fluid. *Clin. Pharmacokinet.*, **43**(**7**), 467–478 (2004).
59. Fang, W.B., Andrenyak, D.M., Moody, D.E. and Nuwayser, E.S., Determination of nalmefene by high-performance liquid chromatography-electrospray ionization-tandem mass spectrometry. *J. Anal. Toxicol*., **29**(**3**), 169–174 (2005).
60. Shang, E., Xiang, B., Liu, G., Xie, S., Wei, W. and Lu, J., Determination of phenazopyridine in human plasma via LC-MS and subsequent development of a pharmacokinetic model. *Anal. Bioanal. Chem.*, **382**(**1**), 216–222 (2005).
61. Miksa, I.R., Cummings, M.R. and Poppenga, R.H., Multi-residue determination of anti-inflammatory analgesics in sera by liquid chromatography–mass spectrometry. *J. Anal. Toxicol.*, **29**(**2**), 95–104 (2005).
62. Hori, Y., Fujisawa, M., Shimada, K., Hirose, Y. and Yoshioka, T., Method for screening and quantitative determination of serum levels of salicylic Acid, acetaminophen, theophylline, phenobarbital, bromvalerylurea, pentobarbital, and amobarbital using liquid chromatography/electrospray mass spectrometry. *Biol. Pharm. Bull.*, **29**(**1**), 7–13 (2006).
63. Vollenbroeker, B., Koch, J.H., Fobker, M., Suwelack, B., Hohage, H. and Muller, U., Determination of cyclosporine and its metabolites in blood via HPLC-MS and correlation to clinically important parameters. *Transplant. Proc.*, **37**(**4**), 1741–1744 (2005).
64. Guddat, S., Thevis, M. and Schanzer, W., Identification and quantification of the plasma volume expander dextran in human urine by liquid chromatography-tandem mass spectrometry of enzymatically derived isomaltose. *Biomed. Chromatogr.*, **19**(**10**), 743–750 (2005).
65. Sekera, M.H., Ahrens, B.D., Chang, Y.C., Starcevic, B., Georgakopoulos, C. and Catlin, D.H., Another designer steroid: discovery, synthesis, and detection of 'madol' in urine. *Rapid Commun. Mass Spectrom*., **19**(**6**), 781–784 (2005).
66. Thevis, M., Thomas, A., Delahaut, P., Bosseloir, A. and Schanzer, W., Qualitative determination of synthetic analogues of insulin in human plasma by immunoaffinity purification and liquid chromatography-tandem mass spectrometry for doping control purposes. *Anal. Chem.*, **77**(**11**), 3579–3585 (2005).
67. Heudi, O. and Fontannaz, P., Determination of vitamin B5 in human urine by high-performance liquid chromatography coupled with mass spectrometry. *J. Sep. Sci.*, **28**(**7**), 669–672 (2005).
68. Tsugawa, N., Suhara, Y., Kamao, M. and Okano, T., Determination of 25-hydroxyvitamin D in human plasma using high-performance liquid chromatography–tandem mass spectrometry. *Anal. Chem.*, **77**(**9**), 3001–3007 (2005).
69. Suhara, Y., Kamao, M., Tsugawa, N. and Okano, T., Method for the determination of vitamin K homologues in human plasma using high-performance liquid chromatography-tandem mass spectrometry. *Anal. Chem.*, **77**(**3**), 757–763 (2005).

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 14

Mass spectrometry of infectious pathogens

PLAMEN A. DEMIREV*

Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, MS 2-217, Laurel, MD 20723–6099, USA

1. Introduction

It is estimated that a little over 1400 pathogen species (217 viruses, 538 bacteria and *Rickettsia,* 307 fungi, 66 protozoa, and 287 helminths) are potential etiological sources for infectious diseases in humans [1]. Among these, more than 90% of the infectious disease mortality world-wide is caused by seven major illnesses: the figures for 2002 are [2], in million: lower respiratory infections—3.9; HIV—2.8; diarrhea—1.6; tuberculosis—1.6; malaria—1.2; and measles—0.6. While it is obvious that a number of socio-economic and environmental factors (poverty, wars, climate, the emergence of drug-resistant pathogen strains, etc.) influence these numbers significantly, particularly in the developing world, naturally occurring

*Tel.: 443-778-7712; Fax: 443-778-6905. E-mail: plamen.demirev@jhuapl.edu (P.A. Demirev).

(or intentionally inflicted) infectious diseases are a major threat for the industrialized countries as well. As a result of the rapidly accelerating trends of globalization, infectious disease pandemics are projected to occur in the not too distant future since geographical and political boundaries offer but trivial impediments to pathogen spread.

It is clear that the development of new and more efficient molecular-level diagnostic tools would certainly improve our ability to fight much more efficiently both current and emerging infections [3]. Effective responses to emerging pathogens require enhanced capabilities for rapid and accurate microorganism detection and identification, discovery of new drugs, finding reliable biomarkers for diseases, and/or creating new vaccines. In the post-genomic era, the emphasis in cellular and molecular biology is being gradually shifted from DNA sequencing projects to large-scale efforts in identification of individual proteins, their 3D structures, expression levels, post-translational modifications, network relationships, and metabolism products. The two major components of such a holistic "systems biology" approach are proteomics and metabolomics, and mass spectrometry (MS) is the most prominent technology currently applied in both fields [4,5].

For more than three decades MS has been a major analytical tool for the characterization of diverse microorganisms in the laboratory [6,7]. MS, among other spectroscopic methods like NMR, has been indispensable for structural elucidation of various classes of natural products, originating from microorganisms, e.g., cyclic peptide antibiotics. After the introduction of the soft ionization MS techniques—matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) [8–10]—a number of new MS applications in life sciences and medicine have emerged. These soft ionization techniques (recognized by the Nobel Prize in Chemistry in 2002) allowed for the first time the ionization and transfer into vacuum of large (>30 kDa) intact non-volatile biomolecules, such as proteins. In a MALDI experiment, a low-mass photo-absorbing organic compound (matrix) is added to a sample prior to irradiation with nanosecond laser pulses to desorb high-mass biomolecular ions [9,10]. In ESI, large multiply charged ions are generated by injecting the protein analyte solution through a capillary needle biased at high voltage (several kilovolts). Several stages of differential pumping and suitable ion optics allow the interfacing of an ESI ion source, operating at atmospheric pressure (AP), with a mass spectrometer operating in high-vacuum conditions [8]. In the post-genomic era, MS approaches, combined with bioinformatics algorithms and genome databases, have been applied to analyze the proteomes of a number of pathogens. A number of large-scale proteomics strategies for systems biology investigation of microorganisms involve multistage chromatography/fractionation and ESI and/or MALDI tandem MS, e.g., *Plasmodium falciparum* [11,12], *Bacillus anthracis* [13], and the SARS virus [14,15] (for a recent collection of articles on proteomics of microbial pathogens, see ref. [16]).

The current paradigm for rapid MS identifications of pathogens relies on the detection and identification of unique biomarker molecules from experimental mass spectra. This paradigm can be traced back to Anhalt and Fenselau [17], who demonstrated that lower mass biomolecules from different pathogenic bacteria, introduced intact into a mass spectrometer, could be vaporized and ionized directly by electron impact. Structural elucidation of the unique chemical biomarkers from different organisms has been achieved by MS [17]. Furthermore, the signature composition and abundances allowed taxonomic distinctions between the microorganisms to be made. In the following, current developments of this paradigm will be illustrated using various forms of laser desorption (LD) MS as a major diagnosis and detection platform. In particular, MALDI MS for rapid identification of intact *Bacillus* spore species, as well as LD MS for detection in blood of *Plasmodium* parasites (the causative agent of malaria), will be discussed.

1.1. Highlights for medical professionals

The burden of infectious diseases to global health continues to increase. Effective diagnosis of current or emerging infectious diseases requires enhanced capabilities for rapid and accurate pathogen characterization. Traditional approaches for pathogen diagnosis involve microorganism growth with all of its associated drawbacks—slow, inadequate sensitivity, and non-viable cultures (sometimes due to previous or concomitant anti-microbial therapy). Novel molecular-level technologies are being developed for reliable detection and identification of both natural and bioengineered microorganisms with applications ranging from science and medicine to homeland security. MS is one such emerging biosensor technology for diagnosis of infectious diseases with several practical advantages—speed, sensitivity, and specificity. MS is rapid: a typical experiment including sample collection and preparation takes minutes—in contrast to days for (often times retrospective) diagnosis via the classical microbiology methods. This is the major advantage of MS. It is also to be compared to other molecular-based pathogen diagnostic methods—e.g., antibody recognition (ELISA) or DNA detection after PCR amplification—that take several hours to complete. MS is broadband: upon appropriate sample preparation it can detect all types of pathogens—viruses, vegetative bacteria and fungi and their spores, as well as parasitic protozoa. MS is sensitive: typically, a signal with a sufficient signal/noise ratio can be generated from a sample, containing less than 10^4 organisms. MS can be interfaced to a variety of sample collection and sample processing modules to allow versatile sampling from different environments—from aerosols to biofluids. MS can be automated and it is also computer-friendly: e.g., latest developments in bioinformatics can be coupled to MS experimental data for robust pathogen diagnosis. Furthermore, MS instruments can be miniaturized and deployed in the field.

The current paradigm in pathogen detection and identification by MS is based on the fact that a mass spectrum of a microorganism contains masses of intact *biomarker molecules,* uniquely characteristic for that organism. Such an experimental mass spectrum (mass/charge ratios of the various biomarker ions vs. their abundances) forms a characteristic "fingerprint" signature of the respective species. These signatures can be derived experimentally (e.g., by acquiring mass spectra from a large number of microorganisms under a variety of conditions) or deduced by bioinformatics means.

In what follows, the uses of MS for pathogen detection would be demonstrated in the examples of intact *Bacillus* spore species, as well as *Plasmodium* parasites—the causative agent of malaria. The types of characteristic biomarker molecules detected by MS for these two very different cases would be discussed as well as the factors influencing mass spectra of intact organisms.

1.2. Highlights for chemists

In the past decade two ionization methods—MALDI and ESI—have formed the basis of the new MS. These methods have allowed for the first time the transfer of large (>30 kDa) intact non-volatile biomolecules, such as proteins, into vacuum. Thus, the molecular masses of individual large biomolecules could be determined with unprecedented accuracy by MS. In the post-genomic era, this new MS, combined with bioinformatics algorithms and genome databases, is the cornerstone of proteomics: the field devoted to characterization of all expressed proteins in a cell. Furthermore, it has been demonstrated that the mass spectrum of an intact microorganism contains the masses of intact *biomarker molecules,* uniquely characteristic for that organism: a mass spectral "fingerprint" signature. For pathogen identification by MS several major classes of biomarker molecules can be exploited: proteins (50% of the dry weight of an individual cell), DNA (one double-strand copy per cell), RNA (0.01–1%), and polar and non-polar lipids (4–9%). The paradigm behind the use of MS for pathogen detection and identification is the *uniqueness of the MS biomarker signature* of the respective species.

While ESI, combined with various mass analyses, is mainly used for large-scale proteomics characterization of different pathogens, LD MS methods form the basis for the use of MS as a biosensor platform for rapid microorganism detection and identification. For instance, in a MALDI experiment a low-mass photoabsorbing organic compound (a matrix) is added to a sample, containing intact microorganisms (viruses, spores, vegetative bacterial cells) or a protein toxin. The solid sample/matrix surface is irradiated with UV (typically 337 nm) nanosecond laser pulses. The photon–matrix molecule interactions induce rapid sample expansion and material ablation into a plume with well-defined aerodynamics characteristics. As a result of (predominantly) ion–molecule, ion–ion, and ion–electron reactions in the expanding plume, microorganism-specific biomarker ions are

formed. They are accelerated in an electric field and subsequently detected by a time-of-flight (TOF) mass analyzer.

A large number of experimental factors influence the MALDI MS of intact microorganisms: various degrees of extraction of individual protein biomarkers from cells (performed *in situ* on the MALDI sample holder by addition of matrix solution), variability of biomarker protein expression levels due to variations in growth conditions (time, growth media), varying ionization efficiencies for different biomarker molecules (e.g., as a function of matrix/biomarker protein ratios), variations in laser pulse energy, detection efficiency variations from instrument to instrument, etc. While variability in peak intensity is observed in MALDI mass spectra of intact pathogens, the mass values of the observed biomarkers remain the same. In this chapter, the basics of LD MS for rapid pathogen characterization are illustrated in the examples of intact bacterial species—*Bacillus* spores, as well as *Plasmodium* parasites—the causative agent of malaria.

2. Methodology

2.1. MALDI MS-based methods for Bacillus spore species characterization

Recently, MALDI MS had received considerable attention as a method for rapid and highly reliable detection and identification of intact microorganisms—viruses, bacteria and bacterial spores, and fungi [18,19]. MALDI MS as a method for pathogen characterization has several advantages. It is rapid—a typical experiment, including sample collection and sample preparation, takes minutes (vs. days for classical microbiology experiments). MALDI MS is broadband, i.e., it can detect not only microorganisms, but also protein and non-protein toxins (e.g., lower mass non-volatile substances such as saxitoxin and palitoxin). The latter feature distinguishes MS from all DNA-based technologies, which would require the presence of DNA from the producing organism. MS is sensitive—typically, a signal with a sufficient signal/noise ratio can be generated from a sample, containing less than 10^4 organisms, or a few femtomoles of a toxin, respectively. MS can be interfaced to a variety of sample collection and sample processing modules to allow versatile sampling from different environments (aerosols, liquids, powders). MALDI MS can be applied directly to intact microorganisms without the need for, e.g., protein separation and isolation. In that case, addition of slightly acidic matrix solution facilitates *in situ* cell lysis and extraction (on the MALDI slide). MS is easily automated and is computer-friendly—e.g., latest developments in bioinformatics and genome databases can be coupled to MS experimental data for robust identification of microorganisms [20]. Furthermore, MS instruments are robust and can be miniaturized [21]. A family of MALDI TOF instruments for pathogen detection have been described that fit, e.g., in a regular suitcase for field-portable use [22,23].

The paradigm for MALDI TOF MS of intact microorganism identification can be illustrated in the example of two *Bacillus* spore species (Fig. 1): *different microorganisms give rise to different mass spectra*. This is due to the presence of different expressed biomarker molecules in different organisms. In this case, the differences between the two spectra (due to different masses of the detected biomarker molecules) allow distinguishing between the two spore species. The characteristic peaks in the respective mass spectra correspond to biomarkers from different classes of molecules. In MALDI MS, the typical classes employed for differentiation between microorganisms (in some instances after fractionation, clean-up, and additional sample preparation procedures) are: proteins (50% of dry weight) [18,19]; DNA (one copy per cell!) and RNA (0.01–1%) [24–26]; and various polar lipids (4–9%) [7]. For the *Bacillus* spores (Fig. 1), $\sim 10^5$ intact spores are deposited from spore suspensions and mixed with a MALDI matrix/10% trifluoroacetic acid solution to extract a set of small ($\sim$70 amino acids) proteins with particularly high abundance in dry spores [27,28]. It takes less than 10 min from suspension of the spores in, e.g., water, to obtaining the actual MALDI mass spectra shown. The amount of liquids used is less than 1 μl. It is of paramount importance to be able to elucidate the nature of the observed biomarkers from intact microorganisms. The proteins detected here—small acid-soluble spore proteins (SASP)—have sequences that differ among different species and as a result the SASPs have different masses as evident in the mass spectra [20]. Another class of biomarkers with masses $\sim$1 kDa can be observed in the spectra of intact spore species as well (Fig. 1). These are lipopeptides, characteristic for specific spore species, and their structures have been elucidated by MS [18,29].

The observed protein biomarkers in MALDI mass spectra from intact bacteria are typically highly expressed proteins with housekeeping functions, such as ribosomal, chaperone, and translation/transcription factor proteins [20,30]. To achieve identification, experimental MALDI mass spectra can be compared with a collection of mass spectra of known organisms—MS fingerprints—compiled into a reference biomarker signature library (Fig. 2a). However, the biomarker fingerprints detected by MALDI MS exhibit variations (e.g., different biomarkers observed under different conditions) as a result of various factors—sample preparation, instrumental parameters, microorganism biochemistry, and environmental conditions, such as diverse biological backgrounds [18,19]. A large number of spectra for each targeted microorganism need to be compiled in the library so that the "fingerprint" approach is effective. Standardized experimental protocols have been developed for methicillin-resistant *Staphylococcus aureus* in order to achieve reliable and reproducible species-level identification and sub-typing from the MALDI MS fingerprint libraries. The effects of a number of parameters—incubation period, method of deposition of cell suspension, matrix solution concentration and drying time, time between sample preparation and analysis, and MALDI TOF instrument parameters—have been examined in order to optimize

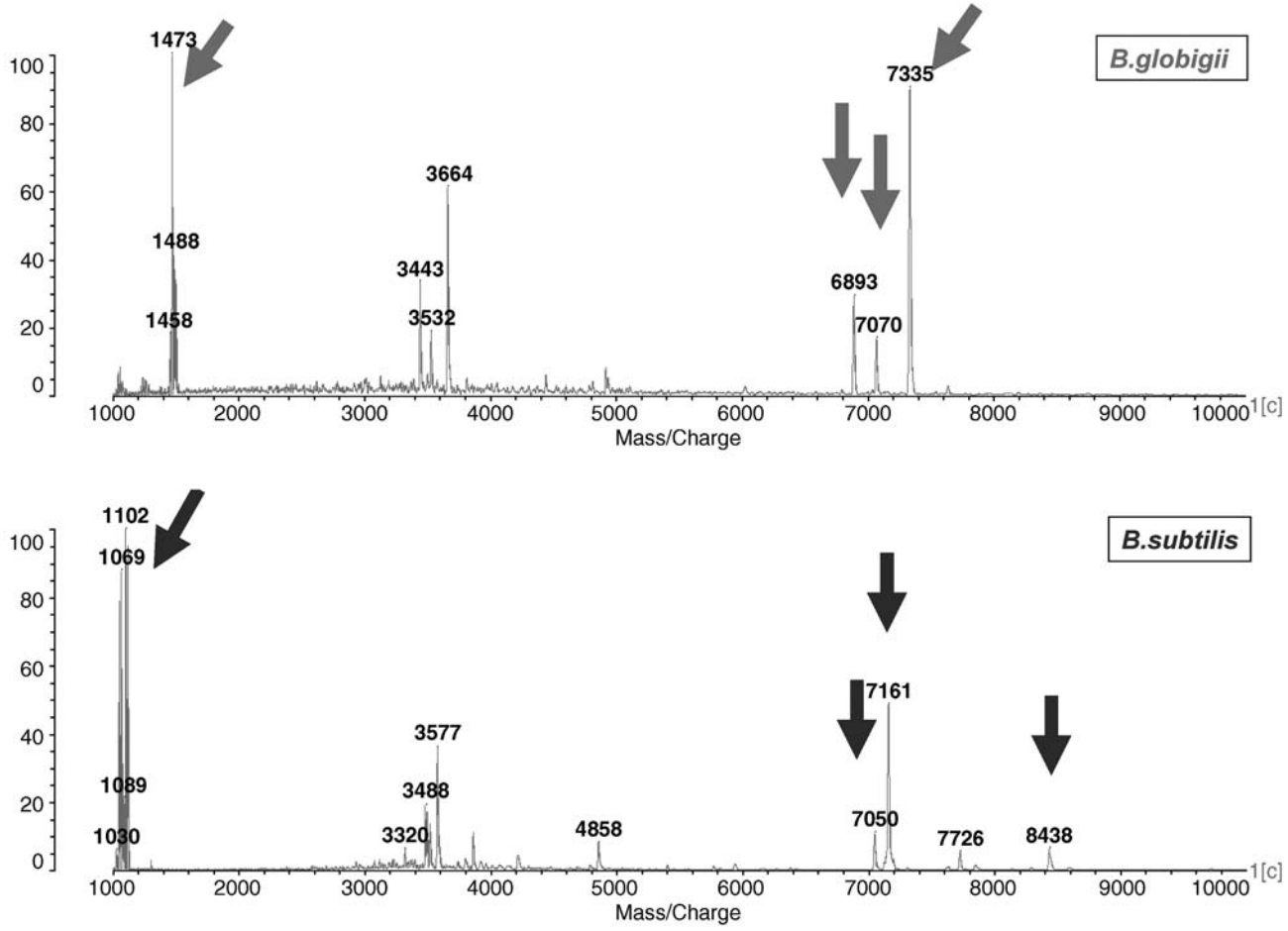


Fig. 1. Positive ion MALDI TOF mass spectra of intact *Bacillus* spore species. The peaks denoted by arrows in the *mass/charge* range ~7000 correspond to *biomarker proteins* (SASP) that differ in amino acid sequence between the two species and allow species differentiation and identification.

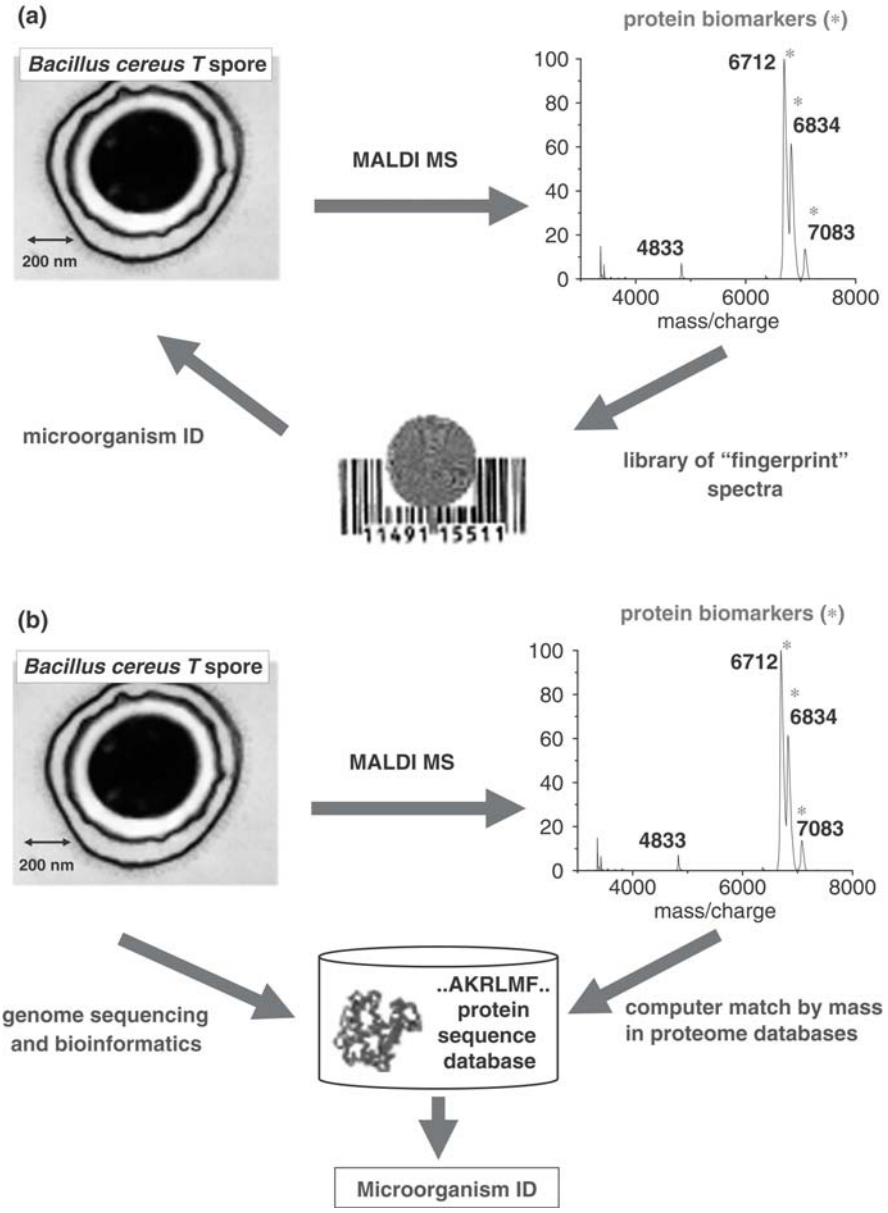


Fig. 2. MS methods for rapid microorganism identification: (a) fingerprint library matching of mass spectra and (b) bioinformatics-based strategies. While the former rely on previously collected experimental mass spectra, the latter rely on *in silico* prediction of biomarker protein sequences from the respective genome sequences.

the protocol [31]. In the case of continuous *Escherichia coli* cultures, maintained in a bioreactor at specific growth rates and pH, the microorganism was correctly identified from the MALDI mass spectral library in all cases except for a biofilm sample collected from the reactor [32]. There are other practical limitations to the fingerprint library approach, e.g., no MS signature may be available at hand for a novel or a highly pathogenic organism.

An alternative bioinformatics-based strategy (Fig. 2b) for generation of protein biomarker signatures has been proposed [20,30,33–35]. The expected protein biomarkers for a microorganism that could be detected in a MALDI spectrum are determined from the masses of a subset of all of its potentially expressible proteins, which in turn are available once the respective organism's genome is sequenced. There is no need to create experimental MS libraries in the bioinformatics-based approach, which is its major difference from the fingerprint approach. In both approaches, the experimental MS data are compared to expected masses (from reference spectra in the traditional approach, derived from the genome in the new approach) and the microorganism that provides the most statistically significant matches is selected. Different sets of proteins can be expressed in a microorganism and experimentally observed by MS (depending on growth stage, growth medium, etc.). The masses of all these proteins can be independently derived *in silico* from their amino acid sequences. The sequences of these proteins can be found in Internet-accessible proteome databases together with their organism sources, provided the genome of the particular pathogen is known. Conditions and requirements that secure the successful application of such a bioinformatics-based approach have been discussed [20,30,33–35]. These database conditions include: *completeness*—availability of the genome sequence for the particular pathogen, and *fidelity*—capability to predict/incorporate various post-translational modifications [35]. Currently (mid-2006), there are more than 300 completely sequenced and publicly available bacterial genomes (The Institute for Genomic Research, www.tigr.org). The list includes all microbial pathogens on the CDC priority agents list. The bioinformatics approach for microorganism identification has been successfully demonstrated in a blind study by constructing *in silico* a database of highly expressed proteins (e.g., ribosomal proteins) for more than 30 sequenced microorganisms [30]. The obtained results scale to a database of ~1000 sequenced different pathogenic organisms being successfully detected at 95% confidence level (Table 1).

The availability of genome information for a particular pathogen allows the adaptation of bottom-up [36,37] or top-down [38–40] proteomics methodologies for microorganism identification by MS [41–45] (Fig. 3). These proteomics-based approaches are based on the initial identification of one or more individual protein biomarkers (from, e.g., their corresponding tryptic peptides and/or tandem mass spectra). By inference, the microorganism from which these proteins originate is then identified. For instance, rapid *in situ* (on a MALDI sample slide) proteolysis of

Table 1
Microorganism identification using bioinformatics-derived signatures

Organism	Ribosomes in mass range	Detection (%)	
		Matrix	
		α-Cyano	Synapinic acid
Bacillus subtilis	31	100	100
Escherichia coli	30	100	100
Pseudomonas aeruginosa	26	100	100
Haemophilus influenzae	25	100	100
Bacillus stearothermophilus	20	60	100
Bacillus halodurans	10	0	0
Salmonella typhimurium	7	0	0
Micrococcus luteus	5	0	0
Acinetobacter cloacoa	0	0	0

Source: Adapted with ACS permission from ref. [30].
Note: In this case, only highly expressed ribosomal proteins are included in the *in silico* generated biomarker database. Results scale to a database of ~1000 organisms at 95% confidence level of identification.

proteins, derived from intact viruses, can facilitate rapid virus identification [41]. Proteolytic peptides from SASP biomarkers in various *Bacillus* spores have been analyzed by several different types of tandem mass spectrometers—a MALDI TOF MS with a curved-field ion reflectron [42], a hybrid ion trap/TOF mass spectrometer [43], or an AP MALDI ion trap instrument [44]. In all these instances, unambiguous spore species identification has been provided after SASP biomarkers were identified from the partial sequences of their proteolytic peptides, combined with proteome-based database queries using the MASCOT search engine.

The capability to identify an intact protein by deducing its partial amino acid sequence (a sequence tag) in a Fourier transform ion cyclotron resonance (FTICR) MS/MS experiment and subsequent homology search in a proteome database was first demonstrated by Mortz et al. [38]. In analogy to bottom-up proteomics, unambiguous identification of one or more intact protein biomarkers by top-down proteomics allows successful microorganism identification (provided the proteome database contains both the respective protein sequences and the respective organism sources). Top-down proteomics approaches for *Bacillus* spore protein biomarkers and from there *Bacillus* species identification have been also described [27,45]. For instance, biomarker proteins from *Bacillus cereus T* spores have been analyzed

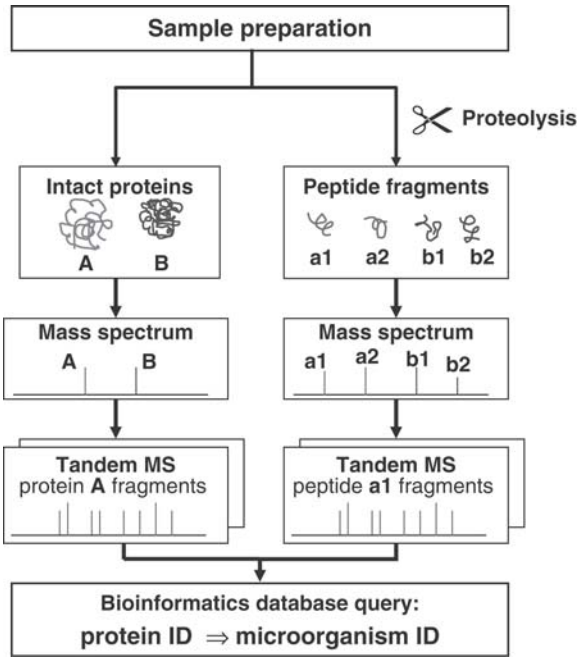


Fig. 3. "Top-down" vs. "bottom-up" MS proteomics-based approaches for microorganism identification.

by high-resolution tandem FTICR MS [27]. Fragmentation-derived sequence tags and BLAST sequence similarity searches in a proteome database identify the major biomarker protein, observed in MALDI MS of intact spores as a SASP. Following individual protein identification, the spore species itself could be unambiguously identified [27]. MALDI TOF/TOF MS of whole (undigested) protein biomarkers has been described recently as a method for direct and rapid identification of individual *Bacillus* spore species, either pure or in a mixture [45] (Fig. 4). A major advantage of this method is that biomarker MS/MS spectra are obtained without the need for biomarker pre-fractionation, digestion, separation, and cleanup.

2.2. LD MS detection of Plasmodium parasites in blood

Recently, a novel physical method for rapid and sensitive malaria detection in blood has been developed [46–50]. This method—ultraviolet LD MS—is based on the detection of heme (iron protoporphyrin) in blood as a qualitative and quantitative malaria biomarker, both *in vitro* [46] and *in vivo* [49,50]. In infected erythrocytes, the parasite sequesters heme from digested hemoglobin in a molecular crystal (malaria pigment or hemozoin). LDMS detects only heme from hemozoin in parasite-infected blood, and not heme, bound to hemoglobin

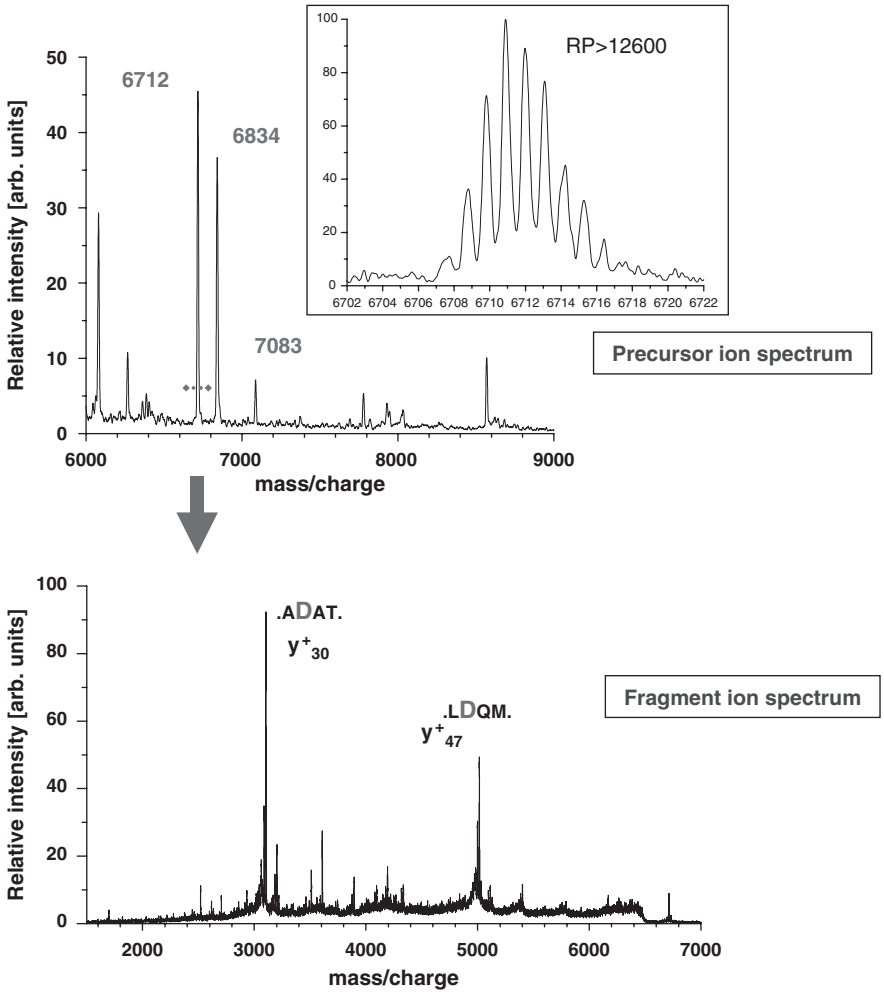


Fig. 4. Top-down proteomics for spore identification from MALDI TOF/TOF mass spectra of biomarker molecular ions: (top) precursor ion spectrum and (bottom) fragment ion spectrum. The inset on top represents the expanded region around the singly charged precursor ion at *m/z* 6712. The resulting 68 fragment ions are matched against proteins in Swiss-Prot and the protein P0A4F4 (SAS2_BACCR), originating from *Bacillus cereus,* is identified as the most plausible candidate (adapted with ACS permission from ref. [45]).

or other proteins in uninfected blood samples (Fig. 5). Formation of hemozoin crystals is a unique evolutionary feature of *Plasmodium* parasites. The parasite presents a volume of high concentration of purified biomarker molecules, uniquely suited for sensitive and specific detection of malaria by LDMS. Thus, the parasite fractionates, purifies, and concentrates the biomarker molecule

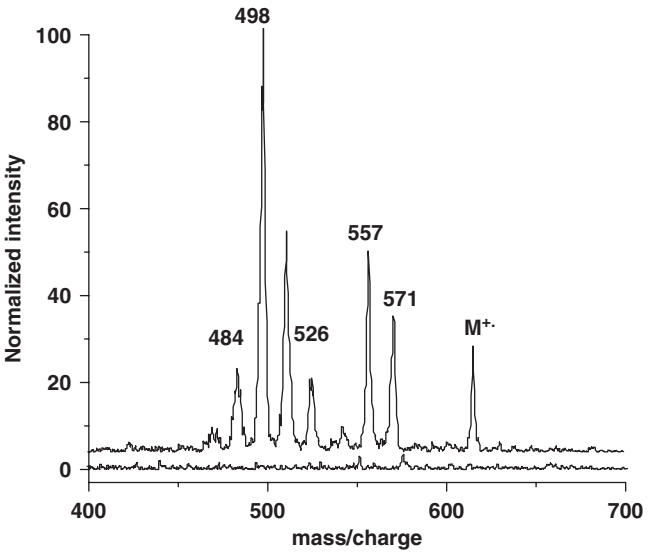


Fig. 5. Positive ion LD TOF mass spectra: (top trace) blood sample from a patient infected with *P. malariae* parasites (to be published) and (bottom trace) control sample (uninfected blood). A commercial LD TOF system is used, and both spectra are normalized to the same detector response value. Each trace represents the average of 100 single laser shot spectra, obtained from linear scanning of an individual sample well. The molecular ion ($M^{+}\cdot$ at *m/z* 616) and several characteristic "fingerprint" fragment ions of desorbed heme are denoted in the upper trace.

(heme), in effect performing itself (rather than a researcher) the complex and time-consuming sample preparation tasks required for detection of a biomarker in MS-based proteomics! The LDMS detection of malaria requires only a drop of blood. The method is pan-malarial, i.e., all four *Plasmodium* species, infecting humans, are detected. In contrast to MALDI, external photo-absorbing matrix does not need to be added to the sample. The heme molecule—a 22 π-electron conjugated protoporphyrin system—is an efficient photo-absorber in the visible and near UV (with an absorption maximum near 400 nm). The heme photo-physical properties, concurrently with its low ionization potential, warrant that direct LDMS possesses extremely low limits for heme detection (less than 10 parasites per 1 μl blood can be detected). In LD, the heme molecular ion dissociates in several structure-specific fragments, providing at least five additional peaks for detection of the biomarker molecule. LDMS detection of heme is quantitative, and heme signal tracks the number of parasites per unit volume of blood. A simplified sample protocol requiring minimal handling combined with miniaturized LD TOF instruments can permit the large-scale deployment of automated screening systems for rapid and affordable malaria diagnosis in large populations.

2.3. Other MS-based methods for pathogen detection/identification

A combined laser fluorescence/laser ionization TOF mass spectrometer has been developed recently as a tool to identify individual airborne, micrometer-sized particles, comprised of a single cell or a small number of clumped cells [51,52]. The technique, termed *bioaerosol mass spectrometry* (BAMS), has been evaluated for real-time detection and identification of individual aerosolized *Bacillus* spore species [51] or *Mycobacterium tuberculosis* particles [52]. This approach is reagent-less, i.e., no sample preparation with the associated liquid handling is required. However, only lower mass ($<m/z$ 200) positive and negative ions are ablated and detected. In the reported studies, two *Bacillus* spore species have been distinguished from one another and from other biological and abiological background materials by BAMS with no false positives at a sensitivity of 92%. In addition, the BAMS mass spectral signatures for aerosolized *M. tuberculosis* particles are distinct from *M. smegmatis, Bacillus atrophaeus,* and *B. cereus* particles. In a background-free environment, BAMS is capable of detecting *M. tuberculosis* at airborne concentrations of $\sim$1 particle/l. This technique is tested as a stand-alone airborne *M. tuberculosis* detector in bioaerosols from an infected patient.

MALDI TOF MS has been also used for simultaneous detection of multiple target microorganisms using bacteriophage amplification [53]. In this approach the target pathogenic bacteria are infected with bacteria-specific bacteriophages (e.g., MS2 and MPSS-1 phages specific for *E. coli* and *Salmonella* spp., respectively). Proteins, indicative of the progeny phages, are detected and utilized as a secondary biomarker for the target pathogen. For instance, *E. coli* when mixed with both MS2 and MPSS-1 produces only a MS2 biomarker protein (13.7 kDa). Mixing *Salmonella* spp. with both phages results in detection of the biomarker (a protein at 13.5 kDa) characteristic of MPSS-1. Amplification of both phages in a mixture of the two bacteria leads to detection of biomarkers characteristic for both MS2 and MPSS-1 (no deleterious effects on bacteriophage amplification have been observed).

An entirely different approach for biological warfare agent detection, combining nucleic acid detection with MS, has been described recently [25,26]. In this approach, analysis of PCR-amplified variable regions of microbial genomes is performed by ESI MS. The approach is termed TIGER (Triangulation Identification for the Genetic Evaluation of Risks), and relies on "intelligent PCR primers" to target broadly conserved regions that flank the variable genome regions. The sample preparation procedure takes more than an hour. The masses of PCR products with lengths between 80 and 140 base pairs must be determined with accuracy better than 20 ppm (i.e., better than $\pm$0.35 Da for a 35 kDa molecule!). Such accuracy should allow unambiguous assignment of the base composition of the amplified regions, which unequivocally determine the microorganism based on comparison with available genome sequences. The sample for analysis by this method can originate from air filtration devices, clinical samples, or other sources. Examples

illustrating the TIGER approach include *B. anthracis,* DNA-genome viruses from the Poxviridae family (whose members include the smallpox virus), and RNA-genome viruses (e.g., alphaviruses) [25]. In another application, a high-throughput and high-resolution genotyping and relative quantification of pathogenic bacteria from complex mixtures in respiratory samples has been performed [26]. High concentrations of several respiratory pathogens (*Haemophilus influenzae, Neisseria meningitidis,* and *Streptococcus pyogenes*) have been revealed confirming the polymicrobial nature of respiratory disease epidemics. Although appealing, so far the TIGER approach has been demonstrated only on FTICR mass spectrometer with a superconducting magnet—a device that can be utilized only in specialized laboratory conditions. Furthermore, PCR—a necessary step in this approach—is time-consuming and sensitive to contaminants.

3. Future trends and prospects

The transformation of MS into a viable tool for biomedical diagnostics has been a long-standing goal of mass spectrometrists [54]. Current developments in MALDI and LD MS for pathogen detection may bring us closer to that goal. Recent instrumental developments have demonstrated the capability of building small field-portable and inexpensive LD mass spectrometers. Additional work is certainly needed in order to develop simplified sample protocols for detection of a particular pathogen infection in bodily fluids, e.g., blood, urine, saliva, etc. To facilitate large-scale rapid and affordable screening and diagnosis for infectious pathogens in large populations, these protocols would require minimal sample preparation and handling. Validation of pathogen-specific disease biomarkers (both *in vivo* and *ex vivo*) by modern MS proteomics technologies is another obvious avenue for research to be pursued. Miniaturized multi-array LD TOF MS instruments and advanced signal processing can be implemented in a laboratory setting for screening of samples to detect the presence of infectious pathogens like *Plasmodium*. MS devices can be further incorporated into a framework of multi-tiered technologies for pathogen detection. In the future, such a framework will merge various molecular-level sensor platforms, e.g., MS, lab-on-a-chip (microfluidics) devices, DNA and protein microarrays, and computer bioinformatics algorithms for rapid and automated infectious pathogen diagnosis at point-of-care facilities.

References

1. Taylor, L.H., Latham, S.M. and Woolhouse, M.E.J., Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B*, **356**, 983–989 (2001).
2. WHO, *World Health Report 2002*, 2004.

3. Holland, C.A. and Kiechle, F.L., Point-of-care molecular diagnostic systems—past, present and future. *Curr. Opin. Microbiol.*, **8**, 504–509 (2005).
4. Washburn, M.P., Wolters, D. and Yates, J.R., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.*, **19**, 242–247 (2001).
5. Aebersold, R. and Mann, M., Mass spectrometry-based proteomics. *Nature*, **422**, 198–207 (2003).
6. Odham, G., Larsson, L. and Mardh, P.-A. (Eds.), *Gas Chromatography–Mass Spectrometry*: *Applications in Microbiology*. Plenum Press, New York, 1984.
7. Fenselau, C. (Ed.), *Mass Spectrometry for the Characterization of Microorganisms*. American Chemical Society, Washington DC, May 1994.
8. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse, C.M., Electrospray ionization for mass-spectrometry of large biomolecules. *Science*, **246**, 64–71 (1989).
9. Karas, M. and Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons. *Anal. Chem.*, **60**(**20**), 2299–2301 (1988).
10. Tanaka, K., The origin of macromolecule ionization by laser irradiation. *Angew. Chem.-Int. Ed.*, **42**, 3860–3870 (2003) (Nobel lecture).
11. Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M.W., Pain, A., Sauerwein, R.W., Eling, W.M.C., Hall, N., Waters, A.P., Stunnenberg, H.G. and Mann, M., Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature*, **419**, 537–542 (2002).
12. Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., Witney, A.A., Wolters, D., Wu, Y.M., Gardner, M.J., Holder, A.A., Sinden, R.E., Yates, J.R. and Carucci, D.J., A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*, **419**, 520–526 (2002).
13. Francis, A.W., Ruggiero, C.E., Koppisch, A.T., Dong, J.G., Song, J., Brettin, T. and Iyer, S., Proteomic analysis of *Bacillus anthracis* Sterne vegetative cells. *Biochim. Biophys. Acta-Proteins Proteomic.*, **1748**, 191–200 (2005).
14. Krokhin, O., Li, Y., Andonov, A., Feldmann, H., Flick, R., Jones, S., Stroeher, U., Bastien, N., Dasuri, K.V.N., Cheng, K.D., Simonsen, J.N., Perreault, H., Wilkins, J., Ens, W., Plummer, F. and Standing, K.G., Mass spectrometric characterization of proteins from the SARS virus. *Mol. Cell. Proteomic.*, **2**, 346–356 (2003).
15. Tan, Y.J., Lim, S.G. and Hong, W.J., Characterization of viral proteins encoded by the SARS-coronavirus genome. *Antiviral Res.*, **65**, 69–78 (2005).
16. Peter, R.J. and Michael, H. (Eds.), Proteomics of microbial pathogens. *Proteomics*, **4**(**10**), 2829–3106 (2004) (special issue).
17. Anhalt, J.P. and Fenselau, C., Identification of bacteria using mass spectrometry. *Anal. Chem.*, **47**, 219–225 (1975).
18. Fenselau, C. and Demirev, P., Characterization of intact microorganisms by MALDI mass spectrometry. *Mass Spectrom. Rev.*, **20**, 157–171 (2001).
19. Lay, J., MALDI-TOF mass spectrometry of bacteria. *Mass Spectrom. Rev.*, **20**, 172–194 (2001).
20. Demirev, P.A., Feldman, A.B. and Lin, J.S., Bioinformatics-based strategies for rapid microorganism identification by mass spectrometry. *Johns Hopkins APL Tech. Digest*, **25**, 27–37 (2004).
21. Graham Cooks, R. and Badman, E.R., Miniature mass analyzers. *J. Mass Spectrom.*, **35**, 659–671 (2000).
22. Bryden, W.A., Benson, R.C., Ecelberger, S.A., Phillips, T.E., Cotter, R.J. and Fenselau, C., The tiny-TOF mass-spectrometer for chemical and biological sensing. *Johns Hopkins APL Tech. Digest*, **16**, 296–310 (1995).
23. Cornish, T.J., Antoine, M.D., Ecelberger, S.A. and Demirev, P.A., Arrayed time-of-flight mass spectrometry for time-critical detection of hazardous agents. *Anal. Chem.*, **77**, 3954–3959 (2005).

24. von Wintzingerode, F., Bocker, S., Schlotelburg, C., Chiu, N.H.L., Storm, N., Jurinke, C., Cantor, C.R., Gobel, U.B. and van den Boom, D., Base-specific fragmentation of amplified 16S rRNA genes analyzed by mass spectrometry: a tool for rapid bacterial identification. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 7039–7044 (2002).
25. Hofstadler, S.A., Sampath, R., Blyn, L.B., Eshoo, M.W., Hall, T.A., Jiang, Y., Drader, J.J., Hannis, J.C., Sannes-Lowery, K.A., Cummins, L.L., Libby, B., Walcott, D.J., Schink, A., Massire, C., Ranken, R., Gutierrez, J., Manalili, S., Ivy, C., Melton, R., Levene, H., Barrett-Wilt, G., Li, F., Zapp, V., White, N., Samant, V., McNeil, J.A., Knize, D., Robbins, D., Rudnick, K., Desai, A., Moradi, E. and Ecker, D.J., TIGER: the universal biosensor. *Int. J. Mass Spectrom.*, **242**, 23–41 (2005).
26. Ecker, D.J., Sampath, R., Blyn, L.B., Eshoo, M.W., Ivy, C., Ecker, J.A., Libby, B., Samant, V., Sannes-Lowery, K.A., Melton, R.E., Russell, K., Freed, N., Barrozo, C., Wu, J.G., Rudnick, K., Desai, A., Moradi, E., Knize, D.J., Robbins, D.W., Hannis, J.C., Harrell, P.M., Massire, C., Hall, T.A., Jiang, Y., Ranken, R., Drader, J.J., White, N., McNeil, J.A., Crooke, S.T. and Hofstadler, S.A., Rapid identification and strain-typing of respiratory pathogens for epidemic surveillance. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 8012–8017 (2005).
27. Demirev, P., Ramirez, J. and Fenselau, C., Tandem mass spectrometry of intact proteins for characterization of biomarkers from Bacillus cereus T spores. *Anal. Chem.*, **73**, 5725–5731 (2001).
28. Hathout, Y., Setlow, B., Cabrera-Martinez, R.M., Fenselau, C. and Setlow P., Small, acid-soluble proteins as biomarkers in mass spectrometry analysis of *Bacillus* spores. *Appl. Environ. Microbiol.*, **69**, 1100–1107 (2003).
29. Hathout, Y., Ho, Y.P., Ryzhov, V., Demirev, P. and Fenselau, C., Kurstakins: A new class of lipopeptides isolated from Bacillus thuringiensis. *J. Nat. Prod.*, **63**, 1492–1496 (2000).
30. Pineda, F., Antoine, M., Demirev, P., Feldman, A., Jackman, J., Longenecker, M. and Lin, J., Microorganism identification by matrix-assisted laser/desorption ionization mass spectrometry and model-derived ribosomal protein biomarkers. *Anal. Chem.*, **75**, 3817–3822 (2003).
31. Jackson, K.A., Edwards-Jones, V., Sutton, C.W. and Fox, A.J., Optimisation of intact cell MALDI method for fingerprinting of methicillin-resistant *Staphylococcus aureus*. *J. Microbiol. Methods*, **62**, 273–284 (2005).
32. Wunschel, D.S., Hill, E.A., McLean, J.S., Jarman, K., Gorby, Y.A., Valentine, N. and Wahl, K., Effects of varied pH, growth rate and temperature using controlled fermentation and batch culture on matrix assisted laser desorption/ionization whole cell protein fingerprints. *J. Microbiol. Methods*, **62**, 259–271 (2005).
33. Demirev, P.A., Ho, Y.P., Ryzhov, V. and Fenselau, C., Microorganism identification by mass spectrometry and protein database searches. *Anal. Chem.*, **71**, 2732–2738 (1999).
34. Pineda, F., Lin, J.S., Fenselau, C. and Demirev, P., Testing the significance of microorganism identification by mass spectrometry and proteome database search. *Anal. Chem.*, **72**, 3739–3744 (2000).
35. Demirev, P., Lin, J., Pineda, F. and Fenselau, C., Bioinformatics and mass spectrometry for microorganism identification: Proteome-wide post-translational modifications and database search algorithms for characterization of intact H. pylori. *Anal. Chem.*, **73**, 4566–4573 (2001).
36. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L.Y., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W.V., Figeys, D. and Tyers, M., Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183 (2002).

37. Bogdanov, B. and Smith, R.D., Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom. Rev.*, **24**, 168–200 (2005).
38. Mortz, E., O'Connor, P., Roepstorff, P., Kelleher, N., Wood, T., McLafferty, F.W. and Mann, M., Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 8264–8267 (1996).
39. Kelleher, N.L., Lin, H.Y., Valaskovic, G.A., Aaserud, D.J., Fridriksson, E.K. and McLafferty, F.W., Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *J. Am. Chem. Soc.*, **121**, 806–812 (1999).
40. Cargile, B.J., McLuckey, S.A. and Stephenson, J.L., Identification of bacteriophage MS2 coat protein from E. coli lysates via ion trap collisional activation of intact protein ions. *Anal. Chem.*, **73**, 1277–1285 (2001).
41. Yao, Z., Demirev, P. and Fenselau, C., Mass spectrometry-based proteolytic mapping for rapid virus identification. *Anal. Chem.*, **74**, 2529–2534 (2002).
42. Warscheid, B. and Fenselau, C., Characterization of Bacillus spore species and their mixtures using postsource decay with a curved-field reflectron. *Anal. Chem.*, **75**, 5618–5627 (2003).
43. Warscheid, B., Jackson, K., Sutton, C. and Fenselau, C., MALDI analysis of Bacilli in spore mixtures by applying a quadrupole ion trap time-of-flight tandem mass spectrometer. *Anal. Chem.,* **75**, 5608–5617 (2003).
44. Pribil, P.A., Patton, E., Black, G., Doroshenko, V. and Fenselau, C., Rapid characterization of Bacillus spores targeting species-unique peptides produced with an atmospheric pressure matrix-assisted laser desorption/ionization source. *J. Mass Spectrom.*, **40**, 464–474 (2005).
45. Demirev, P., Feldman, A., Kowalski, P. and Lin, J., Top-down proteomics for rapid identification of intact microorganisms. *Anal. Chem.*, **77**, 7455–7461 (2005).
46. Demirev, P.A., Feldman, A.B., Kongkasuriyachai, D., Scholl, P., Sullivan Jr., D. and Kumar, N., Detection of malaria parasites by laser desorption mass spectrometry. *Anal. Chem.*, **74**, 3262–3266 (2002).
47. Demirev, P.A., Studies of malaria by mass spectrometry. In: Wilkins, Ch. and Lay, J. (Eds.), *Identification of Microorganisms by MS*. Wiley-Interscience, Hoboken, New Jersey, 2005, pp. 161–180.
48. Demirev, P.A., Mass spectrometry for malaria diagnosis. *Exp. Rev. Mol. Diagn.*, **4**, 821–829 (2004).
49. Scholl, P., Kongkasuriyachai, D., Demirev, P.A., Feldman, A., Lin, J., Sullivan, D. and Kumar, N., Rapid detection of malaria infection in vivo by laser desorption mass spectrometry. *Am. J. Trop. Med. Hyg.*, **71**, 546–551 (2004).
50. Nyunt, M., Pisciotta, J., Feldman, A., Thuma, Ph., Scholl, P., Demirev, P., Lin, J., Shi, L., Kumar, N. and Sullivan, D., Detection of *Plasmodium falciparum* in pregnancy by laser desorption mass spectrometry. *Am. J. Trop. Med. Hyg.*, **73**, 485–490 (2005).
51. Fergenson, D.P., Pitesky, M.E., Tobias, H.J., Steele, P.T., Czerwieniec, G.A., Russell, S.C., Lebrilla, C.B., Horn, J.M., Coffee, K.R., Srivastava, A., Pillai, S.P., Shih, M.T.P., Hall, H.L., Ramponi, A.J., Chang, J.T., Langlois, R.G., Estacio, P.L., Hadley, R.T., Frank, M. and Gard, E.E., Reagentless detection and classification of individual bioaerosol particles in seconds. *Anal. Chem.*, **76**, 373–378 (2004).
52. Tobias, H.J., Schafer, M.P., Pitesky, M., Fergenson, D.P., Horn, J., Frank, M. and Gard, E.E., Bioaerosol mass spectrometry for rapid detection of individual airborne *Mycobacterium tuberculosis* H37Ra particles. *Appl. Environ. Microbiol.*, **71**, 6086–6095 (2005).
53. Rees, J.C. and Voorhees, K.J., Simultaneous detection of two bacterial pathogens using bacteriophage amplification coupled with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, **19**, 2757–2761 (2005).
54. Mann, M., Mass tool for diagnosis. *Nature*, **418**, 731–732 (2002).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 15

Proteomics of viruses

ANNE PUMFERY[a], REEM BERRO[a], and FATAH KASHANCHI[a,b,*]

[a]*Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC 20037, USA*
[b]*The Institute for Genomic Research, Rockville, MD 20850, USA*

*Corresponding author. Address: *Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC 20037, USA*. Tel.: 202-994-1782; Fax: +1-202-994-1780. E-mail: bcmfxk@gwumc.edu.

1. Introduction

The term "proteome" was first coined in late 1994 by Marc Wilkins at the Siena two-dimensional gel electrophoresis (2DE) meeting and defines the entire protein complement in a given cell, tissue, or organism [1]. In its wider sense, proteomics research also assesses protein activities, modifications and localization, and interactions of proteins in complexes. It relies heavily on technology since it needs to identify proteins and protein complexes in biological samples comprehensively and quantitatively with both high sensitivity and fidelity.

Proteomics is a promising approach for the study of viruses. It allows a better understanding of disease processes and the development of new biomarkers for diagnosis and early detection of disease, and accelerates drug development. Areas of proteomics that are particularly promising include the determination of altered protein expression, not only at the whole-cell or tissue level, but also in subcellular structures and biological fluids; the development of novel biomarkers for diagnosis and early detection of disease; and the identification of new targets for therapeutics and the potential for accelerating drug development through more effective strategies to evaluate therapeutic effect and toxicity.

There is a growing interest in applying proteomics to the study of infectious disease. A complicating factor in therapy for infectious disease is the development of resistance to commonly used drugs, which heightens the need for developing effective new therapies. The availability of the complete sequences of a number of viruses has provided a framework for identifying proteins encoded in these genomes using mass spectrometry (MS). Applying proteomics to the study of viruses allows the characterization of subviral proteomes (e.g., secreted proteins, surface proteins, and immunogenic proteins), the comparative analysis of different stains or physiological states, the identification of proteins related to pathogenicity and host–pathogen interactions, and the evaluation of mechanisms of action of antiviral therapies.

1.1. Highlights for medical professionals

Viral infections cause significant morbidity and disease including cancer, immunosuppression, and death. Often infections are not diagnosed until symptoms appear and, in several cases, this may be years or decades after the initial infection. The ability to diagnose infection or cancer before the appearance of symptoms would be of critical importance for effective treatment. Proteomic analysis of serum has been proposed as a means of diagnosing infectious disease and/or the early diagnosis of cancer. There have been some recent exciting findings in the proteomics of the host or pathogen, and the use of standard mass spectrometric technologies has enabled many physicians and scientists to examine more closely the pathological and biological questions that can only be answered using proteomic approaches.

Therefore, in this chapter we will discuss some recent findings on the proteomics of DNA and RNA viral infections that are associated with clinically important diseases in humans, including human cytomegalovirus (HCMV), herpes simplex virus (HSV), Epstein–Barr virus (EBV), human immunodeficiency virus (HIV), hepatitis B and C (HBV and HCV, respectively), and adenovirus, as well as the coronavirus that causes severe acute respiratory syndrome (SARS).

HCMV is the largest member of the human herpesviruses. After initial infection, HCMV remains in a persistent state with the host [2]. Immunity against the virus controls replication, although intermittent viral shedding can still take place in the seropositive immunocompetent person [2]. As replication of cytomegalovirus in the absence of an effective immune response is central to the pathogenesis of disease, complications are primarily seen in individuals whose immune system is immature or suppressed by drug treatment or coinfection with other pathogens [3]. Estimates of the coding capacity of HCMV range from 160 open reading frames (ORFs) to more than 200 ORFs [4]. Recent studies using MS to determine the viral proteome suggest that the number of viral proteins may be even greater than previous estimates [5]. Analysis of proteins from purified HCMV virion preparations has indicated that the particle contains significantly more viral proteins than the previously known 71 HCMV virion proteins. Twelve of the identified proteins were encoded by known viral ORFs previously not associated with virions, and 12 proteins were from novel viral ORFs [6]. Therefore, new protein markers including HCMV tegument and various cellular structural proteins, enzymes, and chaperones are now serving as biomarkers for HCMV infection and as possible drug targets.

Other herpesvirus members have also been explored for the presence of possible biomarkers. EBV is a ubiquitous member of the herpesvirus family that is associated with a variety of lymphomas and lymphoproliferative diseases [7]. It encodes a multitude of genes that drive proliferation or confer resistance to cell death [8]. Infection of human B lymphocytes with EBV induces proliferative B-lymphoblastoid cell lines (LCLs). Recently, proteomic profiles of three LCLs were analyzed comparatively at the early and the late passages of cell culture. The phosphoprotein stathmin was identified, and expression significantly decreased with immortalization of LCLs [9]. Stathmin is critically important not only for the formation of a normal mitotic spindle upon entry into mitosis but also for the regulation of the function of the mitotic spindle in the later stages of mitosis and for the timely exit from mitosis [9]. In another study using standard matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) methods, 20 EBNA2 target proteins were identified, 11 of which were c-*myc* dependent and therefore most probably associated with proliferation of the host cell [10]. These findings further stress the role of EBV viral proteins, namely EBNA and LMP, in disease pathogenesis. Interestingly, when EBV-infected cells were treated with the drug 5′-azacytidine (AZC)—a demethylating agent that induces the expression of silenced genes, i.e., the p16 tumor suppressor gene—21 polypeptides were down-regulated, while

14 showed increased expression. Many of the induced proteins were involved in energy metabolism, organization of cytoskeletal structures, protein synthesis, or cell viability [11]. Therefore, the effect of drugs that activate silenced tumor suppressor genes and their proteomic profile following treatment is of considerable interest.

Finally, an important herpesvirus to consider is HSV. HSV types 1 and 2 are ubiquitous viruses that cause infections in human populations throughout the world. The clinical manifestations of HSV infections are varied, ranging from asymptomatic to life-threatening illness in neonates and immunocompromised hosts [12,13]. HSV-1 infection induces severe alterations of the host translational apparatus, including the phosphorylation of a few ribosomal proteins and the progressive association of several nonribosomal proteins to ribosomes [14–16]. Using a proteomics approach, it was shown that VP19C, VP26, and ICP27 associated with ribosomal proteins [17]. Specifically, immediate early ICP27 protein associated with the cellular translation initiation factor poly A binding protein (PABP), eukaryotic initiation factor 3 (eIF3), and eukaryotic initiation factor 4G (eIF4G) in infected cells, resulting in the stimulation of translation of certain viral mRNAs and inhibiting host mRNA translation [18]. Another study has shown that approximately 50 cellular and viral proteins associate with the HSV-1 ICP8 single-stranded DNA-binding protein, some of which belong to DNA repair and chromatin family members [19], implying that HSV-1 infection results in control of host cellular DNA replication/repair and gene expression machineries.

Proteomic analyses of RNA viruses with regard to diagnosis and novel biomarker detection are also of considerable interest in the medical community. For instance, SARS is a new infectious disease that first emerged in the Guangdong province, China, in November 2002 [20]. A novel coronavirus was later identified in patients with SARS. The detection of the virus in these patients, its absence in healthy controls or other patients with atypical pneumonia, and the reproduction of a similar disease in a relevant animal model indicated that this coronavirus was the causative agent of SARS (SARS-CoV) [21]. Interestingly, the full genome sequence was determined within weeks of the identification of the virus, but the proteome and biomarkers associated with SARS are slowly forthcoming. In a recent study using a mass spectrometric decision tree classification algorithm, Kang et al. identified four biomarkers determined in the training set that could precisely detect 36 of 37 (sensitivity, 97.3%) acute SARS and 987 of 993 (specificity, 99.4%) non-SARS samples [22]. A reasonably complete proteomic analysis was also performed on four patients with SARS at different times of infection, and a total of 38 differential spots were selected for protein identification. Most of the proteins identified were acute phase proteins, and their presence represented the consequence of a serial cascade of inflammatory reactions initiated by SARS-CoV infection. Of significance was the level of plasma peroxiredoxin II in patients with SARS, which was significantly higher in SARS patients and could be secreted by

T cells [23]. Finally, while pursuing new enzyme targets, another study identified 14 putative ORFs, 12 of which were predicted to be expressed from a nested set of eight subgenomic mRNAs. Distant homologs of cellular RNA processing enzymes were identified in group 2 coronaviruses, with four of them being conserved in SARS-CoV. These newly recognized viral enzymes put the mechanism of coronavirus RNA synthesis in a completely new perspective, which has opened the door for new drug targets for the treatment of SARS [24].

The proteomes of three other RNA viruses including HCV, HBV, and HIV-1 have also been studied. Hepatitis C often progresses to chronic infection in the majority of patients and is an emerging cause of viral hepatitis. Clinically, the infection is generally asymptomatic, but may present with a wide variety of symptoms. Cirrhosis, hepatocellular carcinoma (HCC), cryoglobulinemia, autoantibodies, and glomerulonephritis have been strongly associated with HCV infection [25]. When analyzing proteins that interacted with the HCV protein NS5A, Choi et al. found that the cytoplasmic heat shock protein 27 (HSP27) bound to NS5A was concentrated in the ER [26], where drugs for HCV treatment would have easy access (as opposed to drugs delivered into the nucleus). Chronic infection with HBV is associated with the majority of HCC. Using woodchucks as a model system, HCC induced dramatically higher levels of serum-associated core alpha-1,6-linked fucose, as compared with woodchucks without HCC. The coupling of this model system with 2D gel electrophoresis has permitted the identification of several glycoproteins with altered glycosylation as correlated to cancer prognosis. One such glycoprotein, the golgi protein 73 (GP73), was found to be elevated and hyperfucosylated in animals with HCC [27]. Finally, in an effort to identify useful biomarkers for HBV- or HCV-associated HCC, 60 proteins were identified which exhibited significant changes in expression between nontumorous and tumorous tissues. Among these, 14 proteins were commonly changed in all three of the HCC types, but 46 proteins showed a tendency toward viral marker specificity, suggesting that the pathogenic mechanisms of hepatocarcinogenesis may be different according to the viral etiology of HBV or HCV [28].

Diagnosis and treatment strategies for HCV have become extremely important as one-third of HIV-infected individuals in Europe and the USA are coinfected with HCV [29]. Therefore, defining biomarkers in coinfections after highly active antiretroviral therapy (HAART) is currently the focus of many laboratories. HIV accelerates HCV liver disease especially with the progression of HIV-associated immunodeficiency. With the introduction of pegylated interferon in combination with ribavirin, greatly improved treatment options for patients coinfected with HIV and HCV have become available and have led to sustained virological response rates of up to 40% [30]. Furthermore, recent cohort analyses have shown that immune reconstitution induced by HAART can improve the course of hepatitis C infection leading to a decline in liver-related mortality. However, patients with HCV coinfection are at increased risk of hepatotoxicity from HAART [29].

Owing to the high rates of HIV and HCV coinfections worldwide, new improved biomarkers and treatment strategies and guidelines for the management of coinfection remain a major goal. Biomarkers could include protein fingerprints of HIV-1-infected human monocyte-derived macrophages (MDMs) after viral infection, as well as HCV-infected liver cells. Recently, 58 proteins have already been identified to be up- or down-regulated after HIV-1 infection [31].

1.2. Highlights for chemists

The overall awareness of the importance of proteins and peptides in physiology and pathophysiology has increased dramatically over the last few years. With progress in the analysis of whole genomes, the knowledge base in gene sequence and expression data, useful for protein and peptide analysis, has increased considerably. Therefore, the medical need for relevant biomarkers is enormous. This is particularly true for many viral infections and various types of cancer, where there is a lack of useful and adequate diagnostic markers with high specificity and sensitivity.

However, proteomic and peptide-based techniques have evolved in recent years to simplify the search for biomarkers. Peptide-based technologies provide new opportunities for the detection of low-molecular-weight protein biomarkers (peptides) by MS. Improvements in peptide-based research are based on separation of peptides and/or proteins by their physicochemical properties in combination with mass spectrometric detection and identification using sophisticated bioinformatics tools for data analysis. Therefore, peptide-based technologies offer an opportunity to discover novel biomarkers for diagnosis and management of disease including prognosis, treatment decision, and monitoring response to therapy.

There are a number of critical viral infections that have dominated the research and biomarker landscape. Many of these findings rely on somewhat simple or "off the shelf" technologies that are fairly straightforward to use. Perhaps the simplest of these technologies is the surface-enhanced laser desorption/ionization (SELDI) technology. In a study for SARS detection, Kang et al. developed a mass spectrometric decision tree classification algorithm using SELDI-TOF MS. Serum samples were grouped into acute SARS and non-SARS and healthy control cohorts. Diluted samples were applied to WCX-2 ProteinChip arrays (Ciphergen), and the bound proteins were assessed on a ProteinChip Reader (Model PBS II). The results clearly indicated an impressive accuracy for discriminatory classifiers [22]. Another similar study indicated that nine serum markers significantly increased and three significantly decreased in SARS patients as compared to controls [32].

Another ProteinChip assay used to study HIV-1 infection showed a unique MDM protein fingerprint during HIV-associated dementia (HAD) and HAART. Seven unique protein peaks between 3.0 and 20.0 kDa were found in the HAD MDM samples, all of which were abrogated after HAART [33]. A very similar study using specific proteins produced from monocytes from HAD patients

showed a total of 177 protein peaks from 2 to 80 kDa in 31 MDM lysates. Select protein peaks, at 5028 and 4320 Da, separated HIV-1-infected from HIV-1-seronegative subjects with 100% sensitivity and 80% specificity [34].

However, most viral proteomics studies to date have utilized either 2DE and MALDI-TOF MS or LC/MS/MS. The HIV virion is composed of a lipid bilayer that surrounds the viral capsid (Fig. 1A). In a clever study, Fuchigami et al. [35] studied the HIV-1(LAV-1) particles, which were collected by ultracentrifugation, treated with subtilisin, and then purified by Sepharose CL-4B column chromatography to remove microvesicles. The lysate of the purified HIV-1 particles was subjected to 2DE and stained, and the stained spots were excised and digested with trypsin. The resulting peptide fragments were characterized by MALDI-TOF MS. Twenty-five proteins were identified as proteins inside the virion, and the acid-labile formyl group of an amino terminal proline residue of HIV-1(LAV-1) p24(gag) was determined by MALDI-TOF MS before and after weak-acid treatments (0.6 N hydrochloric acid) and confirmed by postsource decay (PSD) of the *N*-formylated N-terminal tryptic peptide (*N*-formylated Pro(1)-Arg(18)). Interestingly, formylation plays a critical role in the formation of the HIV-1 core for conferring HIV-1 infectivity [35].

More recently, the use of liquid chromatography and tandem MS (LC/MS/MS) has also eased purification and recovery methods. For instance, Varnum and colleagues utilized gel-free two-dimensional capillary LC/MS/MS and Fourier transform ion cyclotron resonance MS to identify and determine the relative abundances of viral and cellular proteins in purified HCMV virions and dense bodies. Analysis of the proteins from purified HCMV virion preparations has indicated that the particle contains significantly more viral proteins than previously known. They identified more than 71 HCMV-encoded proteins and 70 host cellular proteins in HCMV virions, which included cellular structural proteins, enzymes, and chaperones [6]. Another study using LC/MS/MS for the adenovirus type 5 proteome found a total of 11 protein species from 154 peptides, at a sensitivity of 10 copies per virus and a detection limit of 70 fmol for two proteins [36].

Two new methods have been used recently to decipher viral proteomes. A method for proteolytic stable isotope labeling was recently used to provide quantitative and concurrent comparisons between individual proteins from two different proteome pools or their subfractions. Using this technique two ^{18}O atoms were incorporated universally into the carboxyl termini of all tryptic peptides during the proteolytic cleavage of proteins in the first pool. Proteins in the second pool were analogously cleaved with the carboxyl termini of the resulting peptides containing two ^{16}O atoms (i.e., no labeling). The method was used to compare two virus strains, adenovirus types 2 and 5. This shotgun approach for proteomic studies with quantitative capability may be a very powerful tool for comparative proteomic studies of very complex protein mixtures [37]. Finally, the isotope-coded affinity tag (ICAT) procedure has also yielded some very interesting results for a

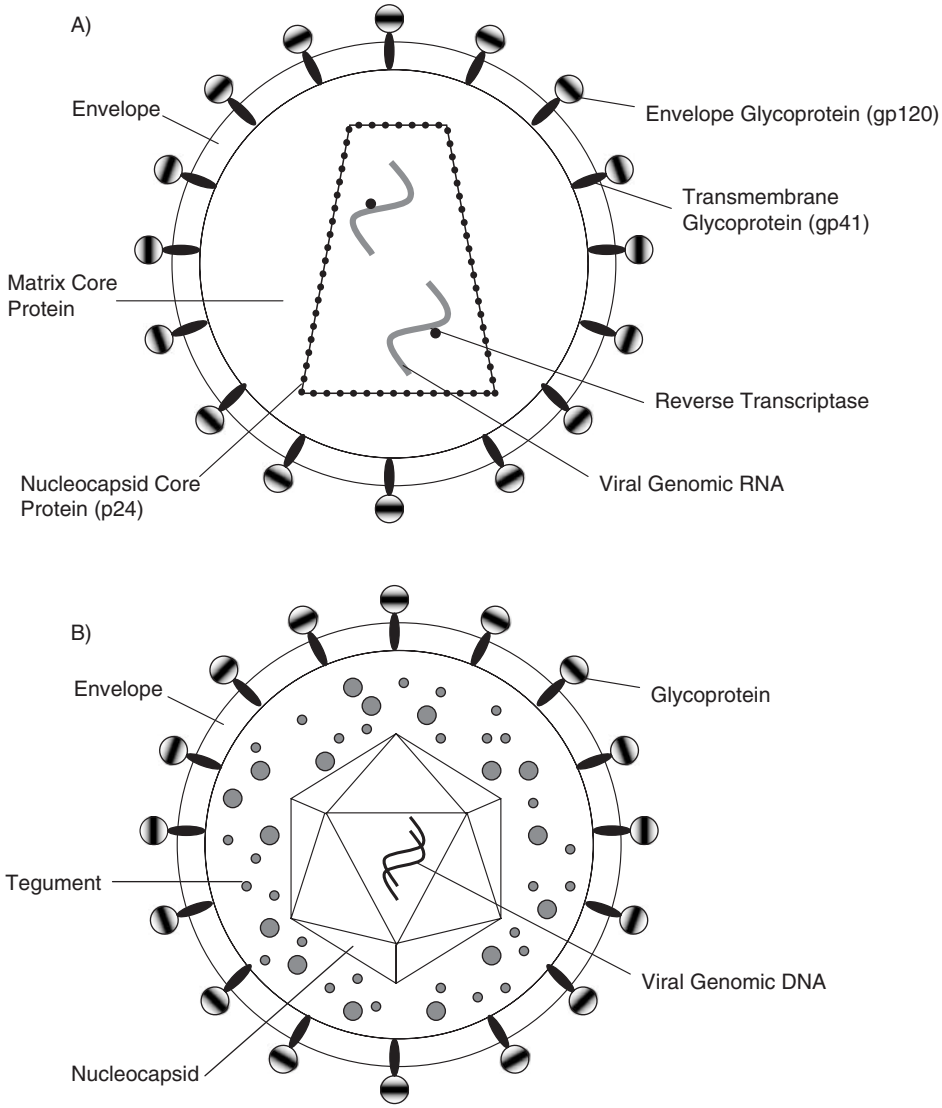


Fig. 1. *Virion structures.* (A) *HIV virion*. The HIV virion is composed of a lipid bilayer membrane (envelope) that surrounds the capsid. Two viral glycoproteins (gp120 and gp41) are part of the envelope and are important for viral binding and entry. The capsid is composed of the matrix core and nucleocapsid (p24) core proteins and surrounds two copies of the viral genomic RNA and reverse transcriptase. (B) *Herpesvirus virion*. The herpesvirus virion is composed of a lipid bilayer membrane (envelope) that surrounds the tegument and the capsid. Viral glycoproteins required for binding to and entering the host cell are imbedded into the envelope. The tegument is an amorphous proteinaceous structure that contains a variety of viral and cellular proteins. The herpesvirus capsid is an icosahedron of 150 hexons and 12 pentons that surrounds the double-stranded DNA genome.

rather complex viral infection setting. For instance, proteins from human liver carcinoma cells, representing transformed liver cells, and cultured primary human fetal hepatocytes (HFH) were extracted and processed for ICAT chromatography. Proteins from hepatitis C virus-infected cells and corresponding control cells were labeled with light and heavy cleavable ICAT reagents, respectively. After the labeled samples were combined, trypsinized, and subjected to cation-exchange and avidin-affinity chromatographies, the resulting cysteine-containing peptides were analyzed by microcapillary LC/MS/MS. Using SEQUEST and other bioinformatics software a total of ~1500 proteins or related protein groups were identified in three subdatasets from uninfected and infected cells [38]. Collectively, these results further emphasize the new targets for biomarkers and drug development for HCV infection.

The described new technologies have collectively added to our arsenal of possible biomarkers when diagnosing various viral infections. However, many of these markers still need to be validated using more rigorous sample methods, biological and biochemical tests, and more sophisticated bioinformatics tools. Bioinformatics tools that have been valuable for viral diagnosis and fast retrieval of DNA or protein sequences include the ORFer program (http://www.proteinstrukturfabrik.de/orfer), 2D proteome database (http://proteome.btc.nus.edu.sg/hccm), the Poxvirus proteomics database (http://contact14.ics.uci.edu/virus/vaccinia.php), and VirGen (http://bioinfo.ernet.in/virgen/virgen.html).

In this chapter, we will explore the importance of proteomics in studying virus–host interactions in several viral systems including HCMV, KSHV, EBV, HSV, HIV-1, HTLV-1, and HCV. We will also describe the methods that have been employed to study viral disease progression using several techniques including 2DE, LC–MS/MS, SELDI, and protein microarrays.

2. Virus–host interactions

Viral proteomics has included the analysis of viral particles to determine all proteins—viral and cellular—that compose the infectious virus, the examination of cellular proteins associated with a single viral protein in the hopes of determining all the functions of that viral protein, or the determination of cellular proteins induced or altered during a particular disease state. Identification of viral proteins requires that the viral genome has been fully sequenced and potential ORFs have been identified. Presently, over 1200 different viral genomes have been sequenced, annotated, and deposited in public sequence databases (GenBank, EMBL, and DDBJ) [39]. Additionally, the National Center for Biotechnology Information (NCBI) has established a Viral Genomes Project to provide standards for viral genomic research [39]. This resource will further the research of virus proteomics. The viral proteome of several herpesviruses, hepatitis C virus, human

T-lymphotropic virus (HTLV), and the HIV have been analyzed and will be reviewed here.

2.1. Proteomics of herpesvirus virions

Viral particles of HCMV and Kaposi's sarcoma-associated herpesvirus/human herpesvirus 8 (KSHV/HHV-8) have recently been examined. During the herpesviral replicative cycle, different viral particles are formed. For HCMV, this includes mature, infectious virions, noninfectious enveloped particles, and dense bodies [6]. Similarly for KSHV, only a portion of the produced virus particles is infectious [40]. Therefore, analysis of infectious virions requires their separation from the noninfectious and immature forms. Density ultacentrifugation gradients are typically used to separate the various forms. Each fraction can be analyzed by electron microscopy to determine the level of purity [6,41] or by assaying for viral DNA and an envelope glycoprotein [40].

The herpesviruses are large enveloped DNA viruses (Fig. 1B). The viral particle consists of a lipid envelope, in which are embedded viral glycoproteins important for infection of target cells. The envelope surrounds an amorphous proteinaceous structure called the tegument [42]. The tegument is often composed of viral proteins critical for the initiation of viral gene expression, for example the VP16 protein of HSV [43], as well as other viral and cellular proteins whose functions are unknown. The tegument surrounds the viral capsid, which is composed of a major capsid protein, one or more minor capsid proteins, and viral DNA. Identification of tegument and capsid proteins can be differentiated from the envelope glycoproteins by their differential sensitivity to trypsin and detergents. The tegument and capsid proteins are resistant to trypsin digestion in the absence of detergents. The envelope glycoproteins, however, are sensitive to trypsin digestion whether or not detergents are present. However, only the surface-exposed portions of glycoproteins are sensitive to trypsin in the absence of detergents.

2.1.1. Identification of proteins in HCMV particles

Following gradient purification of virions, LC/MS/MS was used to identify the components of the HCMV virion [6]. The results were verified by coupling high-accuracy mass measurements with LC and FT-ICR (Fourier transform ion cyclotron resonance) MS. Fifty-nine proteins were identified including 12 proteins encoded by known HCMV ORFs not previously known to reside in virions. The classes of proteins identified included capsid proteins, tegument proteins, glycoproteins, and 12 proteins involved in DNA replication and transcription. Additionally, 12 more viral polypeptides were identified that had not been previously characterized [6].

Using the intensities in the FT-ICR spectra, the relative quantities of the virion proteins were determined, indicating that 50% of the virion was composed of

tegument proteins, 30% were capsid proteins, 13% were envelope proteins, while 7% were undefined proteins. These undefined proteins are likely to be cellular proteins that are incorporated into the virion. Host cellular proteins were detected by comparison with peptides predicted from a human-FASTA database. There were 71 cellular proteins identified to be associated with the HCMV virion. They included cytoskeletal proteins, proteins involved in translation control, and several signal transduction proteins [6]. The identification of cellular proteins involved in translation and signal transduction as components of the HCMV virion suggests that these proteins may have a function in the initiation of viral gene expression or inducing an environment that is suitable for HCMV infection.

2.1.2. Identification of proteins in KSHV particles

KSHV has only been fully sequenced in the last 10 years [44], and therefore not much is known about the composition of the virus particle. Nealon et al. [41] used SDS-polyacrylamide gel electrophoresis (PAGE) and Western blotting to identify the major capsid and scaffolding proteins as components of isolated virions. They then used ion trap MS to identify three additional components of the virion as ORFs 62, 26, and 65. This study, however, was limited and unable to identify all components of the infectious virus. Zhu et al. used a more comprehensive approach to identify virion components. Extracellular virions were purified by double gradient centrifugation. SDS-PAGE revealed 30–40 protein bands whose identity was determined by in-gel trypsin digestion, followed by LC and MS. Both peptide masses and peptide sequences were produced by tandem MS (MS/MS) and used to determine protein identity. The isolated proteins included five capsid proteins, eight glycoproteins, six tegument proteins, and five other KSHV ORFs. Twenty cellular proteins were also identified and, as seen with HCMV, these included cytoskeletal proteins, signal transduction proteins, as well as heat shock proteins [45]. Similar results were seen in the study by Bechtel et al. [40]. However, fewer proteins were identified in this study as a single 7.5% SDS-PAGE gel was used to separate virion proteins. Zhu et al. [45] used three SDS-PAGE gel—a 4–12% gel, a 3–8% gel to separate proteins larger than 50 kDa, and a 12% gel to separate proteins smaller than 50 kDa. These studies underscore the need for good separation methods to be able to identify all proteins in a virus particle.

2.2. Proteomics of Epstein–Barr virus

The Epstein–Barr virus (EBV) is a B-cell lymphotropic herpesvirus that induces a usually asymptomatic infection and is carried by more than 90% of adults. However, EBV is the causative agent for Burkitt's lymphoma and nasopharyngeal carcinoma and is involved in a number of acquired immunodeficiency syndrome (AIDS)-associated lymphomas. EBV can induce immortalization of B cells *in vitro* to generate lymphoblastoid cell lines (LCLs), a model for the carcinogenic potential

of EBV. LCLs are latently infected with EBV; they maintain the virus as an extrachromosomal episome, and have limited viral gene expression. The latently expressed proteins are the six EBV nuclear antigens (EBNAs 1, 2, 3A, 3B, 3C, and -LP) and three latent membrane proteins (LMPs 1, 2A, and 2B). Of the latently expressed proteins, EBNA2 and LMP are required for transformation induced by EBV [46]. Characterizations of cellular proteins associated with EBNA2 or proteins differentially expressed in the early stages of the transformation process will be described in the following text. Results of these studies may lead to a better understanding of EBV-mediated transformation and the identification of cellular targets for therapy.

A proteome database of LCLs, before and after transformation, has been developed to identify the cellular mechanisms of virus-induced immortalization [9,47]. 2DE was used to first separate proteins based on their relative charge (pI) and then based on their molecular weight. Differentially expressed proteins were digested and subjected to electrospray ionization MS. Proteins were identified based on their peptide mass fingerprint and amino acid sequences of peptides determined by Edman degradation. There were 32 differentially expressed proteins and 20 were assigned to known proteins. The expression of several proteins involved in proliferation or nucleotide metabolism was increased in the immortalized cells, which may result in the growth stimulation seen in immortalized cells. A database of 2D gel images as well as the identity of the differentially expressed proteins has been made available to the public at www.proteome.jp/2D/. The availability of these images and the identification of the differentially expressed proteins may prove useful to others in their analysis of EBV-infected cells.

EBNA2 is required for transformation of LCLs by EBV [48,49] and induces the expression of c-*myc* [50]. c-*myc* is an oncogene which drives cell proliferation; however, the proliferation program induced by c-*myc* is different than that observed by the expression of EBNA2 [51], suggesting that other cellular proteins and events are induced by EBNA2 to mediate transformation. Furthermore, there is limited information on the cellular targets of EBNA2. To identify the EBNA2-induced changes as c-myc dependent and c-myc independent, EBNA2 and c-*myc*-conditionally expressing cell lines were used and the proteome of each cell line was compared [10]. Proteins were separated by 2D SDS-PAGE and identified by MALDI-TOF MS. In EBNA2 expressing cells, there were 20 differentially expressed proteins; 12 were induced and 8 were repressed. Of the proteins that were induced following EBNA2 expression, several were involved in nucleotide metabolism, protein synthesis, or the control of apoptosis. Many of these proteins were also induced in c-*myc* expressing cells, though six proteins were found to be EBNA2 specific. Two EBNA2-specific proteins were induced (Bid and IgE-HRF) and four were repressed (Annexin IV, γ-actin, GMFγ, and AF103803). Furthermore, analysis of the activation kinetics demonstrated that expression of EBNA2 preceded that of c-*myc*, which was then followed by the expression of Nm23-H1

(a nucleotide diphosphate kinase that may suppress metastasis [52]), indicating that c-*myc* was a direct target of EBNA2 [10].

2.3. Proteomics of herpes simplex virus

Herpes simplex virus (HSV) has been the most extensively studied of the human herpesviruses owing to its ability to easily infect cells *in vitro* to produce infectious virus. As with all herpesviruses, HSV encodes a number of proteins for efficient viral gene expression, viral DNA replication, and the shutoff of cellular gene transcription and translation [53]. These virally expressed proteins do not function in isolation but associate with a variety of cellular and viral proteins. Furthermore, many have exhibited multiple different functions. In an effort to understand the biology of HSV and the function of its proteins, a proteomics approach has been used to study a critical viral transactivator (ICP27), the alteration of the cellular translation machinery, and components of the viral replication complex, which will be reviewed here.

The ICP27 protein is expressed early in infection and is essential for viral replication and expression of certain early genes and virtually all late genes. It is a multifunctional protein that may function with the virion host shutoff (vhs) protein of HSV to repress cellular protein synthesis. This repression serves to direct cellular resources to the synthesis of viral proteins. Using immunoprecipitation of ICP27 from HSV-infected cells followed by SDS-PAGE and MS, several translation initiation factors were identified, including PABP, eIF3, and eIF4G [18]. The interaction of ICP27 with translation initiation factors may recruit these factors to viral mRNA to facilitate translation of viral mRNAs and also to sequester these factors away from the translation of cellular mRNAs.

HSV infection also induces ribosomal changes and it has been hypothesized that these changes may contribute to HSV-mediated translational control of host and viral gene expression [17]. To identify the changes in ribosomes following HSV infection, ribosomes were purified by ultracentrifugation, the proteins were separated by 2D SDS-PAGE, and their identities were determined by MALDI-TOF MS. Seven additional protein spots were found associated with ribosomes following HSV infection, including several viral proteins: VP19C and VP26—components of the viral capsid, and US11—a tegument protein. Three of the seven spots were phosphorylated forms of US11. One nonribosomal protein, PABP, was also found associated with ribosomes. The association of PABP with ribosomes increased following HSV infection [17]. Interestingly, PABP was also found associated with ICP27 in the previous study [18]. Although Greco et al. [17] did not find ICP27 associated with ribosomes, this is likely due to the different separation procedures used. Greco et al. [17] used 2DE and focused solely on basic proteins with a p*I* greater than 8.6 while Fontaine-Rodriguez et al. [18] separated isolated proteins based on molecular weight. Together, these results suggest that ICP27 likely associates with ribosomes in infected cells.

Herpesvirus DNA replication occurs in intranuclear structures called replication compartments [54,55]. HSV encodes seven proteins that participate in viral DNA replication; however, it is not known what cellular proteins are involved in this process. To identify cellular proteins in HSV replication compartments, ICP8, the HSV single-stranded DNA-binding protein, was immunoprecipitated from infected cells and coprecipitating proteins were separated by SDS-PAGE and identified by ion trap MS [19]. Greater than 50 viral and cellular proteins were identified as copurifying with ICP8. The cellular proteins included those that participate in DNA replication/repair/recombination, chromatin remodeling, RNA binding/splicing, and transcription factors. Several of these proteins require DNA binding to associate with ICP8, including several chromatin-remodeling proteins. The roles of a number of interacting cellular proteins are presently unclear and further studies are needed to determine their exact roles in viral DNA replication.

2.4. Proteomics of retroviruses—HIV and HTLV

HIV encodes a critical transcriptional activator, Tat, which directs a cellular transcription factor, pTEFb, to the HIV LTR to mediate transcription elongation [56,57]. However, it has been shown that the viral genome is bound by nucleosomes that inhibit viral gene expression [58,59]. To determine if Tat interacts with additional cellular proteins to further assist viral gene expression, we used Tat peptides linked to biotin to pull down all Tat-associated proteins [60]. Additionally, acetylated and unmodified peptides were also used because acetylation of Tat has been attributed to alternative functions of Tat [61,62]. We found that many more cellular proteins bound to the unmodified Tat, including proteins involved in modification of chromatin structure (CHD2 and p/CAF) and additional transcription factors (TIF1—a TRIM family member—and SCL—a bHLH transcription factor) [60]. These results indicate that Tat influences viral gene expression at various levels and suggests that targeting these specific interactions may be a viable form of treatment of HIV infection and AIDS.

HIV infects several cell types during the course of infection and progression to AIDS. In HIV-infected patients, the virus establishes a persistent infection in cells of the monocyte/macrophage lineage. Monocytes and macrophages are the first line of defense in the immune system: they phagocytose and kill a range of microorganisms. However, little is know about how HIV persists in these cells. To understand how HIV may persist in these cells, Carlson et al. [31] used a “ProteinChip” and SELDI to identify unique protein signatures in HIV-infected monocytes obtained from different donors. Infection of monocytes isolated from humans was used to mimic the virus–host interactions that would occur in an infected individual. The ProteinChips used in the study were used to partially purify samples. One is a weak cation exchange, and the second is a reverse-phase hydrophobic interaction chip. Charged proteins will bind to the cation ion-exchange chip while hydrophobic, i.e.,

membrane-associated, proteins will bind to the reverse-phase chip. Proteins bound to the chip were then analyzed by MS. Each peak represents a protein of a particular mass; however, the nature of the protein in a peak after SELDI MS is unknown. A different, separation technology and MS are needed to determine protein identities. To determine the identities of proteins up- or down-regulated following HIV infection of monocytes, total protein extracts were subjected to trypsin digestion, LC, and tandem MS to determine sequences of tryptic peptides. Sequences were then analyzed against a Protein Bank to determine the identities of the proteins and given a score [31]. The problem with this type of study is that there is no quantitative assessment of the increase or decrease in the protein levels or changes in posttranslational modifications (PTM) following infection.

HAD affects almost one-third of adults infected with HIV [63]. The exact cause of dementia is not known. There is significant neuronal loss but neurons are not infected with HIV [64]. It has been hypothesized that HIV-infected astrocytes are critical in the development of HIV dementia, and that Tat is a contributor to this disease. Extracellular Tat released from astrocytes induces cell death in neurons, though Tat protects astrocytes from cell death [65]. To understand this dichotomy, proteins differentially expressed in Tat-expressing astrocytes were identified [66]. Total protein extracts of Tat-expressing and control cells were separated by 2D gel electrophoresis and identified by MALDI MS. Interestingly, seven proteins were found to be repressed in Tat astrocytes, including Rho GDP dissociation inhibitor and protein phosphatase 2A (PP2A) inhibitor. Many of these proteins have been shown to be involved in the biology of HIV and interact with Tat [66]. Three proteins identified by a slot blot technique were found to be induced, and included HSP70, heme oxygenase, and inducible nitric oxide synthase (iNOS) [66]. Previously published data have demonstrated a correlation between iNOS and the severity of HAD [67,68]; however, the role of the other differentially expressed proteins in astrocyte survival and HIV dementia will require further study.

The human T-cell leukemia virus type 1 (HTLV-1) causes adult T-cell leukemia (ATL) and HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) [69]. HTLV-1 encodes a transactivator, Tax, that is critical for virus replication and plays a central role in the development of ATL and HAM/TSP [69]. Tax does not bind to DNA directly but functions by interacting with a variety of cellular proteins [69]. Many protein–protein interactions of Tax have been determined by mutational analysis including CREB [70–72] and NF-κB [70,72]. To identify all the cellular proteins that interact with Tax, Wu et al. used chromatography, 2D gel electrophoresis, and mass spectrometric analysis of an HTLV-1-infected cell line (C81) [73]. As Tax functions in both the cytoplasm and the nucleus [70,74], Tax-interacting proteins were identified from both cellular compartments. Some of the cytoplasmic proteins included small GTPases and components of the cytoskeleton while some of the more interesting nuclear proteins included components of the SWI/SNF chromatin remodeling complex [73].

The interaction of Tax with many of the identified cellular proteins may be involved in the ability of Tax to dysregulate cellular functions leading to T-cell transformation and leukemia.

2.5. Proteomics of hepatitis C virus and hepatocellular carcinoma

Hepatocellular carcinoma (HCC) causes approximately one million deaths a year [75]. Two viruses are the main causes of HCC: HBV and HCV [76–78]. Although HBV is the most important cause of HCC, accounting for 80% of HCC cases, an effective vaccine is available [79]. HCV, however, is a major cause of the increasing incidence of HCC in developed countries [80] and no effective vaccine is available. HCC progresses after decades of chronic infection and often is at an advanced stage once it presents clinically [81]. As such, good noninvasive diagnostic markers are needed. This will be discussed further in Section 3. The focus of this section will be on the identification of cellular proteins that are induced following HCV infection or cellular proteins that interact with HCV proteins.

An extensive study by Wirth et al. [82] analyzed normal liver tissue and hepatoma-derived cell lines by 2D gel electrophoresis and N-terminal sequencing and identified a number of proteins that were differentially expressed between normal tissue and hepatoma cell lines. Similar studies have been performed by others as well [83,84]. However, these studies have used cell lines that have been in culture that may not accurately reflect all the changes seen in HCC. Comparison of liver tumor tissue with normal tissue would be ideal; however, tissue heterogeneity is an issue and could confound the results [85]. Until only very recently [86,87], an infectious cell culture model for HCV has not been available. This new model system will allow for the identification of cellular proteins that are induced following HCV infection and further the development of a treatment for HCV.

The HCV genome encodes a large polyprotein that is cleaved to generate 9–10 proteins, including the core and envelope proteins E1 and E2, and the nonstructural proteins NS2, NS3, NS4A, NS4B, NS5A, and NS5B [88]. NS5A has been proposed to act as a cofactor in HCV replication [89], as a transcription activator [90], or as an anti-apoptotic factor [91]. To identify the cellular proteins that interacted with NS5A, Choi et al. [26] coimmunoprecipitated NS5A-interacting proteins using antisera against NS5A, separated them via 2D gel electrophoresis, and determined their identity by MS. One cellular protein was found to interact specifically with NS5a, the heat shock protein (HSP) 27, also known as SRP27. Further analysis indicated that HSP27 interacted with the C-terminal domain of NS5A and both proteins were shown to colocalize around the nucleus. The HSPs are induced following cellular stress to protect cells from apoptosis [92]; yet, overexpression of HSP27 did not protect cells from HCV core-induced apoptosis. As HSP27 expression could not be reduced by siRNA, the role of HSP27 in HCV RNA replication could not be determined. However, due to the multifunctional

nature of NS5A, it is surprising that only one protein was found to interact with NS5A. Because an antibody to NS5A was used to coimmunoprecipitate cellular proteins, the antibody used in the study may be interfering with the binding of other cellular proteins. This problem may be overcome by using different antibodies to NS5A or by using a tagged NS5A and affinity chromatography.

3. Diagnostics

Proteomic analysis has provided a unique tool for the identification of diagnostic biomarkers, evaluation of disease progression, and drug development [93,94]. It is also an important approach for clinical diagnostics. In fact, early diagnosis of disease could be possible through the use of unique protein profiles, consisting of a panel of biomarkers that serves as a surrogate marker of disease. Novel diagnostic tests may be generated through proteomic discoveries, and many more proteins can be identified as potential drug targets. These biomarkers are likely to serve multiple purposes, including the assessment of drug efficacy and drug toxicity, and diagnosis. We will review the various methodologies used for viral diagnostics and discuss the advantages and disadvantages of each technique.

3.1. 2DE-MS: SARS, HBV, HCV, and HIV-1

3.1.1. Description

The classical proteomics platforms include 2DE and MS [85]. 2DE is employed to separate proteins in a mixture in the first dimension by their isoelectric points and then in the second dimension by molecular mass. The resulting gel can be stained with a variety of protein dyes to reveal a pattern of spots. In the first dimension, isoelectric focusing (IEF) is performed by using IPG strips which are based on the use of bifunctional immobiline reagents, a series of 10 chemically well-defined acrylamide derivatives that copolymerize with the acrylamide matrix, to generate extremely stable pH gradients forming a series of buffers with different pK values between 1 and 13. Subsequently, linear or nonlinear wide (IPG 3–12), medium (IPG 4–7), narrow (IPG 4.5–5.5), and ultra-narrow (IPG 4.9–5.3) pH range IPGs can be cast [95]. We suggest the reader to refer to the review by Gorg et al. for more details about IPG strip rehydration, sample application, and IPG strip equilibration [96]. The second dimension consists of using SDS-PAGE to separate proteins according to their molecular weight. However, the analysis of low-molecular-weight ($<$15 kDa) and high-molecular-weight ($>$150 kDa) proteins is challenging since there is no standard 2DE system that effectively allows separation of proteins over the entire range between 5 and 500 kDa. A common approach is to combine several gels optimized for different molecular weight ranges instead of using a single standard 2DE system.

3.1.2. Application for virus studies

Current methods for the diagnosis of HCC rely on serological markers such as α-fetoprotein (AFP) [97] and certain liver enzymes as well as Des gamma carboxyprothrombin (DCP) [98]. This type of diagnosis lacks the sensitivity to detect HCC at an early stage when therapy can be more effective. To find markers of disease progression, 2DE was employed to resolve and compare proteins present in serum obtained from individuals infected with HBV or HCV and with varying risks for the development of HCC [99,100]. In several studies, proteins expressed at different levels among diseased individuals as compared to those of healthy ones were identified as markers for disease progression as well as proteins with different *N*-glycosylation patterns [99–101]. In another study, 2D-MS was also employed to analyze altered plasma proteins due to SARS-CoV infection. Thirty-eight different plasma proteins from SARS patients were identified, most of which were associated with acute phase proteins [23].

3.1.3. Advantages

One advantage of 2D gels is their resolution since they can resolve as many as 2000 proteins simultaneously and proteins can be detected at greater than 1 ng in one spot [96]. 2DE is currently the only technique that can be routinely applied for parallel quantitative expression profiling of large sets of complex protein mixtures such as whole cell lysates. In addition, 2DE produces a map of intact proteins, which reflects changes in protein expression level, different isoforms, or PTM. In fact, a great advantage of this methodology is its capability to study proteins that have undergone some form of PTM (such as phosphorylation, glycosylation, or limited proteolysis) that can be detected visually on the 2DE gels as they appear as distinct spot trains in the horizontal and/or vertical axis of the 2DE gel. This is in contrast with other methods, including LC-based methods, which perform analysis on peptides, where molecular weight and p*I* information is lost, and stable isotope labeling is required for quantitative analysis [96].

3.1.4. Drawbacks

Although 2D gel electrophoresis is a standard technology, it suffers from several problems that may limit its utility. These include issues with reproducibility, as well as the inability to separate hydrophobic proteins, which are poorly soluble. Although the use of IPG strips increases the reproducibility of 2DE, various problems with 2D separation remain such as streaking, poor focusing, and the variable occurrence of gaps [85]. Although 2DE allows for high resolution of individual spots, a single spot may not correspond to a single protein, since proteins can comigrate as a single spot on a 2D gel [102]. Furthermore, 2DE requires milligram

quantities of protein, reflecting the low sensitivity of this method. To further enhance the utility of 2DE-MS, enrichment of samples for low-abundance proteins by improved methods is required. Enrichment can include prefractionation of samples, as well as more sensitive detection and quantitation methods, or the use of alternative methods including laser capture microdissection [103] for heterogeneous tissues. In most of the studies mentioned previously, the resolution problem was overcome by narrowing the pH range allowing for greater focusing. However, the reduced pH range in IEF can lead to the elimination of a large number of proteins that may be informative. Comparing hundreds of protein spots across gel images taken from a large number of different samples is extremely time-consuming, even with specialized software. For this reason, although 2D electrophoresis is a promising tool, it is not very practical for clinical application. The challenge is to develop this technique into a system capable of automation, high throughput, and high sensitivity.

3.2. LC–MS: HIV-1 and HCV

3.2.1. Description

Multidimensional LC/MS/MS involves solution proteolysis of a complex mixture of proteins, which are then fractionated by high-performance liquid chromatography (HPLC). Peptides are then analyzed by tandem MS consisting of two phases. In the first phase, peptides in each chromatographic fraction are electrosprayed and ionized producing a mass spectrum characteristic of the molecular weight of each peptide in the sample. In the second phase, the first mass analyzer of the instrument is used to select a single $(M + H)^+$ ion from the mixture and to transmit it to a collision chamber, where the peptide undergoes collisions with argon atoms and suffers fragmentation. The resulting fragment ions are then transferred to a second analyzer, which separates them according to mass [104]. The end result is a mass spectrum containing ions characteristic of the sequence of amino acids in the selected peptide. When mixtures are extremely complex, online reverse-phase LC is used to concentrate and separate the peptides before sequencing by MS [105]. An online capillary LC/MS/MS system consists of conventional HPLC pumps, transfer tubing, a precolumn flow splitter, a liquid junction, a reverse-phase microcapillary column, and a tandem mass spectrometer [106].

3.2.2. Application for virus studies

Two studies have used LC/MS/MS to identify differential protein expression in HIV- or HCV-infected cells. In the first study, traditional HPLC (ion exchange and reverse-phase columns) coupled to an ultrasensitive ion trap MS was employed to

identify proteins that were unique to MDM and to identify proteins present in HIV-1-infected MDM lysates by microsequencing [31]. The second study used normal hepatocytes and immortalized human hepatocytes that can be induced to express the entire HCV ORF. The two different cell types were labeled with an isotopically light (^{12}C for stimulated) or heavy (^{13}C, for control cells) reagent called isotope-coded affinity tag (ICAT); the two differentially labeled samples were then combined and digested with trypsin. Digested peptides were separated by strong cation-exchange chromatography, affinity purified with an avidin cartridge, and subjected to LC-ESI-MS/MS. This study led to the identification of 2159 unique proteins that could be used as markers for disease progression.

3.2.3. Advantages

Some of the advantages include automation in sample application, ability to switch columns, and sensitivity, as this method is able to identify proteins at very low levels [107]. In addition, this method has been extensively used for the determination of drugs and hormone levels in human serum [108–111], making it a promising tool for the detection of disease prognosis markers.

3.2.4. Drawbacks

2DE-based proteome analysis provides information about protein abundance at the gel level by comparing staining intensities. However, when peptide mixtures are analyzed directly by LC/MS/MS techniques, the original quantitative information is lost. For this reason, one of the drawbacks of using LC/MS/MS is the dependence on incorporating stable isotope labeling for quantitative proteome analysis involving the addition of a chemically identical form of the analyte(s) containing stable heavy isotopes (e.g., ^{2}H, ^{13}C, ^{15}N, etc.) to the sample.

3.3. SELDI ProteinChip: SARS, HIV, and hepatitis

3.3.1. Description

SELDI-TOF is a proteomic technology that aims at the quantitative analysis of protein mixtures. This technique relies on the use of trapping surfaces that allow differential capture of proteins based on intrinsic properties of the proteins themselves to identify proteins from crude samples without the need for an initial separation step. A small amount of sample can be directly applied to a biochip coated with specific chemical matrices (hydrophobic, cationic, or anionic) or specific biochemical materials, including DNA fragments or purified proteins. Bound proteins can then be analyzed by MS to obtain either the protein fingerprints or the amino acid sequence when interfaced with a tandem MS.

3.3.2. Application for virus studies

The SELDI ProteinChip approach has been employed to study the protein profiles of cells infected with viruses, including severe acute respiratory syndrome coronavirus (SARS-CoV), HIV-1, and chronic hepatitis B virus infection (CHB) [31]. SARS is a viral respiratory illness caused by SARS-CoV. SARS was recognized as a global threat in March 2003, after first appearing in Southern China in November 2002 (http://www.cdc.gov; [20]). Current serological methods used for laboratory diagnosis of SARS fail to guarantee early diagnosis since most are based on the detection of antibodies that are produced 17–20 days after the onset of symptoms. ELISA-based antigen detection tests offer high specificity and reproducibility, but they lack sensitivity. On the contrary, PCR-based methods, including reverse transcription-PCR, lack sensitivity and specificity [112]. For this reason, there is a need to develop a diagnostic methodology that can detect SARS before the onset of the symptoms to allow for specific prevention and treatment measures for SARS. According to recent studies, SELDI-TOF seems to be a promising approach to study the protein profile unique for SARS. Sera from acute SARS patients or from healthy donors were examined to identify serum marker that could distinguish SARS from non-SARS patients. In this study, analysis of spectra accurately classified 36 of 37 (97.3%) SARS specimens and accurately classified 987 of 993 (99.4%) of the controls as non-SARS. In addition, the classification algorithm successfully distinguished acute SARS from other type of infections with very high precision [22]. The same approach was also employed for the discovery of diagnostic proteomic signatures in the sera of patients with CHB having liver fibrosis and cirrhosis. Results show that 30 serum proteomic features formed a unique fingerprint for fibrosis that correlated with the different stages of fibrosis from minimal fibrosis to cirrhosis [66].

In another study that evaluated the protein fingerprints of HIV-1-infected MDM, cell lysates were directly applied on two types of protein chips: weak cation exchange and reverse-phase hydrophobic interaction. After washing to remove the unbound proteins, bound proteins were ionized and their molecular mass/charge ratio was determined using TOF analysis. Analysis of the obtained profiles showed distinct patterns between uninfected and infected MDM [33].

3.3.3. Advantages

The SELDI ProteinChip approach allows for high-throughput protein analysis of crude protein mixtures without the need for a separation step. It is sensitive since it takes advantage of the analytical capacity of MS combined with novel surface chemistry. It can provide a phenotypic fingerprint of complex mixtures. Sample requirements are dramatically reduced, and because this approach employs MS

for its readout, attomolar to femtomolar concentrations of proteins can be detected. Additionally, reproducibility is greater than that of other techniques such as 2D gels; proteins at extreme p*I*s can be identified, a condition that is problematic under normal 2D gel electrophoresis conditions; and finally there is a greater sensitivity and accuracy for low-molecular-weight proteins (<25 kDa) using SELDI, especially below 10 kDa, which is particularly troublesome for 2D gels.

3.3.4. Drawbacks

This method needs a very robust algorithm to ensure specificity of the profile, in that it can distinguish the pattern between disease and healthy individuals with high accuracy, taking into account variations in profiles between healthy individuals as well as persons with a variety of different infections at different time periods in their course of illness. Two additional drawbacks of this approach are the following: (i) The identity of the proteins cannot be discovered and (ii) as the absolute intensity of the peaks is measured in relationship to the most abundant peaks, peaks in low abundance will be masked by the more abundant ones. In addition, this method employs the direct analysis of tissues or biological fluids by MALDI. The main drawbacks of this approach are the preferential detection of proteins with a lower molecular mass and the difficulty in determining the identity of proteins owing to PTM obscuring the correspondence of measured and predicted masses.

3.4. Protein microarray: vaccinia virus

3.4.1. Description

A protein microarray relies on high-throughput amplification of each predicted ORF by using gene-specific primers, followed by *in vivo* homologous recombination into a T7 expression vector. The proteins are expressed in an *Escherichia coli*-based cell-free *in vitro* transcription/translation system. The protein products from the unpurified reactions are printed directly onto nitrocellulose microarrays without further purification [113].

3.4.2. Application for virus studies

This approach was used to determine the complete antigen-specific humoral immune-response profile from infected humans and animals. The vaccinia virus proteome containing 185 individual viral proteins was printed on a chip after cloning and expression. The chips were then used to determine the antibody profile in serum from vaccinia-virus-immunized humans, primates, and mice [113].

3.4.3. Advantages

Once it has been developed and produced, a protein microarray can be a very rapid method (3 days for most of the genes) to comprehensively scan the humoral immune response of vaccinated or infected individuals.

3.4.4. Drawbacks

The generation of a complete proteome is technically challenging. One problem is the amplification of long genes. Furthermore, expression of some proteins in heterologous systems is not efficient. This technique also does not take into account PTM of viral proteins that are expressed in bacteria. Lastly, expression in *E. coli* might lead to folding problems of the protein.

4. Discussion

Proteomic analysis of cellular protein samples began with the development of PAGE [114] and later with the development of two-dimensional gel electrophoresis (2D-PAGE) [115]. These techniques allowed for the separation of proteins based on size (PAGE) or charge and size (2D-PAGE). These methods, however, did not allow for direct identification of these protein bands. Indirect methods such as Western blotting with specific antibodies were required for identification—a slow and laborious process. However, by combining a variety of mass spectrometric methods with PAGE, identification of a larger number of proteins has become possible. These methods have proven invaluable in furthering various avenues of viral research. Proteomic analysis of viruses has included identification of proteins in virus particles, characterization of virus–host protein–protein interactions, and analysis of serum proteins for biomarkers of disease.

One aspect of viral proteomics has been the characterization of virus particles and virally infected cells. Characterization of purified virions has led to the identification of viral proteins that were not originally identified with the virion as well as the identification of cellular proteins associated with the purified virus. For example, analysis of HCMV viral particles identified 12 additional ORFs not previously known to reside in virions as well as the identification of 71 cellular proteins [6]. The importance of these cellular and viral proteins in viral replication or pathogenesis awaits further analysis. Additionally, 12 unique polypeptides were identified that did not correspond to previously identified ORFs [6], illustrating the fact that despite intensive sequence analysis, sequence characteristics of viral promoters and ORFs are still not entirely understood. Analysis of virally infected cells has also led to the characterization of events leading to EBV-induced transformation [9,10,47], identification of cellular proteins induced in HIV-infected

macrophages [31], and identification of cellular proteins that may be involved in AIDS-associated dementia [66].

The characterization of virus–host protein–protein interactions has been intensely studied. Originally most studies have relied on the analysis of the interaction of two proteins or used the yeast two-hybrid system to identify new protein partners of a protein of interest. These studies, however, are quite labor intensive. Furthermore, the yeast two-hybrid system is susceptible to false-positive identifications, cannot be used to identify multiprotein complexes, and typically does not take into account possible PTM that may influence protein binding. Proteomic analysis, however, can be used to identify multiprotein complexes and, when used in the analysis of infected cells, will take into account any PTM that occur in infected cells. Proteomic analysis of infected cells has resulted in the identification of cellular proteins that may mediate HSV IC27-induced repression of cellular protein synthesis [18], and the identification of over 50 cellular and viral proteins in HSV DNA replication [19]. Furthermore, analysis of the HIV Tat and HTLV Tax proteomes identified members of chromatin remodeling complexes as components of these viral transactivator multiprotein complexes [60,73]. Many of these studies will allow for further understanding of virus transcription, replication, and transformation. Additionally, these studies may lead to the identification of unique drug targets. For example, the p-TEFb complex has been shown to be critical for HIV gene expression [56,60,116] and HIV-infected cells are uniquely sensitive to the transcription suppressing effects of the p-TEFb inhibitor flavopiridol [116–118].

A number of viruses are the causative agents of cancer, including EBV, hepatitis B virus, and hepatitis C virus (HCV). HCV is a major cause of the increasing incidence of liver cancer in developed countries [80], though events leading to transformation are not well understood. Until recently [86,87], an infectious cell culture model of HCV has not been available. The lack of a cell culture model has prevented the systematic analysis of changes induced by HCV infection. Alternative approaches to studying HCV transformation have been the comparison of liver tissue and hepatoma-derived cell lines [82–84] and analysis of a single virus (NS5a)–host (HSP27) protein–protein interaction [26]. Further analysis of additional HCV proteins and infected cells will provide additional insights into the nature of this virus and its ability to cause cancer.

One aspect of viral proteomics that is of interest to physicians is the analysis of serum for protein biomarkers of disease. Studies have been performed on patients infected with SARS-CoV, HIV, HCV, HBV, and HIV-1 using a variety of approaches. Some of the methods that have been used are 2D-PAGE followed by MS, LC/MS/MS, SELDI ProteinChips, and protein microarrays. These methods have their advantages and disadvantages. 2D gel electrophoresis allows for resolution of greater than 1000 protein species, can be used for quantitative analysis of expression, and reflects changes in PTM and the identification of isoforms. However, several issues with 2D gel electrophoresis are its lack of reproducibility, the difficulty in detecting hydrophobic proteins, low sensitivity, and the inability

to use a high-throughput method to analyze a large number of samples. LC–MS has the advantages of solubilization of the majority of proteins, automation, ability to switch columns, and sensitivity; however, the ability to quantify changes in protein levels is lost with LC–MS.

The SELDI ProteinChip is unique in that it allows for differential separation of complex protein mixtures based on chemical characteristics such as hydrophobicity or charge, resulting in a decrease in the complexity of the sample analyzed. However, SELDI is considered a soft-ionization method and the results obtained are patterns of protein peaks and not the identification of peptide masses. To ensure the specificity of peak profile for a particular disease state, a robust algorithm is needed. Lastly, protein microarrays have been developed to determine the immune response to a viral infection. The method requires the expression and printing of all ORFs of a pathogen and cross-linking them to a solid support. Protein microarrays would allow for the rapid diagnosis of a particular viral infection. However, expression of a complete proteome is a challenging task. As the proteins are expressed in bacteria, potentially important PTM are lost and proteins may not be properly folded.

Serum is a complex mixture of proteins that is dominated by two proteins—albumin and immunoglobulin (Ig) [119]. The abundance of these proteins means that analysis of serum for potential biomarkers of disease requires either very sensitive methods or separation of albumin and Ig from serum. Several albumin and/or Ig depletion methods have been developed to resolve this issue. Pieper et al. [120] developed a series of chromatographic columns to separate immunoglobulins based on their affinity for proteins A and G as well as columns containing antibodies with specificities for individual proteins such as albumin, fibrinogen, and transferrin. The columns were successful in depleting serum samples of their respective proteins, and use of several columns significantly decreased the complexity of the sample analyzed [120]. Additionally, a mixed-bed column was developed that allowed the simultaneous separation of several proteins, which would allow for automated processing of samples. A similar approach has been developed by Bio-Rad (Affi-Gel Blue) to deplete samples of albumin, enhancing the detection of other proteins in the sample [121]. Affi-Gel Blue has affinity for hydrophobic, aromatic, or sterically active binding sites of protein. Although this product has high affinity for albumin, it may bind other proteins as well, limiting its usefulness. Lastly, Baussant et al. [122] developed a peptide-based approach to deplete albumin. Their approach was based on the fact that although protein G has affinity for the Fc region of IgG, it can also bind albumin less specifically. Baussant et al. modified a peptide of protein G to have a much higher affinity for albumin, which significantly and specifically depleted the serum of albumin; however, other hydrophobic proteins, i.e., apolipoproteins, were also captured [122].

Despite their relatively small size, viruses are fairly complex and encode between a dozen and more than 200 proteins. Many of these proteins are post-translationally modified and interact with other viral and host proteins to function.

Identifying the proteins that are encoded by viruses and the proteins with which they interact will greatly further the understanding of viral replication and pathogenesis and proteomic approaches will greatly facilitate these studies. Lastly, the ability to diagnose cancer or viral infections at early stages will allow for early treatment and reduce the morbidity and mortality associated with these diseases. Proteomic analysis of biological markers in serum should allow for the early non-invasive diagnosis of cancer. Although good reliable methods are available for the analysis of the serum proteome, the abundance of a few proteins, i.e., albumin and Ig, and the low abundance of many other proteins will require methods for separating out the high-abundance proteins and instruments and methods with enough sensitivity to identify proteins at low concentration.

5. Future trends

It is becoming increasingly clear that the field of proteomics may require better and more robust separation methods, sensitive instrumentation, and unbiased bioinformatic tools. 2DE has historically provided a rapid means for separating thousands of proteins from cell and tissue samples in one run. Although this is a powerful research tool and has been enthusiastically applied in many fields of biomedical research, accurate analysis and interpretation of the data have provided many challenges. Several analysis steps are needed to convert the large amount of noisy data obtained with 2DE into reliable and interpretable biological information. The goals of such analysis steps include accurate protein detection and quantification, consistent comparative visualization methods, as well as the identification of differentially expressed proteins between samples run on different gels. To achieve these goals, systematic errors such as geometric distortions between the gels must be corrected by using computer-assisted methods. A wide range of computer software has been developed, but no general consensus exists as a standard for 2DE data analysis protocols.

In search for new diagnostic and therapeutic targets, 2DE has been used to study differential expression of peptides and proteins in various disease entities. However, 2DE usually requires large amounts of starting material, is time-consuming, and reveals only a fraction of the proteins present in a given sample. More recently, the ProteinChip technology coupled with bioinformatics has gained considerable attention. This technique uses SELDI-TOF/MS to screen protein sources for putative disease biomarkers in a spectrum from 2 to 20 kDa. Several studies have provided evidence that ProteinChip technology is capable of detecting early-stage cancer by its unique cancer-specific proteomic fingerprints, with sensitivities and specificities reaching far beyond well-established serum-based tumor markers [123]. However, as in most rapid diagnosis tests, SELDI technology can still not detect the nature of the amino acid biomarkers or their

PTM in a consistent and reproducible manner. Clearly other technologies such as the LC/MS/MS and the LC-FTICR are far more sensitive and better in defining the composition of these biomarkers.

Finally, very recently, much effort has gone into the concept of "Lab-on-a-chip." These chips involve micron-sized channels embedded in glass or silicon chips. Attempts have been made to carry out two-dimensional gel-based experiments on chips. Microchips that are able to carry out microfluidic experiments are being developed (e.g., Nanogen Inc., DiagnoSwiss, Caliper Technologies), which are faster and more accurate than the conventional gel technology. If such technologies were made 2DE compatible then it would offer immense research potential. Especially promising are advancements in detecting low-abundance proteins and PTM.

6. Conclusions

In this chapter, we have discussed the latest new proteomics findings that relate to some of the most important viral infections known to humans. These included HCMV, HSV, EBV, KSHV, HIV, HTLV, HBV, HCV, and SARS infections. In many instances we have seen a mere description of the viral or the infected host cell proteome; however, most of the data to date are descriptive in nature and very few studies have correlated phenotype of the infection to the pathology or drug treatment. Although in some cases investigators have found new enzyme targets as markers (i.e., SARS-CoV), no serious attempts have been made to functionally identify their significance in the pathology of the virus. This is mainly because the field of viral proteomics is at its early stages of development and much confirmatory information would be required from animal or human model studies, which are currently either in progress or will need to be developed in near future. Therefore, a new field of functional viral proteomics is developing in both industrial and academic settings to address issues related to functional biomarkers, drug-resistance viruses, and host/pathogen relations that pertain to disease prognosis, treatment decision, and monitoring response to therapy.

Another challenging consideration is the mixed infections seen in AIDS patients who not only may have varying HIV-1 clade infections (more than seven clades, and close to 1500 genetically distinct HIV-1 genotypes) but also are coinfected with other viruses such as HCV or KSHV. The complication of identifying biomarkers in these patients, or in some instances animal models, has never been properly addressed in the current literature, nor is there enough awareness between various compartments of patient bench to bedside practices. Therefore, a better flow of information using solid epidemiological data followed by better diagnostics for the viral etiology would allow a meaningful identification of the proteome biomarkers seen in these patients. These multiple biomarkers would serve as invaluable tools for multiple drug treatments and better control of mixed infections.

Finally, the issue of frontend purification for the collected test material is perhaps the most important aspect of sample preparation. Currently there are various methods that utilize standard separation techniques to remove most abundant proteins prior to MS, i.e., removal of some 20 high-abundance proteins and better visualization of low-abundance proteins (Sigma-Aldrich kits). However, in most cases the removal of these proteins may in fact compromise the detection of biomarkers or their partners, since in many instances, viral infection leads to overexpression of the most abundant proteins such as the actins, keratins, tubulins, cyclophilins, vimentin, and HSPs among others. Therefore, future attempts at the identification of biomarkers would have to define not only the most high- and low-abundance proteins but also their partners and possible modifications.

Acknowledgments

This work was supported in part by grants from the George Washington University REF funds to Drs. Fatah Kashanchi and Akos Vertes, and NIH grants AI44357, AI43894 and 13969 to Dr. Fatah Kashanchi. We would like to thank Cynthia de la Fuente and Kylene Kehn for their contributions and review of the literature and Lynne Mied for the development of figures.

References

1. Wilkins, M.R., Williams, K.L., Appel, R.D. and Hochstrasser, D.F., *Proteome Research: New Frontiers in Functional Genomics*. Springer, Berlin, 1997.
2. Gandhi, M.K. and Khanna, R., Human cytomegalovirus: clinical aspects, immune regulation, and emerging treatments. *Lancet Infect. Dis.*, **4**, 725–738 (2004).
3. Moss, P. and Khan, N., CD8(+) T-cell immunity to cytomegalovirus. *Hum. Immunol.*, **65**, 456–464 (2004).
4. Prichard, M.N., Penfold, M.E., Duke, G.M., Spaete, R.R. and Kemble, G.W., A review of genetic differences between limited and extensively passaged human cytomegalovirus strains. *Rev. Med. Virol.*, **11**, 191–200 (2001).
5. Britt, W.J. and Boppana, S., Human cytomegalovirus virion proteins. *Hum. Immunol.*, **65**, 395–402 (2004).
6. Varnum, S.M., Streblow, D.N., Monroe, M.E., Smith, P., Auberry, K.J., Pasa-Tolic, L., Wang, D., Camp, D.G., 2nd, Rodland, K., Wiley, S., Britt, W., Shenk, T., Smith, R.D. and Nelson, J.A., Identification of proteins in human cytomegalovirus (HCMV) particles: the HCMV proteome. *J. Virol.*, **78**, 10960–10966 (2004).
7. Ambinder, R.F., Epstein–Barr virus-associated lymphoproliferative disorders. *Rev. Clin. Exp. Hematol.*, **7**, 362–374 (2003).
8. Cohen, J.I., The biology of Epstein–Barr virus: lessons learned from the virus and the host. *Curr. Opin. Immunol.*, **11**, 365–370 (1999).
9. Toda, T., Sugimoto, M., Omori, A., Matsuzaki, T., Furuichi, Y. and Kimura, N., Proteomic analysis of Epstein–Barr virus-transformed human B-lymphoblastoid cell lines before and after immortalization. *Electrophoresis*, **21**, 1814–1822 (2000).

10. Schlee, M., Krug, T., Gires, O., Zeidler, R., Hammerschmidt, W., Mailhammer, R., Laux, G., Sauer, G., Lovric, J. and Bornkamm, G.W., Identification of Epstein–Barr virus (EBV) nuclear antigen 2 (EBNA2) target proteins by proteome analysis: activation of EBNA2 in conditionally immortalized B cells reflects early events after infection of primary B cells by EBV. *J. Virol.*, **78**, 3941–3952 (2004).
11. Poirier, F., Pontet, M., Labas, V., le Caer, J.P., Sghiouar-Imam, N., Raphael, M., Caron, M. and Joubert-Caron, R., Two-dimensional database of a Burkitt lymphoma cell line (DG 75) proteins: protein pattern changes following treatment with 5′-azycytidine. *Electrophoresis*, **22**, 1867–1877 (2001).
12. Whitley, R.J. In: Howley, P.M. (Ed.), *Fields Virology* (Vol. 2, 4th edition). Lippincott Williams & Wilkins, Philadelphia, 2001, pp. 2461–2510.
13. Kimberlin, D.W. and Whitley, R.J., Neonatal herpes: what have we learned. *Semin. Pediatr. Infect. Dis.*, **16**, 7–16 (2005).
14. Kawaguchi, Y., Matsumura, T., Roizman, B. and Hirai, K., Cellular elongation factor 1delta is modified in cells infected with representative alpha-, beta-, or gammaherpesviruses. *J. Virol.*, **73**, 4456–4460 (1999).
15. Kawaguchi, Y., Kato, K., Tanaka, M., Kanamori, M., Nishiyama, Y. and Yamanashi, Y., Conserved protein kinases encoded by herpesviruses and cellular protein kinase cdc2 target the same phosphorylation site in eukaryotic elongation factor 1delta. *J. Virol.*, **77**, 2359–2368 (2003).
16. Walsh, D. and Mohr, I., Phosphorylation of eIF4E by Mnk-1 enhances HSV-1 translation and replication in quiescent cells. *Genes. Dev.*, **18**, 660–672 (2004).
17. Greco, A., Bienvenut, W., Sanchez, J.C., Kindbeiter, K., Hochstrasser, D., Madjar, J.J. and Diaz, J.J., Identification of ribosome-associated viral and cellular basic proteins during the course of infection with herpes simplex virus type 1. *Proteomics.*, **1**, 545–549 (2001).
18. Fontaine-Rodriguez, E.C., Taylor, T.J., Olesky, M. and Knipe, D.M., Proteomics of herpes simplex virus infected cell protein 27: association with translation initiation factors. *Virology*, **330**, 487–492 (2004).
19. Taylor, T.J. and Knipe, D.M., Proteomics of herpes simplex virus replication compartments: association of cellular DNA replication, repair, recombination, and chromatin remodeling proteins with ICP8. *J. Virol.*, **78**, 5856–5866 (2004).
20. Prevalence of IgG antibody to SARS-associated coronavirus in animal traders— Guangdong Province, China, 2003. *MMWR Morb Mortal Wkly Rep,* **52**, 986–987 (2003).
21. Berger, A., Drosten, C., Doerr, H.W., Sturmer, M. and Preiser, W., Severe acute respiratory syndrome (SARS)—paradigm of an emerging viral infection. *J. Clin. Virol.*, **29**, 13–22 (2004).
22. Kang, X., Xu, Y., Wu, X., Liang, Y., Wang, C., Guo, J., Wang, Y., Chen, M., Wu, D., Bi, S., Qiu, Y., Lu, P., Cheng, J., Xiao, B., Hu, L., Gao, X., Liu, J., Song, Y., Zhang, L., Suo, F., Chen, T., Huang, Z., Zhao, Y., Lu, H., Pan, C. and Tang, H., Proteomic fingerprints for potential application to early diagnosis of severe acute respiratory syndrome. *Clin. Chem.*, **51**, 56–64 (2005).
23. Chen, J.H., Chang, Y.W., Yao, C.W., Chiueh, T.S., Huang, S.C., Chien, K.Y., Chen, A., Chang, F.Y., Wong, C.H. and Chen, Y.J., Plasma proteome of severe acute respiratory syndrome analyzed by two-dimensional gel electrophoresis and mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 17039–17044 (2004).
24. Snijder, E.J., Bredenbeek, P.J., Dobbe, J.C., Thiel, V., Ziebuhr, J., Poon, L.L., Guan, Y., Rozanov, M., Spaan, W.J. and Gorbalenya, A.E., Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol.*, **331**, 991–1004 (2003).
25. Dickson, R.C., Clinical manifestations of hepatitis C. *Clin. Liver Dis.*, **1**, 569–585 (1997).

26. Choi, Y.W., Tan, Y.J., Lim, S.G., Hong, W. and Goh, P.Y., Proteomic approach identifies HSP27 as an interacting partner of the hepatitis C virus NS5A protein. *Biochem. Biophys. Res. Commun.*, **318**, 514–519 (2004).
27. Block, T.M., Comunale, M.A., Lowman, M., Steel, L.F., Romano, P.R., Fimmel, C., Tennant, B.C., London, W.T., Evans, A.A., Blumberg, B.S., Dwek, R.A., Mattu, T.S. and Mehta, A.S., Use of targeted glycoproteomics to identify serum glycoproteins that correlate with liver cancer in woodchucks and humans. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 779–784 (2005).
28. Kim, W., Oe Lim, S., Kim, J.S., Ryu, Y.H., Byeon, J.Y., Kim, H.J., Kim, Y.I., Heo, J.S., Park, Y.M. and Jung, G., Comparison of proteome between hepatitis B virus- and hepatitis C virus-associated hepatocellular carcinoma. *Clin. Cancer Res.*, **9**, 5493–5500 (2003).
29. Tossing, G., Management of chronic hepatitis C in HIV-co-infected patients—results from the First International Workshop on HIV and Hepatitis Co-infection, 2nd–4th December 2004, Amsterdam, Netherlands. *Eur. J. Med. Res.*, **10**, 43–45 (2005).
30. Rockstroh, J.K. and Spengler, U., HIV and hepatitis C virus co-infection. *Lancet. Infect. Dis.*, **4**, 437–444 (2004).
31. Carlson, K.A., Ciborowski, P., Schellpeper, C.N., Biskup, T.M., Shen, R.F., Luo, X., Destache, C.J. and Gendelman, H.E., Proteomic fingerprinting of HIV-1-infected human monocyte-derived macrophages: a preliminary report. *J. Neuroimmunol.*, **147**, 35–42 (2004).
32. Yip, T.T., Chan, J.W., Cho, W.C., Wang, Z., Kwan, T.L., Law, S.C., Tsang, D.N., Chan, J.K., Lee, K.C., Cheng, W.W., Ma, V.W., Yip, C., Lim, C.K., Ngan, R.K., Au, J.S., Chan, A. and Lim, W.W., Protein chip array profiling analysis in patients with severe acute respiratory syndrome identified serum amyloid a protein as a biomarker potentially useful in monitoring the extent of pneumonia. *Clin. Chem.*, **51**, 47–55 (2005).
33. Wojna, V., Carlson, K.A., Luo, X., Mayo, R., Melendez, L.M., Kraiselburd, E. and Gendelman, H. E., Proteomic fingerprinting of human immunodeficiency virus type 1-associated dementia from patient monocyte-derived macrophages: a case study. *J. Neurovirol.*, **10(Suppl. 1)**, 74–81 (2004).
34. Luo, X., Carlson, K.A., Wojna, V., Mayo, R., Biskup, T.M., Stoner, J., Anderson, J., Gendelman, H.E. and Melendez, L.M., Macrophage proteomic fingerprinting predicts HIV-1-associated cognitive impairment. *Neurology*, **60**, 1931–1937 (2003).
35. Fuchigami, T., Misumi, S., Takamune, N., Takahashi, I., Takama, M. and Shoji, S., Acid-labile formylation of amino terminal proline of human immunodeficiency virus type 1 p24(gag) was found by proteomics using two-dimensional gel electrophoresis and matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry. *Biochem. Biophys. Res. Commun.*, **293**, 1107–1113 (2002).
36. Chelius, D., Huhmer, A.F., Shieh, C.H., Lehmberg, E., Traina, J.A., Slattery, T.K. and Pungor, E., Jr., Analysis of the adenovirus type 5 proteome by liquid chromatography and tandem mass spectrometry methods. *J. Proteome. Res.*, **1**, 501–513 (2002).
37. Yao, X., Freas, A., Ramirez, J., Demirev, P.A. and Fenselau, C., Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal. Chem.*, **73**, 2836–2842 (2001).
38. Yan, W., Lee, H., Deutsch, E.W., Lazaro, C.A., Tang, W., Chen, E., Fausto, N., Katze, M.G. and Aebersold, R., A dataset of human liver proteins identified by protein profiling via isotope-coded affinity tag (ICAT) and tandem mass spectrometry. *Mol. Cell. Proteomics*, **3**, 1039–1041 (2004).
39. Bao, Y., Federhen, S., Leipe, D., Pham, V., Resenchuk, S., Rozanov, M., Tatusov, R. and Tatusova, T., National center for biotechnology information viral genomes project. *J. Virol.*, **78**, 7291–7298 (2004).

40. Bechtel, J.T., Winant, R.C. and Ganem, D., Host and viral proteins in the virion of Kaposi's sarcoma-associated herpesvirus. *J. Virol.*, **79**, 4952–4964 (2005).
41. Nealon, K., Newcomb, W.W., Pray, T.R., Craik, C.S., Brown, J.C. and Kedes, D.H., Lytic replication of Kaposi's sarcoma-associated herpesvirus results in the formation of multiple capsid species: isolation and molecular characterization of A, B, and C capsids from a gammaherpesvirus. *J. Virol.*, **75**, 2866–2878 (2001).
42. Roizman, B. and Pellett, P.E. In: Howley, P.M., (Ed.), *Fields Virology* (Vol. 2, 4th edition). Lippincott Williams & Wilkins, Philadelphia, 2001, pp. 2381–2398.
43. Triezenberg, S.J., Kingsbury, R.C. and McKnight, S.L., Functional dissection of VP16, the trans-activator of herpes simplex virus immediate early gene expression. *Genes. Dev.*, **2**, 718–729 (1988).
44. Russo, J.J., Bohenzky, R.A., Chien, M.C., Chen, J., Yan, M., Maddalena, D., Parry, J.P., Peruzzi, D., Edelman, I.S., Chang, Y. and Moore, P.S., Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 14862–14867 (1996).
45. Zhu, F.X., Chong, J.M., Wu, L. and Yuan, Y., Virion proteins of Kaposi's sarcoma-associated herpesvirus. *J. Virol.*, **79**, 800–811 (2005).
46. Young, L.S. and Rickinson, A.B., Epstein–Barr virus: 40 years on. *Nat. Rev. Cancer*, **4**, 757–768 (2004).
47. Toda, T. and Sugimoto, M., Proteome analysis of Epstein–Barr virus-transformed B-lymphoblasts and the proteome database. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **787**, 197–206 (2003).
48. Sixbey, J.W., Shirley, P., Sloas, M., Raab-Traub, N. and Israele, V., A transformation-incompetent, nuclear antigen 2-deleted Epstein–Barr virus associated with replicative infection. *J. Infect. Dis.*, **163**, 1008–1015 (1991).
49. Kempkes, B., Spitkovsky, D., Jansen-Durr, P., Ellwart, J.W., Kremmer, E., Delecluse, H.J., Rottenberger, C., Bornkamm, G.W. and Hammerschmidt, W., B-cell proliferation and induction of early G1-regulating proteins by Epstein–Barr virus mutants conditional for EBNA2. *Embo. J.*, **14**, 88–96 (1995).
50. Kaiser, C., Laux, G., Eick, D., Jochner, N., Bornkamm, G.W. and Kempkes, B., The proto-oncogene c-myc is a direct target gene of Epstein–Barr virus nuclear antigen 2. *J. Virol.*, **73**, 4481–4484 (1999).
51. Polack, A., Hortnagel, K., Pajic, A., Christoph, B., Baier, B., Falk, M., Mautner, J., Geltinger, C., Bornkamm, G.W. and Kempkes, B., c-myc activation renders proliferation of Epstein–Barr virus (EBV)-transformed cells independent of EBV nuclear antigen 2 and latent membrane protein 1. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 10411–10416 (1996).
52. Leone, A., Flatow, U., King, C.R., Sandeen, M.A., Margulies, I.M., Liotta, L.A. and Steeg, P.S., Reduced tumor incidence, metastatic potential, and cytokine responsiveness of nm23-transfected melanoma cells. *Cell.*, **65**, 25–35 (1991).
53. Roizman, B. and Knipe, D.M., *Herpes simplex viruses and their replication (*4th edition). Howley, P.M., Philadelphia: Lippincott Williams and Wilkins, 2001.
54. Phelan, A., Dunlop, J., Patel, A.H., Stow, N.D. and Clements, J.B., Nuclear sites of herpes simplex virus type 1 DNA replication and transcription colocalize at early times postinfection and are largely distinct from RNA processing factors. *J. Virol.*, **71**, 1124–1132 (1997).
55. Liptak, L.M., Uprichard, S.L. and Knipe, D.M., Functional order of assembly of herpes simplex virus DNA replication proteins into prereplicative site structures. *J. Virol.*, **70**, 1759–1767 (1996).
56. Mancebo, H.S., Lee, G., Flygare, J., Tomassini, J., Luu, P., Zhu, Y., Peng, J., Blau, C., Hazuda, D., Price, D. and Flores, O., P-TEFb kinase is required for HIV Tat transcriptional activation *in vivo* and *in vitro*. *Genes Dev.*, **11**, 2633–2644 (1997).

57. Zhu, Y., Pe'ery, T., Peng, J., Ramanathan, Y., Marshall, N., Marshall, T., Amendt, B., Mathews, M.B. and Price, D.H., Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation *in vitro*. *Genes Dev.*, **11**, 2622–2632 (1997).
58. Van Lint, C., Emiliani, S., Ott, M. and Verdin, E., Transcriptional activation and chromatin remodeling of the HIV-1 promoter in response to histone acetylation. *Embo. J.*, **15**, 1112–1120 (1996).
59. Verdin, E., Paras, P., Jr. and Van Lint, C., Chromatin disruption in the promoter of human immunodeficiency virus type 1 during transcriptional activation. *Embo. J.*, **12**, 3249–3259 (1993).
60. Pumfery, A., Deng, L., Maddukuri, A., de la Fuente, C., Li, H., Wade, J.D., Lambert, P., Kumar, A. and Kashanchi, F., Chromatin remodeling and modification during HIV-1 Tat-activated transcription. *Current HIV Res.*, **1**, 343–362 (2003).
61. Deng, L., de la Fuente, C., Fu, P., Wang, L., Donnelly, R., Wade, J.D., Lambert, P., Li, H., Lee, C.G. and Kashanchi, F. Acetylation of HIV-1 Tat by CBP/P300 increases transcription of integrated HIV-1 genome and enhances binding to core histones. *Virology*, **277**, 278–295 (2000).
62. Deng, L., Wang, D., de la Fuente, C., Wang, L., Li, H., Lee, C.G., Donnelly, R., Wade, J.D., Lambert, P. and Kashanchi, F., Enhancement of the p300 HAT activity by HIV-1 Tat on chromatin DNA. *Virology*, **289**, 312–326 (2001).
63. Janssen, R.S., Nwanyanwu, O.C., Selik, R. M. and Stehr-Green, J.K., Epidemiology of human immunodeficiency virus encephalopathy in the United States. *Neurology*, **42**, 1472–1476 (1992).
64. Masliah, E., Ge, N. and Mucke, L. (1996). Pathogenesis of HIV-1 associated neurodegeneration. *Crit. Rev. Neurobiol.*, **10**, 57-67
65. Chauhan, A., Turchan, J., Pocernich, C., Bruce-Keller, A., Roth, S., Butterfield, D.A., Major, E.O. and Nath, A., Intracellular human immunodeficiency virus Tat expression in astrocytes promotes astrocyte survival but induces potent neurotoxicity at distant sites via axonal transport. *J. Biol. Chem.*, **278**, 13512–13519 (2003).
66. Pocernich, C.B., Boyd-Kimball, D., Poon, H.F., Thongboonkerd, V., Lynn, B.C., Klein, J.B., Calebrese, V., Nath, A. and Butterfield, D.A., Proteomics analysis of human astrocytes expressing the HIV protein Tat. *Brain Res. Mol. Brain Res.*, **133**, 307–316 (2005).
67. Adamson, D.C., McArthur, J.C., Dawson, T.M. and Dawson, V.L., Rate and severity of HIV-associated dementia (HAD): correlations with Gp41 and iNOS. *Mol. Med.*, **5**, 98–109 (1999).
68. Adamson, D.C., Wildemann, B., Sasaki, M., Glass, J.D., McArthur, J.C., Christov, V.I., Dawson, T.M. and Dawson, V.L., Immunologic NO synthase: elevation in severe AIDS dementia and induction by HIV-1 gp41. *Science*, **274**, 1917–1921 (1996).
69. Mahieux, R. and Gessain, A., HTLV-1 and associated adult T-cell leukemia/lymphoma. *Rev. Clin. Exp. Hematol.*, **7**, 336–361 (2003).
70. Kehn, K., Berro, R., de la Fuente, C., Strouss, K., Ghedin, E., Dadgar, S., Bottazzi, M.E., Pumfery, A. and Kashanchi, F., Mechanisms of HTLV-1 transformation. *Front Biosci.*, **9**, 2347–2372 (2004).
71. Suzuki, T., Fujisawa, J.I., Toita, M. and Yoshida, M., The trans-activator tax of human T-cell leukemia virus type 1 (HTLV-1) interacts with cAMP-responsive element (CRE) binding and CRE modulator proteins that bind to the 21-base-pair enhancer of HTLV-1. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 610–614 (1993).
72. Yoshida, M., Mechanism of transcriptional activation of viral and cellular genes by oncogenic protein of HTLV-1. *Leukemia*, **8**(**Suppl. 1**), S51–S53 (1994).
73. Wu, K., Bottazzi, M.E., de la Fuente, C., Deng, L., Gitlin, S.D., Maddukuri, A., Dadgar, S., Li, H., Vertes, A., Pumfery, A. and Kashanchi, F., Protein profile of tax-associated complexes. *J. Biol. Chem.*, **279**, 495–508 (2004).

74. Oda, T., Watanabe, S., Sumii, H., Nakamura, T., Arakaki, Y. and Shimotohno, K., Immunoelectron microscopic localization of the pX gene products in human T-cell leukemia virus type 1-producing cells. *Cancer Res.*, **47**, 2077–2082 (1987).
75. Bosch, F.X., Ribes, J. and Borras, J., Epidemiology of primary liver cancer. *Semin. Liver Dis.*, **19**, 271–285 (1999).
76. Beasley, R.P., Hepatitis B virus. The major etiology of hepatocellular carcinoma. *Cancer*, **61**, 1942–1956 (1988).
77. el-Refaie, A., Savage, K., Bhattacharya, S., Khakoo, S., Harrison, T.J., el-Batanony, M., Soliman el, S., Nasr, S., Mokhtar, N., Amer, K., Scheuer, P.J. and Dhillon, A.P., HCV-associated hepatocellular carcinoma without cirrhosis. *J. Hepatol.*, **24**, 277–285 (1996).
78. Colombo, M., Hepatitis C virus and hepatocellular carcinoma. *Semin. Liver Dis.*, **19**, 263–269 (1999).
79. Lavanchy, D., Hepatitis B virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures. *J. Viral Hepat.*, **11**, 97–107 (2004).
80. El-Serag, H.B. and Mason, A.C., Rising incidence of hepatocellular carcinoma in the United States. *N. Engl. J. Med.*, **340**, 745–750 (1999).
81. Steel, L.F., Mattu, T.S., Mehta, A., Hebestreit, H., Dwek, R., Evans, A.A., London, W.T. and Block, T., A proteomic approach for the discovery of early detection markers of hepatocellular carcinoma. *Dis. Markers*, **17**, 179–189 (2001).
82. Wirth, P.J., Hoang, T.N. and Benjamin, T., Micropreparative immobilized pH gradient two-dimensional electrophoresis in combination with protein microsequencing for the analysis of human liver proteins. *Electrophoresis*, **16**, 1946–1960 (1995).
83. Seow, T.K., Ong, S.E., Liang, R.C., Ren, E.C., Chan, L., Ou, K. and Chung, M.C., Two-dimensional electrophoresis map of the human hepatocellular carcinoma cell line, HCC-M, and identification of the separated proteins by mass spectrometry. *Electrophoresis*, **21**, 1787–1813 (2000).
84. Yu, L.R., Zeng, R., Shao, X.X., Wang, N., Xu, Y.H. and Xia, Q.C., Identification of differentially expressed proteins between human hepatoma and normal liver cell lines by two-dimensional electrophoresis and liquid chromatography-ion trap mass spectrometry. *Electrophoresis*, **21**, 3058–3068 (2000).
85. Hanash, S.M., Biomedical applications of two-dimensional electrophoresis using immobilized pH gradients: current status. *Electrophoresis*, **21**, 1202–1209 (2000).
86. Heller, T., Saito, S., Auerbach, J., Williams, T., Moreen, T.R., Jazwinski, A., Cruz, B., Jeurkar, N., Sapp, R., Luo, G. and Liang, T.J., An *in vitro* model of hepatitis C virion production. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2579–2583 (2005).
87. Lindenbach, B.D., Evans, M.J., Syder, A.J., Wolk, B., Tellinghuisen, T.L., Liu, C.C., Maruyama, T., Hynes, R.O., Burton, D.R., McKeating, J.A. and Rice, C.M., Complete replication of hepatitis C virus in cell culture. *Science,* **5734,** 623–626 (2005).
88. Major, M.E., Rehermann, B. and Feinstone, S.M., In: Howley, P.M. (Ed.), *Fields Virology* (Vol. 2, 4th edition). Lippincott Williams & Wilkins, Philadelphia, 2001, pp. 1127–1183.
89. Blight, K.J., Kolykhalov, A.A. and Rice, C.M., Efficient initiation of HCV RNA replication in cell culture. *Science*, **290**, 1972–1974 (2000).
90. Satoh, S., Hirota, M., Noguchi, T., Hijikata, M., Handa, H. and Shimotohno, K., Cleavage of hepatitis C virus nonstructural protein 5A by a caspase-like protease(s) in mammalian cells. *Virology*, **270**, 476–487 (2000).
91. Lan, K.H., Sheu, M.L., Hwang, S.J., Yen, S.H., Chen, S.Y., Wu, J.C., Wang, Y.J., Kato, N., Omata, M., Chang, F.Y. and Lee, S.D., HCV NS5A interacts with p53 and inhibits p53-mediated apoptosis. *Oncogene*, **21**, 4801–4811 (2002).
92. Concannon, C.G., Gorman, A.M. and Samali, A., On the role of Hsp27 in regulating apoptosis. *Apoptosis*, **8**, 61–70 (2003).

93. Hanash, S., Disease proteomics. *Nature*, **422**, 226–232 (2003).
94. Boguski, M. and McIntosh, M., Biomedical informatics for proteomics. *Nature*, **422**, 233–237 (2003).
95. Gorg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R. and Weiss, W., The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **21**, 1037–1053 (2000).
96. Gorg, A., Weiss, W. and Dunn, M., Current two-dimensional electrophoresis technology for proteomics. *Proteomics*, **4**, 3665–3685 (2004).
97. Chen, J., Ueda, K., Sakakibara, S., Okuno, T., Parravicini, C., Corbellino, M. and Yamanishi, K., Activation of latent Kaposi's sarcoma-associated herpesvirus by demethylation of the promoter of the lytic transactivator. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 4119–4124 (2001).
98. Lamerz, R., Runge, M., Stieber, P. and Meissner, E., Use of serum PIVKA-II (DCP) determination for differentiation between benign and malignant liver diseases. *Anticancer Res.*, **19**, 2489–2493 (1999).
99. Steel, L., Shumpert, D., Trotter, M., Seeholzer, S., Evans, A., London, W., Dwek, R. and Block, T., A strategy for the comparative analysis of serum proteomes for the discovery of biomarkers for hepatocellular carcinoma. *Proteomics*, **3**, 601–609 (2003).
100. Comunale, M.A., Mattu, T.S., Lowman, M.A., Evans, A.A., London, W.T., Semmes, O.J., Ward, M., Drake, R., Romano, P.R., Steel, L.F., Block, T.M. and Mehta, A., Comparative proteomic analysis of de-N-glycosylated serum from hepatitis B carriers reveals polypeptides that correlate with disease status. *Proteomics*, **4**, 826–838 (2004).
101. Yokoyama, Y., Kuramitsu, Y., Takashima, M., Iizuka, N., Toda, T., Terai, S., Sakaida, I., Oka, M., Nakamura, K. and Okita, K., Proteomic profiling of proteins decreased in hepatocellular carcinoma from patients infected with hepatitis C virus. *Proteomics*, **4**, 2111–2116 (2004).
102. Parker, B.W., Kaur, G., Nieves-Neira, W., Taimi, M., Kohlhagen, G., Shimizu, T., Losiewicz, M.D., Pommier, Y., Sausville, E.A. and Senderowicz, A.M., Early induction of apoptosis in hematopoietic cell lines after exposure to flavopiridol. *Blood*, **91**, 458–465 (1998).
103. Bonner, R., Emmert-Buck, M., Cole, K., Pohida, T., Chuaqui, R., Goldstein, S. and Liotta, L., Laser capture microdissection: molecular analysis of tissue. *Science*, **278**, 1481–1483 (1997).
104. Hunt, D.F., Yates, J.R., 3rd, Shabanowitz, J., Winston, S. and Hauer, C.R., Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 6233–6237 (1986).
105. Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. and Aebersold, R., Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 9390–9395 (2000).
106. Lam, L.T., Pickeral, O.K., Peng, A.C., Rosenwald, A., Hurt, E.M., Giltnane, J.M., Averett, L.M., Zhao, H., Davis, R.E., Sathyamoorthy, M., Wahl, L.M., Harris, E.D., Mikovits, J.A., Monks, A.P., Hollingshead, M.G., Sausville, E.A. and Staudt, L.M., Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol. *Genome Biol*, **2**, RESEARCH0041.0041–RESEARCH00041.00011 (2001).
107. Wall, D.B., Kachman, M.T., Gong, S., Hinderer, R., Parus, S., Misek, D.E., Hanash, S.M. and Lubman, D.M., Isoelectric focusing nonporous RP HPLC: a two-dimensional liquid-phase separation method for mapping of cellular proteins with identification using MALDI-TOF mass spectrometry. *Anal. Chem.*, **72**, 1099–1111 (2000).
108. Maurer, H.H., Multi-analyte procedures for screening for and quantification of drugs in blood, plasma, or serum by liquid chromatography-single stage or tandem mass spectrometry (LC-MS or LC-MS/MS) relevant to clinical and forensic toxicology. *Clin. Biochem.*, **38**, 310–318 (2005).
109. Wu, S.L., Amato, H., Biringer, R., Choudhary, G., Shieh, P. and Hancock, W.S., Targeted proteomics of low-level proteins in human plasma by LC/MSn: using human growth hormone as a model system. *J. Proteome Res.*, **1**, 459–465 (2002).

110. Nelson, R.E., Grebe, S.K., DJ, O.K. and Singh, R.J., Liquid chromatography-tandem mass spectrometry assay for simultaneous measurement of estradiol and estrone in human plasma. *Clin. Chem.*, **50**, 373–384 (2004).
111. Tuthill, C.W., Rudolph, A., Li, Y., Tan, B., Fitzgerald, T.J., Beck, S.R. and Li, Y.X., Quantitative analysis of thymosin alpha1 in human serum by LC-MS/MS. *AAPS Pharm. Sci. Tech.*, **1**, E11 (2000).
112. Poon, T.C., Chan, K.C., Ng, P.C., Chiu, R.W., Ang, I.L., Tong, Y.K., Ng, E.K., Cheng, F.W., Li, A.M., Hon, E.K., Fok, T.F. and Lo, Y.M., Serial analysis of plasma proteomic signatures in pediatric patients with severe acute respiratory syndrome and correlation with viral load. *Clin. Chem.*, **50**, 1452–1455 (2004).
113. Davies, D.H., Liang, X., Hernandez, J.E., Randall, A., Hirst, S., Mu, Y., Romero, K.M., Nguyen, T.T., Kalantari-Dehaghi, M., Crotty, S., Baldi, P., Villarreal, L.P. and Felgner, P.L., Profiling the humoral immune response to infection by using proteome microarrays: high-throughput vaccine and diagnostic antigen discovery. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 547–552 (2005).
114. Laemmli, U.K., Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680–685 (1970).
115. O'Farrell, P.H., High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.*, **250**, 4007–4021 (1975).
116. Pumfery, A., de la Fuente, C.R. Berro, Nekhai, S.F. Kashanchi and Chao., S.H. Potential use of pharmacological cyclin dependent kinase inhibitors as anti HIV therapeutics. *Curr. Pharm. Des.* **12**, 1949–1961 (2005).
117. de la Fuente, C., Maddukuri, A., Kehn, K., Baylor, S.Y., Deng, L., Pumfery, A. and Kashanchi, F., Pharmacological cyclin-dependent kinase inhibitors as HIV-1 antiviral therapeutics. *Curr. HIV Res.*, **1**, 131–152 (2003).
118. Chao, S.H., Fujinaga, K., Marion, J.E., Taube, R., Sausville, E.A., Senderowicz, A.M., Peterlin, B.M. and Price, D.H., Flavopiridol inhibits P-TEFb and blocks HIV-1 replication. *J. Biol. Chem.*, **275**, 28345–28348 (2000).
119. Georgiou, H.M., Rice, G.E. and Baker, M.S., Proteomic analysis of human plasma: failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. *Proteomics*, **1**, 1503–1506 (2001).
120. Pieper, R., Su, Q., Gatlin, C.L., Huang, S.T., Anderson, N.L. and Steiner, S., Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics*, **3**, 422–432 (2003).
121. Ahmed, N., Barker, G., Oliva, K., Garfin, D., Talmadge, K., Georgiou, H., Quinn, M. and Rice, G., An approach to remove albumin for the proteomic analysis of low abundance biomarkers in human serum. *Proteomics*, **3**, 1980–1987 (2003).
122. Baussant, T., Bougueleret, L., Johnson, A., Rogers, J., Menin, L., Hall, M., Aberg, P.M. and Rose, K., Effective depletion of albumin using a new peptide-based affinity medium. *Proteomics*, **5**, 973–977 (2005).
123. Petricoin, E.F. and Liotta, L.A., SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. *Curr. Opin. Biotechnol.*, **15**, 24–30 (2004).

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 16

Neonatal research

BRUNO CASETTA*

Applied Biosystems, Monza, Italy

*Tel.: +39-039-8389-1; Fax: +39-039-8389-497.
E-mail: Bruno.Casetta@eur.appliedbiosystems.com (B. Casetta).

1. Introduction

Generally speaking, and from the mass spectrometry perspective, "neonatal research" is viewed mainly as newborn screening (NBS) which can be defined as the presymptomatic identification of the most commonly known inborn errors of metabolism (IEM).

Mass spectrometric techniques for the identification of metabolic disorders have been developed since the early 1970s. Before the recent advent of tandem mass spectrometry coupled to liquid chromatography (LC–MS/MS), these methodologies were based on mass spectrometry coupled to gas chromatography (GC–MS). They were time-consuming and unable to handle non-volatile compounds unless a chemical derivatization step was performed. Today GC–MS techniques are still used in the clinical laboratories but they are more devoted for confirmation of diagnosis, now the LC–MS/MS unquestionably being recognized as the "tool" for any NBS program.

A lot of papers have recently appeared either as reviews (a very exhaustive one is in ref. [1]) or as chronicles [2] concerning the technical developments of NBS with tandem mass spectrometry (LC–MS/MS).

If deeply interested in some specific details, the reader can refer to those papers. The present chapter would like to give the essential information to both chemists and medical professionals and make them able to leverage this incontrovertible opportunity in detecting metabolic disorders.

The largest part of the present work is devoted to the original format of NBS methodology (centered on acyl-carnitine (AC) profile and most of the amino acids (AA)) with particular attention to some practical issues. A second part will cover the characterization of very long chain fatty acids, steroids, bile acids (BA), and other markers which are gaining relevance in the NBS context.

1.1. Essential medical concepts

Mutations of genes encoding enzymes or transport proteins lead to IEM. Disruption of most of the enzyme functions causes in the body fluids an accumulation of the substrate of the affected enzyme reaction. This accumulation can even trigger alternate biochemical reactions leading to some unusual metabolites. The aim of the NBS approach is to detect these metabolites (either as substrate of the affected specific enzyme activity or as unnatural metabolite secondarily produced in the biochemical reaction chain).

In the original protocol (still the most used one in the large-scale programs), the investigated markers are the AA and the AC. Their detection enables the identification of the involved inherited metabolic disorders (also termed IEM) in the neonatal period.

Recent reports claim the ability to screen more than 30 metabolic disorders with a single analysis. Early medical intervention can avoid irreversible damages like physical and mental retardation, if not death.

Most of the times, for the new adopters of LC–MS/MS in NBS, protocol is viewed in parallel with one of the most known screening tests for screening the phenylketonuria (PKU), introduced in 1963 by Guthrie and Susi [3]. It leverages a bacterial growth inhibition assay for quantifying the phenylalanine in neonate blood. Despite being questionable about precision and accuracy, the test has been widely accepted because it is fast, easy in collecting and transferring the specimen (heel-blood of the infant is dried on a filter paper), and cheap.

NBS is sharing with LC–MS/MS the same sample collection (filter paper collection and handling is now referred as dried blood spot (DBS)) and sometimes this creates some confusion in the terminology. For example, "Guthrie" card was intended to be the filter paper where blood is collected and today, with the adoption of LC–MS/MS, this definition is still used, even if LC–MS/MS has nothing to do with the "Guthrie" test (bacterial growth inhibition assay).

1.2. Basic concept of using LC–MS/MS technology in the clinical domain

Historically NBS can be considered as the Trojan horse for making LC–MS/MS technology accepted in the clinical domain, and for making it now widely considered.

LC–MS/MS technology introduces a new concept in the clinical laboratory, no more one parameter per test, but more parameters per test. NBS with LC–MS/MS, as depicted in the original protocol, is not limited in characterizing, for example, the PKU metabolic disease, but some 30 different altered metabolic functions, all with a running time of 2–3 min.

LC–MS/MS, as widely evidenced in other analytical domains like the pharmaceutical one, is characterized by high sensitivity and high linear range, features which are elective for any quantitation job, all associated with unsurpassed specificity.

Mass spectrometry detects the mass of any required molecule through a "mass analyzer" and measures as well the "amount" of that specific molecule. Considering the huge number of any possible molecules existing around us, sometimes mass spectrometry is not selective enough: a lot of "small" molecules have the same mass. For gaining the required selectivity, mass spectrometry implements two "mass analyzers" with a special interleaved cell where fragmentation of the analyzed molecule is induced (now tandem mass spectrometry or MS/MS). High specificity is now provided by the two mass analyzers working in tandem and filtering, one the mass of the analyzed molecule and the other the mass of the fragment generated by that molecule when its dissociation is performed in the interposed collision cell (Fig. 1).

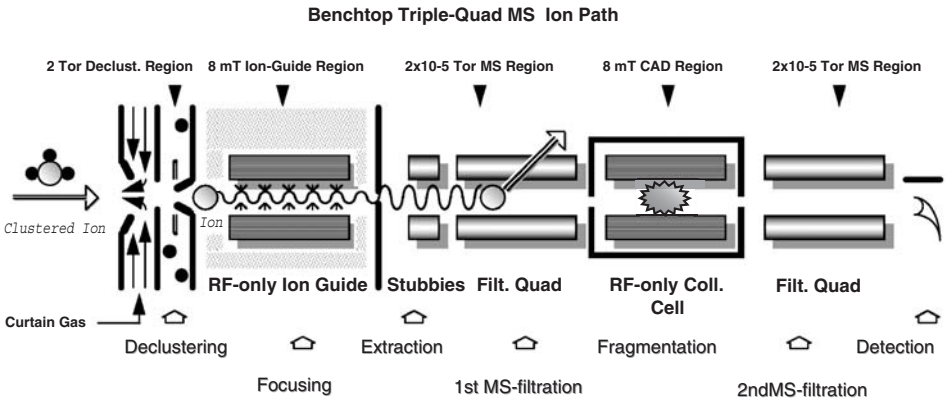


Fig. 1. Schematic of a commercial tandem mass spectrometer.

Such a high specificity accounts for fast running tests (like NBS) since chromatography separation can be skipped. The high specificity enables as well the multi-analyte acquisitions (more analytes measured within the same run).

Staying with the comparison with the classical "Guthrie" test, LC–MS/MS quantifies not only phenylalanine, but also several other "strategically important" AA (for example, tyrosine, and the AA involved in the urea cycle) and, in the same run, the AC all with high precision and accuracy.

2. Principle of the methodology

As with the original definition, neonatal screening refers to a rapid mass spectrometric measurement of all the most prominent AC and the most of the AA, all as markers of possible IEM (namely AA disorders, fatty acid oxidation disorders, and lysosomal disorders).

Chace and Millington [4–7] can be considered the pioneers of this methodology, proposed some 20 years ago. At the very beginning, fast atom bombardment (FAB) instrumentation was used. It is at the beginning of 1990s that methodology acquired the actual design (butylation of the sample extract plus electrospray ionization (ESI) with a triple-quad instrument performing tandem mass measurement through precursor ion scan and neutral loss scan). Besides the above-mentioned pioneers, Rashed [8] must be cited as one of the prominent users of this technology. He leveraged the protocol for characterizing the relative high incidence of cases in his country, acquiring a significant experience on "positive" cases. Today more and more researchers must be accounted for having developed significant expertise in the collection of statistically relevant number of tested babies [9]. Some of them have also extended the study over the adult population [10].

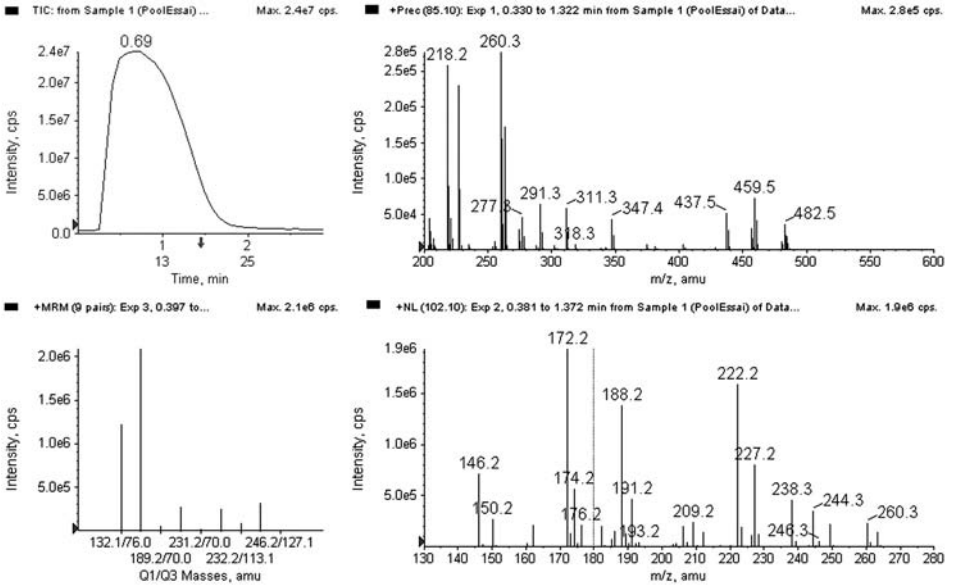


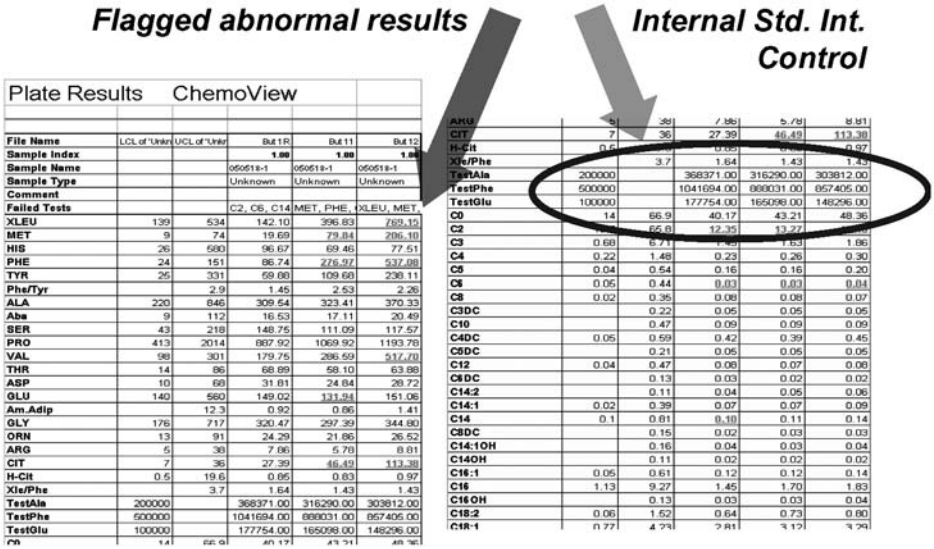
Fig. 2. Example of raw data generated by a NBS protocol measurement. (Upper left panel) TIC trace. (Upper right panel) Precursor ion scan for AC profile. (Lower right panel) Neutral loss scan for AA profile. (Lower left panel) MRM readings for glycine, ornithine, citrulline, arginine, and homocitrulline.

The actual format is still the most used one, even if more and more users are now implementing some modifications, as described below.

Filter paper hosting the DBS is punched and extracted with methanol containing a cocktail of either AA or AC, all as isotopically labeled. The extract is butylated and the resulting butylated esters are subjected to the mass spectrometric measurement without any prior chromatographic separation (flow-injection analysis (FIA)) but implementing either a neutral loss scan for characterizing the AA profile, a precursor ion scan for depicting the AC profile, or a multiple reaction monitoring for quantifying some specific AA such as the "basic" ones (arginine, ornithine, citrulline, homocitrulline, etc.) (Fig. 2). Collected raw data are automatically processed for expressing the original concentration of all the desired analytes and for conveniently flagging the results falling out of the normal range as defined by the medical professionals (Fig. 3).

3. Quick reference for practical implementation of the methodology

As mentioned, the number of published analytical protocols is in the same order of magnitude as the number of papers that has appeared on this specific topic. Each researcher has slightly or deeply modified his own method and, for experience of



File Name	LCL of "Unkn	UCL of "Unkr	But1R	But11	But12
Sample Index			1.00	1.00	1.00
Sample Name			050518-1	050518-1	050518-1
Sample Type			Unknown	Unknown	Unknown
Comment					
Failed Tests			C2, C6, C14	MET, PHE,	XLEU, MET,
XLEU	139	534	142.10	396.83	769.15
MET	9	74	19.69	79.84	206.10
HIS	26	580	96.67	69.46	77.51
PHE	24	151	86.74	276.97	537.08
TYR	25	331	69.86	109.68	238.11
Phe/Tyr		2.9	1.45	2.53	2.26
ALA	220	846	309.54	323.41	370.33
Aba	9	112	16.53	17.11	20.49
SER	43	218	148.75	111.09	117.57
PRO	413	2014	887.92	1069.92	1193.78
VAL	98	301	179.75	286.59	517.70
THR	14	86	68.89	58.10	63.88
ASP	10	68	31.81	24.84	28.72
GLU	140	560	149.02	131.94	151.06
Am.Adip		12.3	0.92	0.86	1.41
GLY	176	717	320.47	297.39	344.80
ORN	13	91	24.29	21.86	26.52
ARG	5	38	7.86	5.78	8.81
CIT	7	36	27.39	46.49	113.38
H-Cit	0.5	19.6	0.85	0.83	0.97
Xle/Phe		3.7	1.64	1.43	1.43
TestAla	200000		368371.00	316290.00	303812.00
TestPhe	500000		1041694.00	888031.00	857405.00
TestGlu	100000		177754.00	165098.00	148296.00
C0	14	66.9	40.17	43.21	48.36
C2					
C3	0.68	6.71		1.83	1.86
C4	0.22	1.48	0.23	0.26	0.30
C5	0.04	0.54	0.16	0.16	0.20
C6	0.05	0.44	0.03	0.03	0.04
C8	0.02	0.35	0.08	0.08	0.07
C3DC		0.22	0.05	0.05	0.05
C10		0.47	0.09	0.09	0.09
C4DC	0.05	0.59	0.42	0.39	0.45
C5DC		0.21	0.05	0.05	0.05
C12	0.04	0.47	0.08	0.07	0.08
C6DC		0.13	0.03	0.02	0.02
C14:2		0.11	0.04	0.05	0.06
C14:1	0.02	0.39	0.07	0.07	0.09
C14	0.1	0.81	0.10	0.11	0.14
C8DC		0.15	0.02	0.03	0.03
C14:1OH		0.16	0.04	0.03	0.04
C14OH		0.11	0.02	0.02	0.02
C16:1	0.05	0.61	0.12	0.12	0.14
C16	1.13	9.27	1.45	1.70	1.83
C16OH		0.13	0.03	0.03	0.04
C18:2	0.06	1.52	0.64	0.73	0.80
C18:1	0.77	4.73	2.81	3.12	3.29

Fig. 3. Example of report produced for medical professionals. Abnormal data are conveniently flagged.

the author in dealing with several laboratories, no one is the clone of any other, either because of the different equipments, the different strategies (see below), and/or the different experimental parameters.

Hereby we present a trace, reflecting somehow the original format, which should be a good starting point for any newcomer in this technology. Vessel types are proposed for beginners and for minute-scale assays: in routine they are swapped to 96-well titer plates.

3.1. Extraction

A circle 3 mm in diameter (roughly corresponding to 3 μL of original blood) is punched out from each spot by means of a standard hole puncher into 1.5 mL polypropylene tube. Specimen is pricked-heel blood subsequently dried on Schleicher&Schuell filter paper (in Europe Grade 903 is used at the most).

The spot is treated at room temperature for 20 min with methanol (200 μL) containing known amounts of stable-isotope-labeled internal standards for AA and AC as commercially available in a cocktail. The supernatant liquid, containing the sample extract, is transferred to glass autosampler vial (screw-capped).

3.2. Butylation

Solvent is evaporated at 55°C under a gentle stream of nitrogen. Once perfectly dried, sample is then redissolved in the derivatizing reagent (80 μL of 3N HCl in *n*-butanol). Vial is capped, rotated, and heated at 65°C for 20 min.

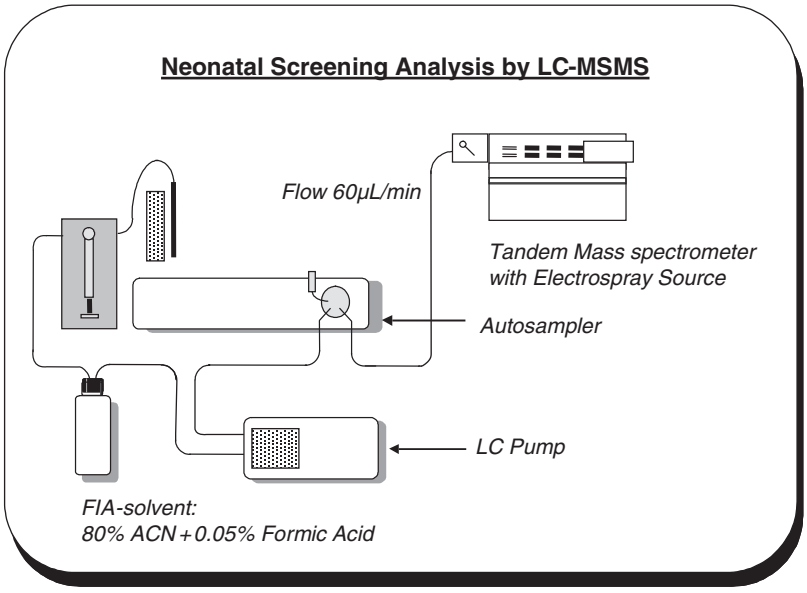


Fig. 4. Example of plumbing for performing the NBS protocol.

The resulting mixture is once again dried (at 55°C under a gentle stream of nitrogen), reconstituted with 200 µL of solvent (80% acetonitrile, 20% water, and 0.05% formic acid), and put in the autosampler, ready for mass spectrometric measurement.

3.3. Analytical equipment

A tandem mass spectrometer must be equipped with an electrospray source and is plumbed as in Fig. 4. LC-pump and an appropriate LC autosampler are used for solvent delivery and automated sample introduction. The mobile phase is acetonitrile:water mixture (80:20, v/v) with 0.05% formic acid at a flow rate of 60 µL/min. The autosampler is programmed to inject a volume of 40 µL of the sample.

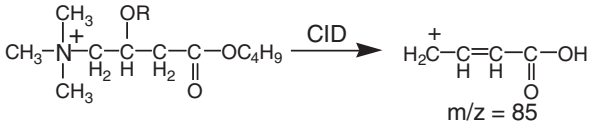
3.4. Analytical measurement

Tandem mass spectrometric reading exploits some specific features of butylated AC and AA. By fragmentation, the AC produce a prominent fragment ion at *m/z* 85, common to all of them (see scheme in Fig. 5). When fragmented, the butylated AA produce a fragment which is in mass 102 Th less than the precursor ion, due to a loss of the neutral moiety corresponding to butyl formiate (see scheme in Fig. 5).

Consequently, for AC profiles, the precursor ion scan for the product ion at *m/z* 85 is performed in the range *m/z* 200–600 and with appropriate collision

MS/MS Mechanisms: Acylcarnitines

- Butyl esters of carnitine or acylcarnitine fragment to form the 85^+ ion
- Precursor scan, where Q3 is set to 85^+ and Q1 scanned, reveals all acylcarnitines present in the blood spot extract



MS/MS Mechanisms: Amino Acids

- Butyl esters of amino acids fragment to lose $HCO_2C_4H_9$ (102 Da)
- Neutral loss scan, where Q3 and Q1 are scanned with a 102 Da mass difference, reveals the amino acids present in the blood spot extract

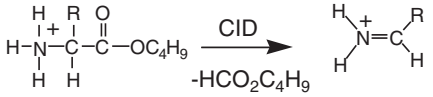


Fig. 5. Mechanisms of MS/MS fragmentation of AC and AA.

energy. Some commercial MS/MS instruments allow to ramp the collision energy and other compound-dependent parameters during the scan in order to have them optimized either for low- or high-mass AC, e.g., short and long chains, respectively (in our laboratory, declustering voltage from 60 to 80 V, collision energy from 35 to 65 eV).

For AA profiles, a neutral loss scan of *m/z* 102 is collected in the range *m/z* 130–280 and with an appropriate collision energy.

For the basic AA (citrulline, homocitrulline, ornithine), glycine and arginine, data are acquired in the multiple reaction monitoring (MRM) mode by monitoring specific transitions with specific collision energies as optimized for the specific instrument.

With a suitable instrument, the above three acquisition experiments (precursor ion scan, neutral loss scan, and MRM) can be cycled as having during all the experiment time (usually between 2 and 3 min) the interleave of each of them.

Raw data are processed either after or during sample batch acquisition.

3.5. Expected performances

Fig. 2 shows a typical total ion chromatogram (TIC) trace obtained by injecting a sample. The neighbor panels show the graphs related to the three concurrent

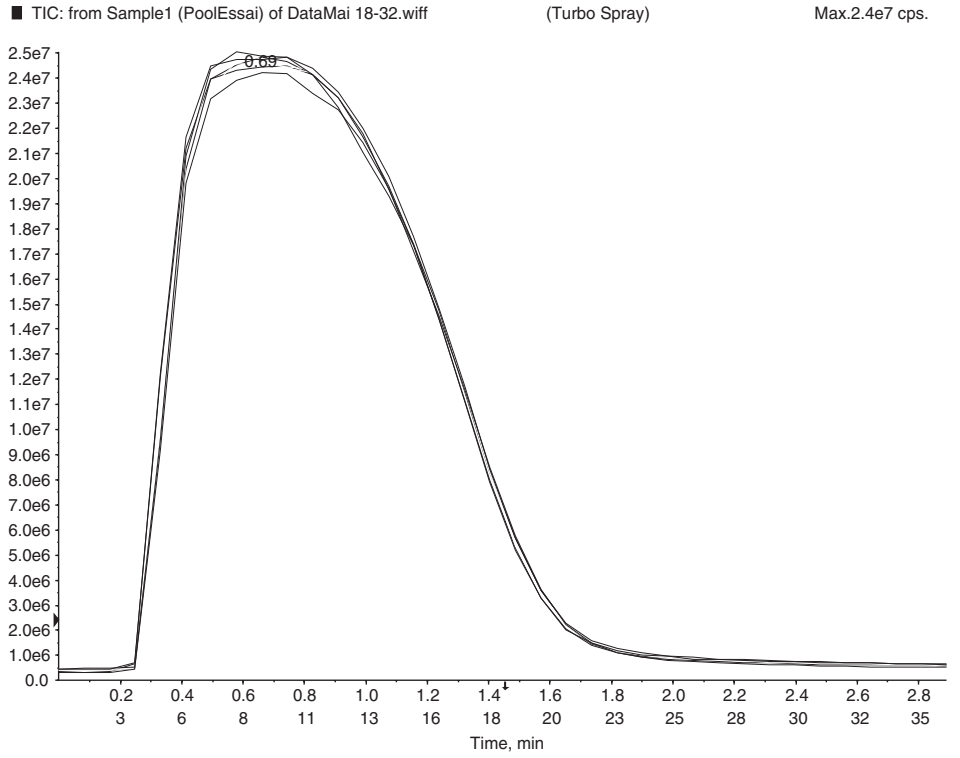


Fig. 6. Overlay of different TIC collected in replicating the MS/MS readings.

experiments. Fig. 6 shows that TIC traces from several injections are essentially superimposable graphically underscoring the high degree of precision of the method and the apparatus despite the repeated injection of crude blood spot extracts.

Linearity test can be assessed by spiking analytes in a specimen and evaluating the resulting calibration curves. Two such curves are shown in Figs. 7 and 8 for octanoyl-carnitine and methionine, respectively. The full set of results is found in the table of Fig. 9. The linear correlations show slopes of between 0.91 and 1.1 and correlation coefficients of between 0.994 and 1.000.

Fig. 10 documents the precision obtained by separately extracting and analyzing six sets of duplicate blood samples at normal concentration levels. Exhibited CV's are between 2.5% for free carnitine at a concentration of 26 μM and 12.6% for octanoyl-carnitine at a concentration of 0.087 μM. Generally, the higher the concentration of the analyte, the better the precision.

Accuracy can be evaluated by a comparison of MS/MS and HPLC results made for phenylalanine. The results are shown in Fig. 11, which demonstrates a linear correlation with a coefficient of 0.997.

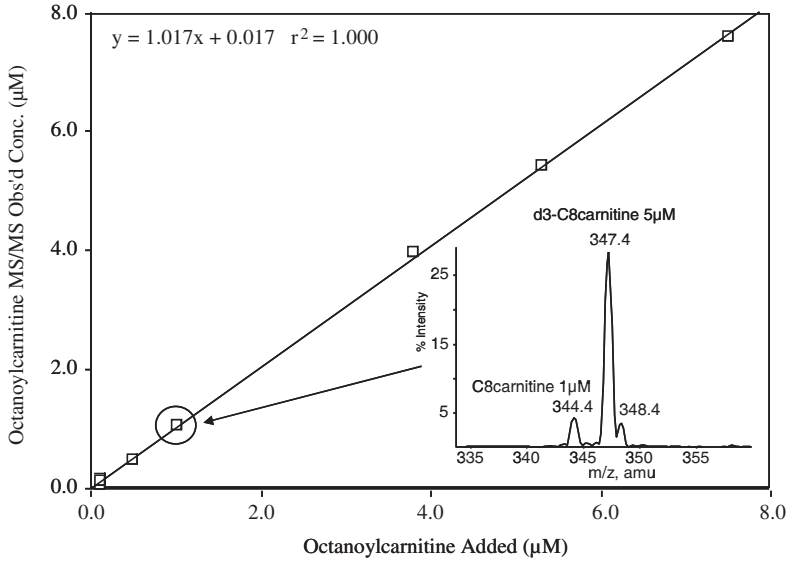


Fig. 7. Linearity test on octanoyl-carnitine (see also table in Fig. 9).

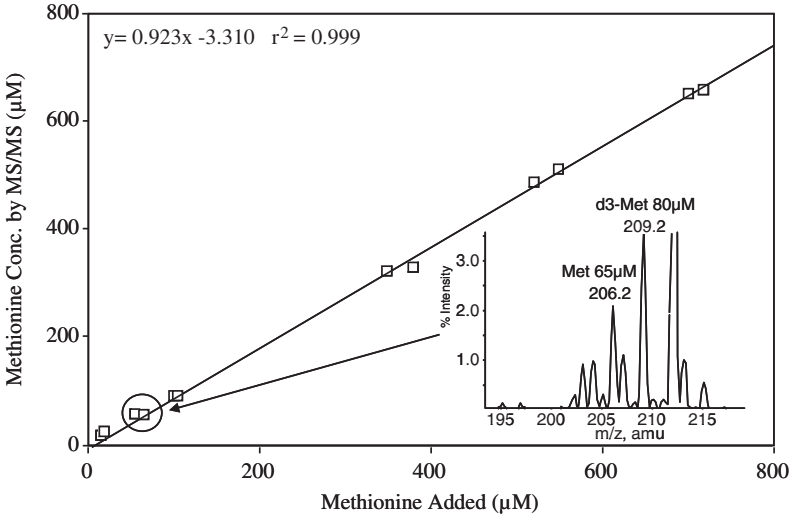


Fig. 8. Linearity test on methionine (see also table in Fig. 9).

Analyte	Slope	Fit (r^2)	CV(%)
Free Carnitine	1.097	1.000	2.7% at 23µM (n=6)
Acetylcarnitine	1.004	0.997	3.3% at 7.6µM (n=3)
Octanoylcarnitine	1.017	1.000	10.9% at 0.1µM (n=6)
Palmatoylcarnitine	0.914	0.994	6.6% at 0.4µM (n=6)
Phenylalanine	0.949	0.994	
Tyrosine	0.933	0.995	
Methionine	0.923	0.999	
Xle	0.990	0.996	

Fig. 9. Results on calibrations curves obtained by spiking a specimen for evaluating the linearity.

Analyte	Average Concentration µM (n=12)	CV (n=6 pairs)
C0	26.025	2.5%
C2	5.720	3.1%
C3	0.412	4.9%
C4	0.157	5.2%
C6	0.064	11.6%
C8	0.087	12.6%
C10	0.049	2.9%
C16	0.353	3.5%
C18	0.178	10.2%
C18:1	0.463	5.6%
	Mean	**6.2%**

Analyte	Average Concentration µM (n=12)	CV (n=6 pairs)
Tyr	59.14	5.4%
Phe	444.34	3.0%
Ala	142.74	2.3%
Val	105.90	3.5%
Xle	171.48	4.8%
Met	19.25	6.3%
Phe/Tyr	9.26	4.9%
	Mean	**4.3%**

Fig. 10. Precision data on acyl-carnitines in replicate blood samples (normal concentration levels, left panel) and on amino acids in replicate blood samples (five out of six are PKU, right panel).

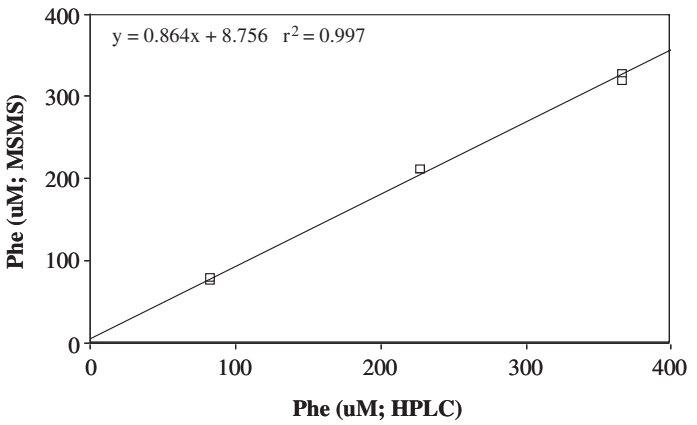


Fig. 11. Correlation between MS/MS and the more classical HPLC assay for phenylalanine.

4. Key points of the methodology

4.1. Quantitation assessment

Measurement is done by simulating the isotope dilution (ID) strategy. For quantifying with good accuracy the endogenous analyte, same molecule but isotopically labeled is spiked at known concentration in the sample itself. Both endogenous and isotopically labeled compounds are characterized by the mass spectrometer, and the quantitation assessment is performed by comparing the two intensities and referring to the known concentration of the spiked standard (also called "internal standard"). Doing so, any error in the manipulation or loss in the subsequent analytical steps is compensated for since any deficiency affects at the same extent either the endogenous analyte or the spiked standard (it is assumed that isotopically labeled molecule has almost all the same physico-chemical properties of the unlabeled compound).

Indeed, in the specific case of NBS, ID strategy is just simulated, it being not possible to spike the internal standards directly in the blood at the collection time but only in the subsequent methanolic extract.

This is one of the most noticeable pitfalls of this analytical protocol. For making accountable the "concentration" of the spiked standards, the estimation of the amount of sample (blood volume) is derived from the diameter of spot originated by drying the collected blood. Since the physical properties of the filter paper and the hematocrit of the sampled blood are influencing the spot diameter of the DBS, it is of paramount importance for the laboratory analyst to correctly estimate the original volume of blood corresponding to the sampled dried spot. Skipping this step leads to significant bias in the final concentration results.

Some guidelines and standardization suggestions are in the way for facilitating the task, such as to preventively select the appropriate vendor of the filter paper and the type of it (in Europe grade 903 from Schleicher&Schuell). A hematocrit of 50% is assumed to be the one corresponding to the heel blood of a neonate and any sampled blood should refer to that value (in North America a hematocrit value approaching 55% is proposed as reference).

Besides this standardization guideline, it is very important that each laboratory makes some tests for assessing the correspondence of sampled spot diameter to the original blood volume.

4.2. Derivatization

More and more users (one example is given in ref. [11]) are proposing to omit the butylation step, owing to the increased sensitivity displayed by the actual commercial instrumentation. Some users prefer as well to skip any spectrum acquisition (precursor ion scan and neutral loss scan) and make all the quantitation measurements exclusively through MRM readings. As reported later, this strategy is enhancing the sensitivity further.

Rationale behind derivatization relates to the chemistry of the analyzed compounds (AA and AC). Both have either a basic functional group (amine for the AA, tertiary amine for the carnitines—free or acylated forms) or an acidic group. Since the yield of the ESI process is related to the overall proton affinity of the molecule itself, by neutralizing the acidity of the carboxylic moiety with esterification, the resulting proton affinity increases and therefore sensitivity is enhanced.

The second and less evident benefit of the derivatization is an equalization of the specific sensitivity of each AC.

The cocktail of isotopically labeled standards is not covering any possible detectable AC. Therefore, for some of them, the internal standard to refer to is not the isotopic homolog but some other compound, very close in terms of structure and mass. For making the measurement free from a significant bias, it is desirable that specific sensitivities displayed by either the analyte or the used internal standard should be as similar as possible. Failing in that, results are significantly biased.

A typical case is represented by the glutaryl-carnitine (C5DC) which cannot count on the presence of its isotopically labeled homolog in the today commercially available internal standard cocktail. Due to the presence of two acidic moieties, its specific sensitivity is lower than any other mono-acidic AC and makes questionable the use of any neighbor compound as internal standard. Butylation is making less exacerbated the difference in proton affinity between mono-acidic and bi-acidic compounds, and therefore it enables the use of a neighbor AC (namely the d3-C8 or d9-C14) as internal standard.

In the case of any dicarboxylic-AC, it is worthy to note that derivatization involves a double butylation reaction (the carnitine carboxylic moiety and the free second carboxylic moiety of the acyl group). Comparing a spectrum from unbutylated specimen with that of a butylated one, shift on mass scale is not 56 Th as expected for the majority of the AC but twice that value (112 Th).

Fig. 12 shows that shift in mass between experiments with butylation and unbutylation is not 56 Th like for the rest of AC but 102 Th (388.4 Th vs. 276.3 Th). By zooming the pertinent spectrum area it is easy to realize the sensitivity difference (a factor of 3 with the instrument so far used), which affects the real detection limit for positive GA-I cases. In the same figure, assuming to take octanoyl-carnitine (C8) as the internal standard for C5DC quantitation in both the experiments, it is interesting to note that in the experiment with butylation, intensity of C5DC is 3.8 times lower than the chosen internal standard; meanwhile, in the experiment without derivatization, C5DC is 8.4 times lower. Therefore, in the case of experiment without butylation, glutary-carnitine result should be biased by underestimation.

In our experience, skipping the butylation step raises a special issue regarding the final solvents used for the analytical measurement. Provided that the DBS extraction is made with methanol, it has been proved that in large-scale routine operation the direct injection of a methanolic solution in the usual flow of ACN/H_2O causes a distortion of the FIA-peak profile, making questionable a long-term stability (Fig. 13).

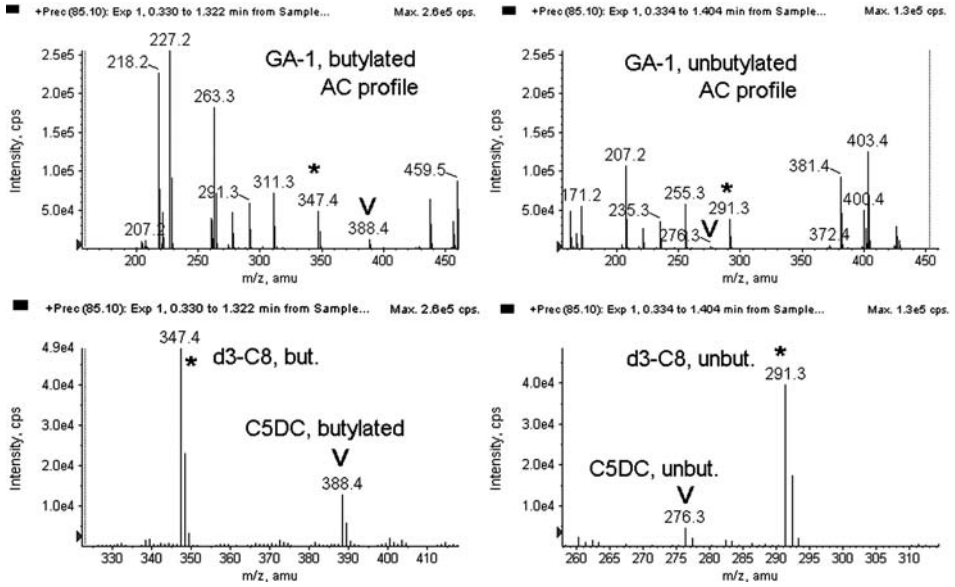


Fig. 12. Comparison of AC profiles on the same sample characterized by a glutaric acidemia type I, with and without butylation. Arrows point where C5DC is nested in the full spectrum. Shift in mass between butylation and unbutylation is not 56 Th like for the rest of AC but 102 Th (388.4 Th vs. 276.3 Th). Assuming to take d3-octanoyl-carnitine (d3-C8) as the internal standard for its quantitation (star-labeled ion) in both the experiments, it is interesting to note that in the experiment with butylation, intensity of C5DC is 3.8 times lower; meanwhile, in the experiment without derivatization, C5DC is 8.4 times lower than the chosen internal standard.

Drying the methanolic extract and reconstituting it with an ACN/H_2O mixture restores the good FIA-peak profile.

However, especially for unbutylated AC, it is mandatory to keep the ACN concentration of the reconstituting mixture quite high (ACN $\geqslant$ 80%) for avoiding any segregating loss of them during the storage in the well plate. We have evidence that polystyrene well plates are prone to make unbutylated AC disappear on long term if not conveniently dissolved (Fig. 14).

Another minor advantage of the butylation is that the chemical treatment of the extract with an acidic media is ending up with a cleaner solution. Repetitive injections of a cleaner solution account for a better long-term robustness.

4.3. FIA flow rate regime

As mentioned, NBS measurement is performed without any LC separation (FIA). The injected plug is moved to the ionization source of the instrument thanks to a carrier flow supplied by an LC-pump. The duty time for acquiring the measurement is dependent on the injected volume, the flow rate, and the dispersion induced by the tubing and the dead volumes downstream the injection valve.

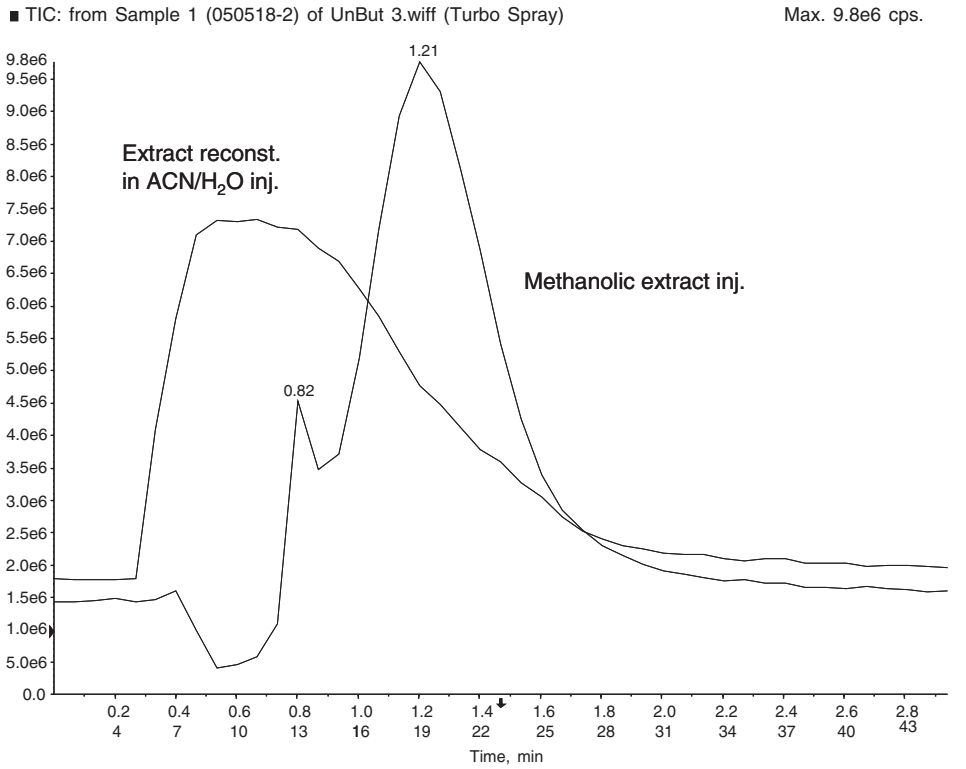


Fig. 13. Peak profiles obtained on the same underivatized sample. Naked methanolic extract, when injected in the ACN/H_2O-mixture carrier, is producing a quite disrupted profile. The same extract, dried and reconstituted with the same ACN/H_2O mixture, leads to a typical smooth FIA-peak profile.

For the analytical reading perspective, a long duration is preferable (more readings on time scale lead to more stable analytical results). On top of that some electrospray sources display higher sensitivity and less ion suppression effect when fed by very low flow rates. However, the big disadvantage is a significant carryover in between the injected samples, which demands a remarkable delay before the next injection, unless in implementing a flow programming in the LC-pump. Program includes: to keep a low-flow-rate regime while the injected plug is penetrating the source and afterwards to move to a high-flow-rate regime for speeding up the washing in between the injections.

With some instruments this is good compromise for having a good yield at the ionization source, while for some others the source is more consistent in the attained sensitivity regardless of the flow rate regime.

In the case of a flow rate programming, it is of paramount importance to have consistency in the appearance and duration times, e.g., starting and ending times of the plug peaks should be reproducible. Lack of it should lead to corrupted results especially in those cases where cycling between the different scanning acquisitions is performed in subsequent periods and not in simultaneous mode (see rationale in the caption of Fig. 15).

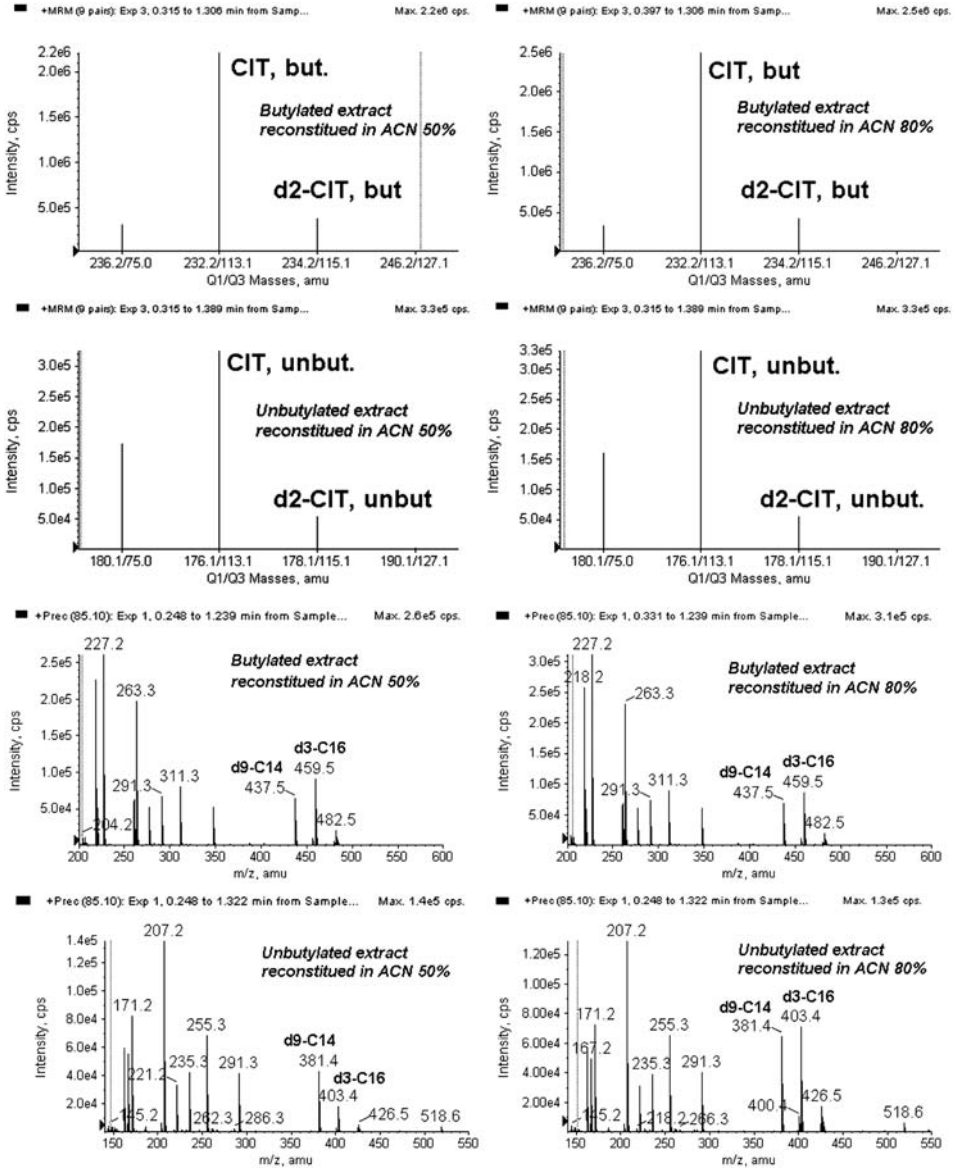


Fig. 14. Comparison of citrulline and AC profiles on the same sample characterized by a citrullinemia, with and without butylation. Upper panel shows as the basic amino acid citrulline quantitation is not dependent on derivatization, except for the absolute sensitivity (sevenfold better with butylation). Either the reconstitution mixture (mixtures of ACN 50% or ACN 80% in water) is not affecting the CIT/d2-CIT ratio. Lower panel shows as with butylation, AC profile is unaffected by the reconstitution mixture composition (upper panes). Without butylation (lower panes) and storing the sample in the well-plate for 1 h @ 4°C, some segregation is occurring at the high-mass AC (compare the ratio d9-C14/d3-C16) when methanolic extract has been dried and reconstituted with ACN 50%. With ACN 80%, ratio is restored and is very similar to the butylated profiles.

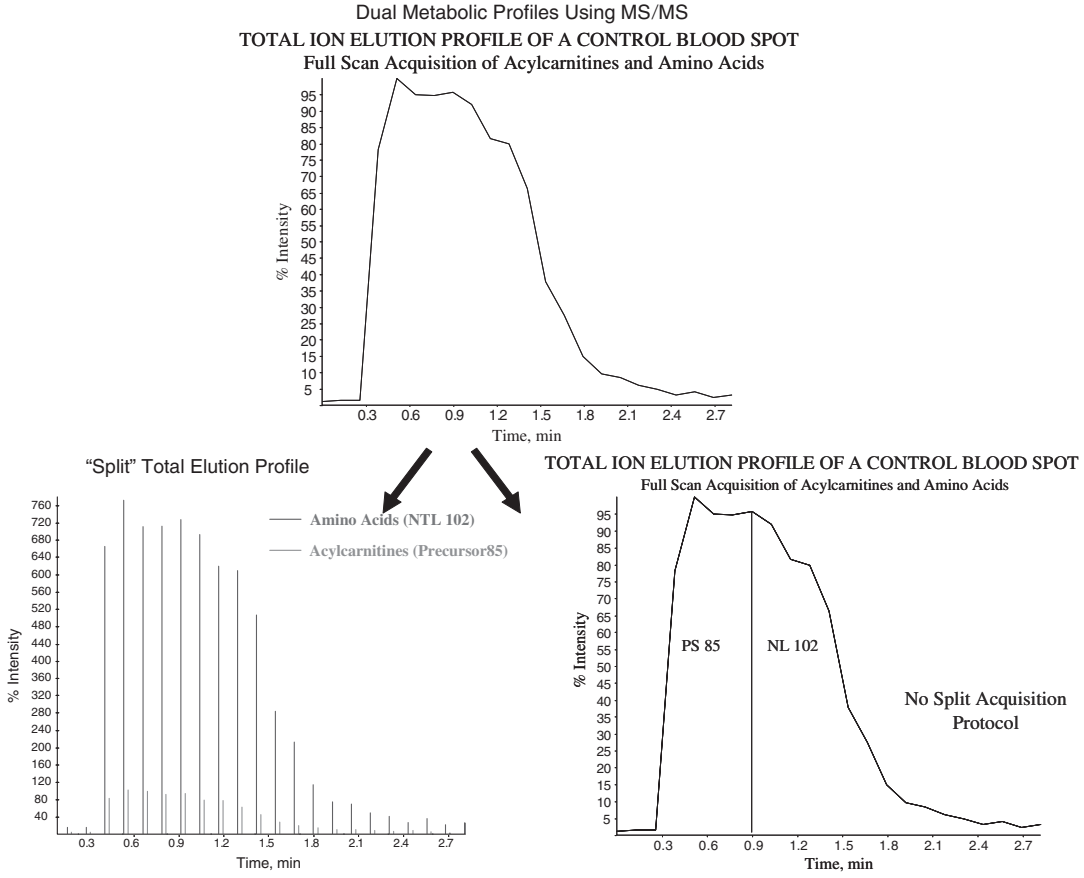


Fig. 15. Different approaches for collecting the various scanning modes from the FIA experiment. For simplicity, just two scanning acquisitions are depicted. Within the same FIA plug reading (upper panel), some instruments are able to cycle the multiple scanning acquisitions in simultaneous mode (lower panel, left). This feature is precluded to some other commercial instruments: in this case the different scanning acquisitions are performed in subsequent periods (lower panel, right). In this case it is of paramount importance to preserve the integrity of the peak shape, lack of it leading to an information degradation from one of the scanning acquisitions.

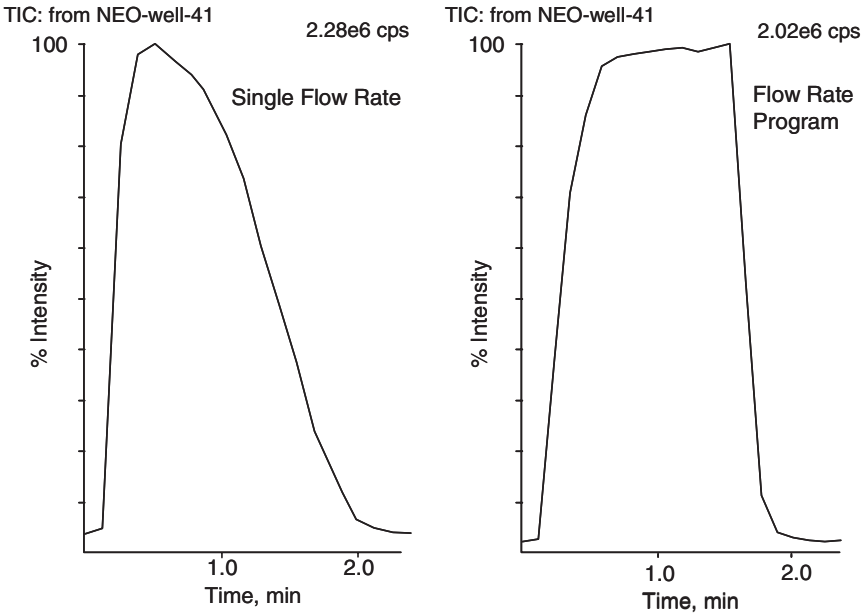


Fig. 16. TIC profiles from single vs. programmed flow experiments. In both experiments data acquisition was over in ca. 2.3 min interval. The type of electrospray source did not display significant differences sensitivity-wise between a constant FIA-flow (left-hand trace) and a programmed flow regime (right-hand trace). In the latter, trace showed a ca. 1 min wide flat-topped TIC peak, which reached baseline after 1.9 min. For both experiments, measured carryover was very little (typically $<1\%$). However, the data obtained using the flow-rate program exhibited about half the carryover of that of the single flow-rate experiment.

As already mentioned, some instruments do not show big sensitivity changes with the carrier flow regime (see Fig. 16). With constant-flow-rate regime reproducibility is much more preserved (Fig. 17).

4.4. MS scanning strategies

For improving the sensitivity, more and more authors [12] prefer to make all the measurements concerning the NBS by MRM, owing to the dwell time per analyte being longer than it should be when a real scanning is implemented.

MRM reading implies the setting of the two analyzers at predefined masses for each attributable analyte, with the first mass representing its pseudo-molecular ion and the second mass the prominent ion resulting from the fragmentation of that specific analyte.

In this configuration the tandem mass spectrometer displays the best performances in quantitation jobs. In fact, sensitivity and reproducibility are directly

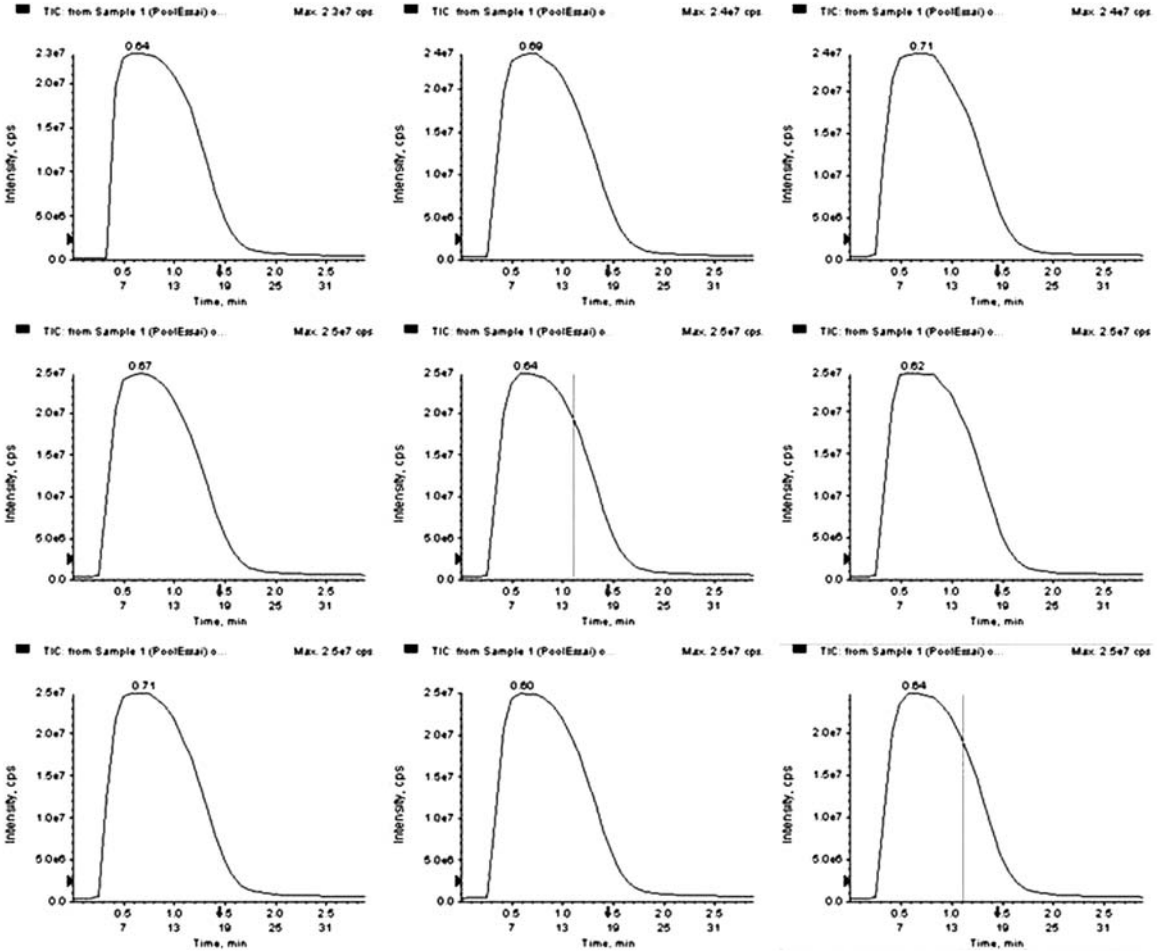


Fig. 17. Evidence of long-term reproducibility of the FIA-peak profile in an NBS program.

related to the time spent for any single analyte measurement. In collecting a spectrum, time is spent for any single-step unit of the spectrum, regardless of whether the bin represents a putative analyte ion or just a background in between other analytes.

In the original NBS protocol, two kinds of scanning are proposed: a precursor ion scan over the fragment at 85 Th for characterizing the AC and a 102 Th-neutral loss scan for characterizing the main AA. For the latter, swapping from the neutral loss scan into a panel of the correspondent MRM does not affect the results quality: the selection of expected AA is predictable.

However, for the AC profile, skipping the scanning (in precursor ion) can have a severe impact on the final results' interpretation by the medical professionals.

The number of discernable analytes in the AC profile is unpredictable: in addition to the common ones (e.g., free −C0 and acetyl-carnitine −C2−, some long chain ones —> C16 –), there can be some generated by the known fatty acid oxidation disorders, some others from "external interferences" (several authors have reported the presence of some carnitines as generated by some specific patient regimen or by administered medication drugs), and some from very rare diseases.

All the above can be captured by a full scan reading (precursor ion scan); meanwhile, just MRM readings can miss some of them if not preventively programmed.

In addition to that, a further benefit in inspecting a spectrum is to free the interpretation from aberrations. A case which has been recently experienced is a glutamate formimino-transferase deficiency characterized by a high concentration of formimino glutamic acid (FIGLU) [13]. As documented by Fig. 18, if reading was made by MRM, this rare disorder should be interpreted as a short-chain acyl-CoA dehydrogenase (SCAD) deficiency since a positive signal is produced at the transition 288 > 85, usually assigned to the C4 carnitine. By coincidence FIGLU produces a fragment ion at 85 Th (same nominal fragment ion mass as the signature fragment ion of AC) but its pseudo-molecular ion is at 287 Th.

By performing a full-scan acquisition in precursor ion mode it has been easy to realize that the prominent ion was at 287 Th and the 288 Th ion was just its C13-isotopomer, therefore avoiding to attribute the high level of the 288 Th ion erroneously to a C4 carnitine and consequently to an SCAD disorder.

4.5. Data processing

In order to make results quickly readable by the medical professionals, mass spectrometric raw data are conveniently processed for relieving the final user from the burden to deal with the minute interpretation of any MS signal produced (see Fig. 2).

Different instrument vendors are supplying different application software for achieving the task, each of them with very different features and functionalities.

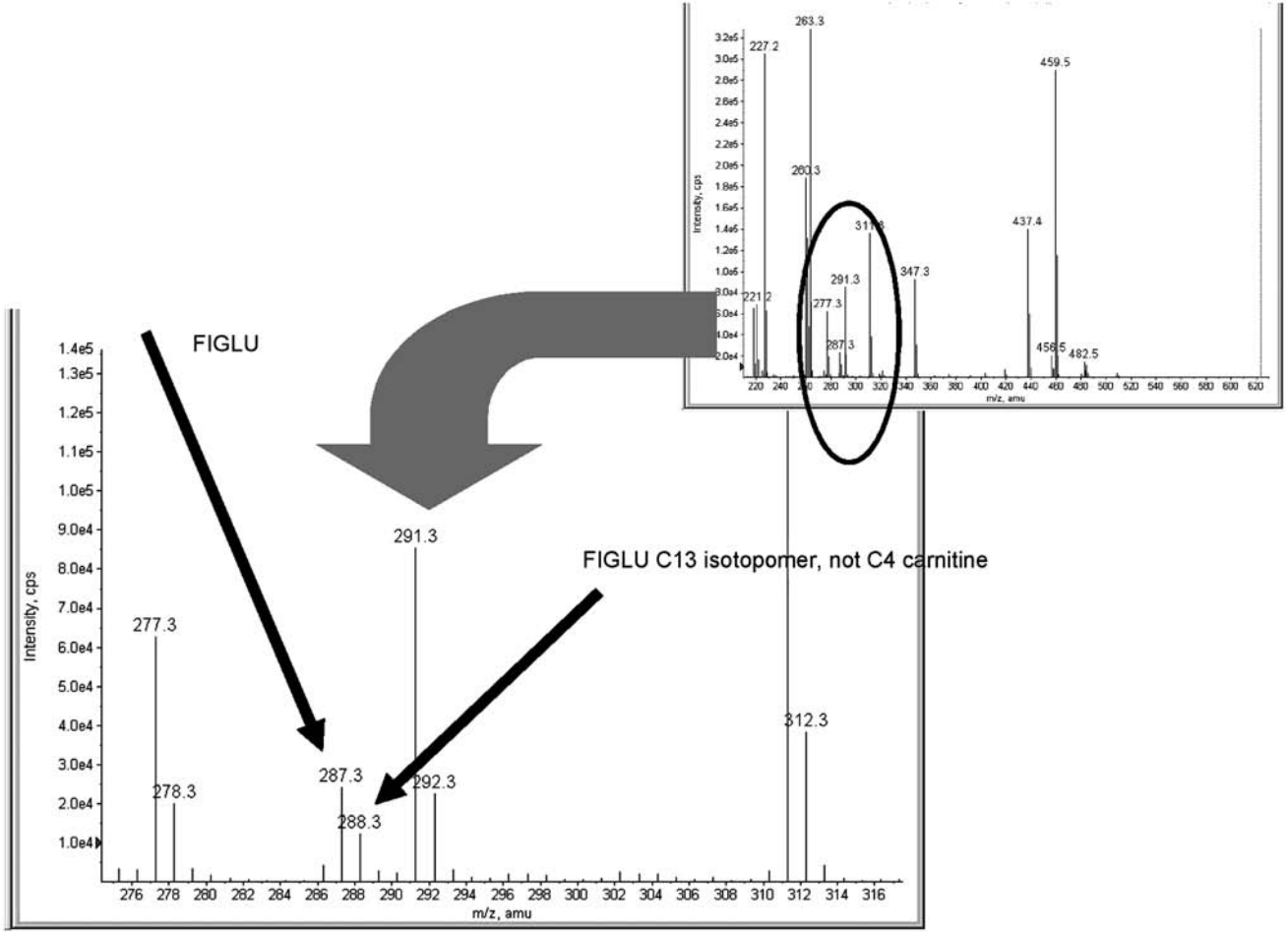


Fig. 18. AC profile from a case of glutamate formimino-transferase deficiency. Ion at 288 Th is not due to C4-carnitine but it is just the C13 isotopomer of FIGLU with main ion at 287 Th.

For example, some applications are able to flag the final results according to different limit suites, each of them specific for a specimen type (e.g., premature, newborn, one year old, etc.).

Beyond the goodies provided by the different application packages, it is worthy to leverage two key features for guaranteeing good performances in large-scale routine programs and avoiding false positives: internal standard intensity monitoring and graphical evidence.

The analytical result is calculated upon the ratio of the intensities of the targeted analyte and its internal standard (isotopically labeled standard). Intensity ratio is trustable as long as the internal standard is not approaching a zero value, the latter caused by some experimental errors in the measurement step or in the upstream sample processing. A final reported result for any analyte is worthy as long as the internal standard intensity is adequate (and not approaching zero, implying that the analyte value is getting mistakenly high). Some application packages are able to flag the samples when the internal standard intensities are getting lower than prefixed thresholds (experimentally and statistically found) for any reason and consequently faith on the displayed final results is questionable (see Fig. 3).

Second key point is related to the full-spectrum acquisition for AC. In the case some AC are evidenced as "abnormal," it is mandatory that the final result report is substantiated by the full spectrum in order to reveal any unpredictable profile (or single ion) which can reveal the source of the abnormal result. Fig. 19 documents, for example, that the isotopically labeled internal standard itself can be a source of artifacts because of some degradation.

5. Ongoing extensions of NBS

5.1. Extended panel of amino acids

List of AA covered by the original NBS protocol is quite limited (roughly 20 AA). It should be appealing to encompass all the AA and the AA-like compounds as done by the most commonly used IEX-ninhydrin-based method.

Several authors have proposed different approaches for enabling all the AA measured by LC–MS/MS, either through a derivatization step [14] or without [15]. In all the proposed protocols, a chromatographic separation is envisaged for resolving the isobaric AA (for example, HYP, allo-ILE, ILE, LEU, or β-ALA, ALA, SAR, or LYS, GLN), implying 15–30 min per analytical run.

Sensitivity, precision, and long-term robustness are getting comparable if not better than the traditional IEX-ninhydrin methodology.

Fig. 20 shows a typical tracing obtained in our laboratory on plasma serum analyzed by LC–MS/MS without derivatization and exploiting, with some modifications, the protocol recently proposed [15,16].

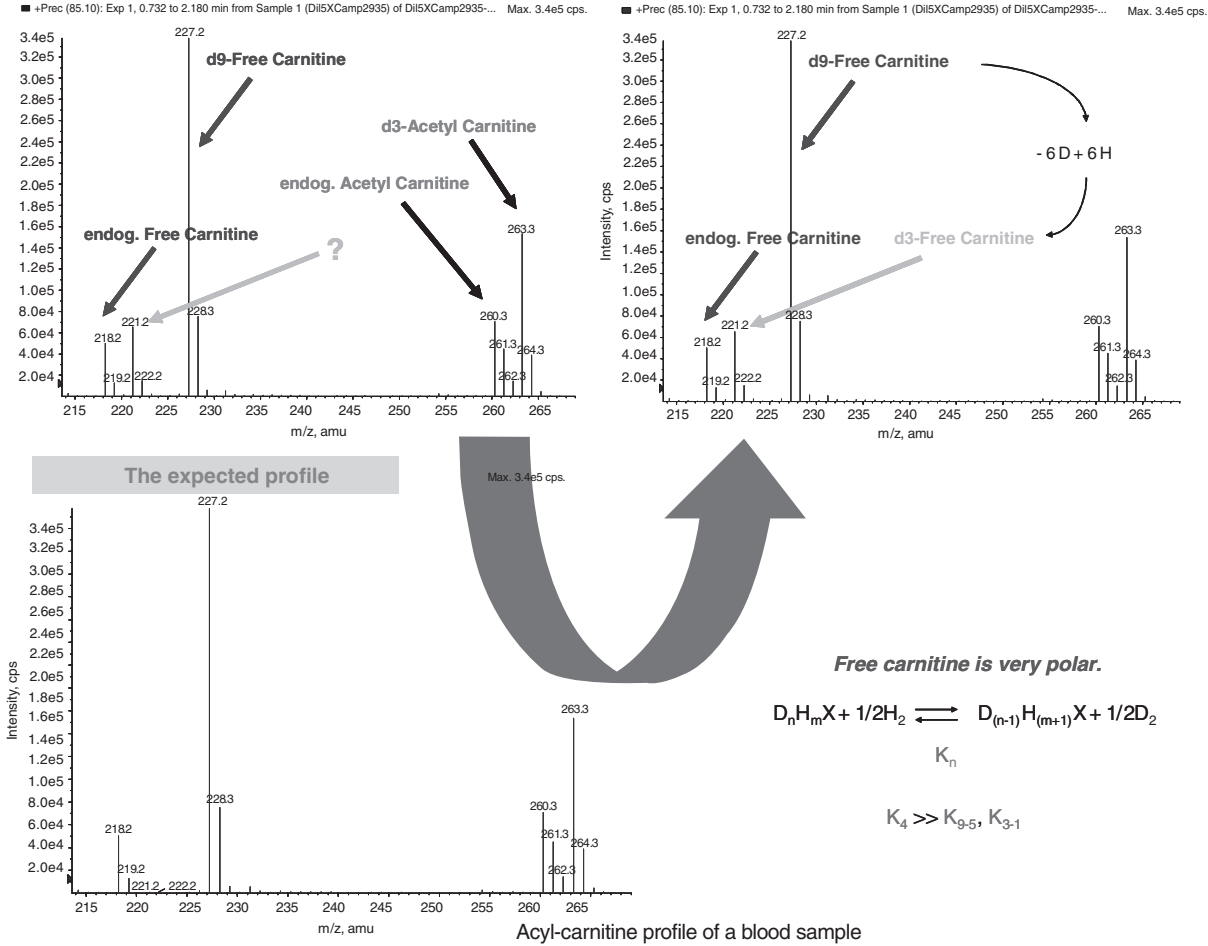


Fig. 19. Speculation on the appearance of the 221 Th ion in the AC profile in some NBS samples. Due to partial degradation of the isotopically labeled d9-acetyl carnitine (d9-C2), 221 Th ion is a new artifact, the d3-acetyl carnitine (d3-C2), produced by a padial D/H back-exchange.

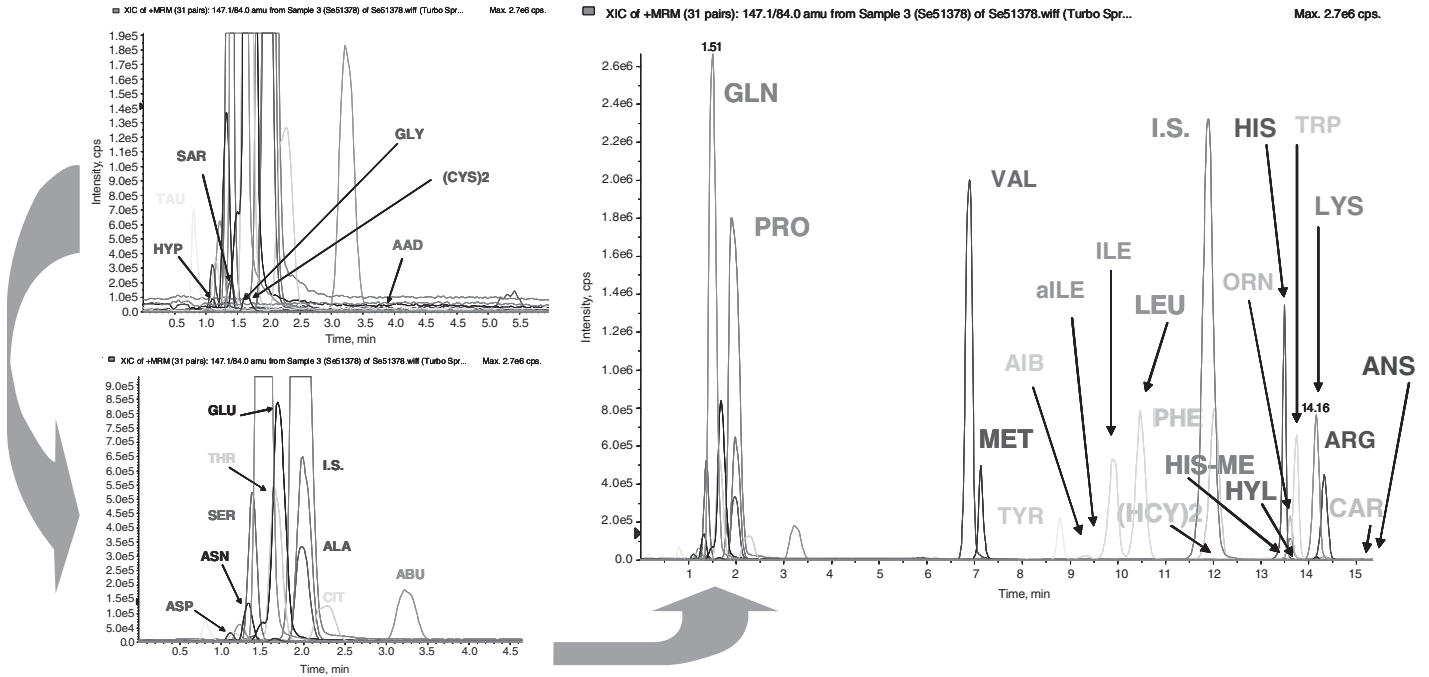


Fig. 20. MRM trace obtained on a plasma sample for AA without derivatization. The two insets are magnifying some details in the starting part of the chromatographic run.

So far, time scale is not compatible yet with the typical NBS throughput (2–3 min/sample), the analytical running time being roughly 1 order of magnitude longer. Quite promising is a very new strategy recently presented, where a special multiplex derivatization is performed enabling the concurrent reading of four different samples simultaneously injected. It leverages the tagging of the AA with four different isotopic labels, each of them generating a different signature fragment. Consequently, four different AA mixtures can be tagged separately with one of the four different labels and then pooled together before the injection in the LC–MS/MS. The time spent for the analytical run is shared by four samples. Therefore, the throughput is increased by a factor of 4.

5.2. Very long chain fatty acids

Peroxisomal disorders are characterized by impaired, reduced, or total absence of peroxisomes in cells. These disorders imply an accumulation of very long chain fatty acids (VLCFA) such as tetracosanoic and hexacosanoic acids in plasma and red blood cells. Some variants of these disorders are characterized by an accumulation of phytanic acid.

Up to now quantification of VLCFA has been done by GC or GC–MS. So far these methods are time-consuming and quite demanding in terms of sample preparation.

VLFAC measurements are accompanied by the calculation of some significant ratios like C26:0/C22:0 and C24:0/C22:0.

Some authors are pursuing the VLCFA characterization through their AC profile [17]. Only 5% of the VLCFA are incorporated within AC, and therefore the resulting detectability is questionable.

Recently Johnson [18] has proposed an interesting procedure employing LC–MS/MS for a rapid screening.

This approach targets all the VLCFA (free, in-phospholipid incorporated and ester forms). Due to the measurements made in flow-injection mode, the isobaric forms cannot be distinguished and the overall detection limits for some critical compounds like pristanic acid (at very low concentration when at normal levels) are not satisfactory.

Leveraging the sample preparation as proposed by Johnson, a methodology centered on a simple and robust LC–MS/MS hardware configuration, involving a chromatographic step and the use of a non-isotopically labeled internal standard, has been presented [19]. Fig. 21 shows that quantitation is viable by LC–MS/MS with an external calibration and without any special isotopically labeled compound.

5.3. Steroids

It is now well recognized that LC–MS/MS is becoming a pivotal tool for the steroid profiling in clinical research studies. Up to now steroids have been analyzed using immunoassay or GC–MS.

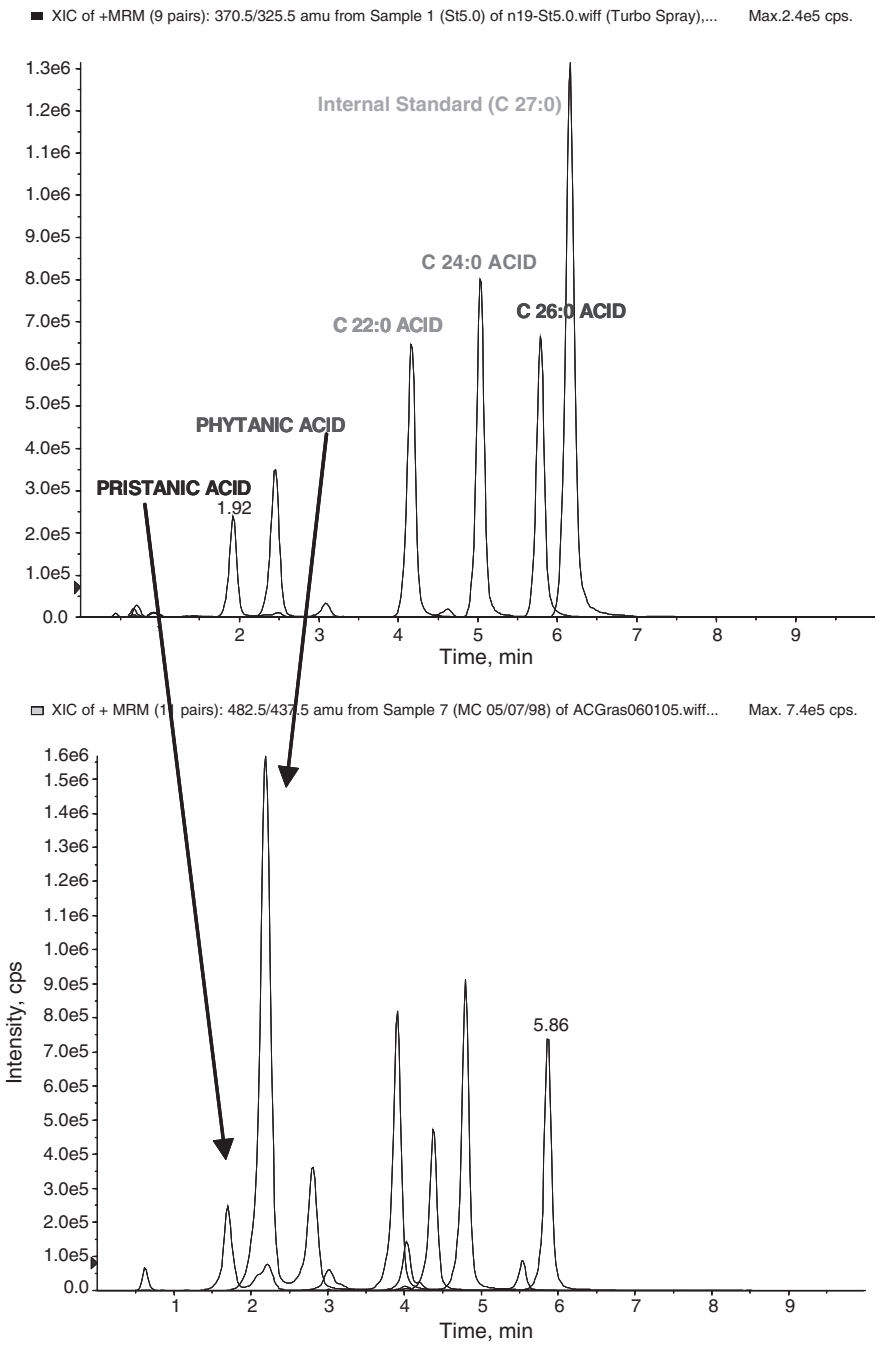


Fig. 21. Chromatographic traces obtained by injecting a standard solution of derivatized pristanic, phytanic, docosanoic, tetracosanoic, and hexacosanoic acids, and the derivatized internal standard (upper panel). Lower panel depicts abnormal concentrations of pristanic and phytanic acids in a real sample (18.9 and 72.6 μM, respectively).

Immunoassay is among the most sensitive analytical method. However, this technique shows a lack of specificity due to cross-reactivity and is particularly costly due to the expense of reagents.

With immunoassay each steroid must be analyzed separately and with very limited ranges, focusing on known levels of the drug of interest. Immunoassays also tend to have high variances at low concentration levels that can lead to error and misleading results.

GC–MS methods have also been used for steroid quantitation but these usually require extraction and purification steps along with derivatization of the steroid before measurement, which complicates the process and involves time.

LC–MS/MS is sensitive, specific, and allows an easier approach to sample preparation without sample derivatization steps. It can encompass the analysis of a virtually unlimited number of analytes in the same shot.

Up to now several papers leveraging LC–MS/MS have been published [20–25]. For attaining the necessary sensitivity, some sample pretreatment is required, basically centered either on a liquid–liquid extraction (LLE) or on an off-line solid-phase extraction (SPE).

Besides those options, Soldin's group [26] has proposed a strategy implementing an on-line single-step solid-phase-like extraction step coupled to a sensitive instrumental set-up (LC–MS/MS with atmospheric pressure photoionization source (APPI)). APPI (called "PhotoSpray" by some manufacturers) has been recently shown to be more sensitive to certain compounds, especially non-polar and aromatic species in biological matrices such as some steroids.

Any of the above strategies is valuable for routine use as long as the usual performance parameters (sensitivity and specificity) are associated to a good robustness and an easy sample preparation test.

Fig. 22 gives a flavor of what is attainable today with a tandem mass spectrometer for the routine quantitation of aldosterone, cortisone, cortisol, 21-deoxycortisol, corticosterone, substance S (11-deoxycortisol), δ-4-androstenedione, 21-hydroxy-progesterone, and 17-hydroxy-progesterone [27].

Fig. 23 shows the tracing obtained on a serum from a patient with 21-hydroxylase deficiency. A high concentration of 17-hydroxy-progesterone has been calculated at 28.9 ng/mL with the methodology. Value obtained using immunoassay was 25 ng/mL. Normal value should be less than 5 ng/mL. Peaks corresponding to substance S, cortisol, cortisone, corticosterone, and aldosterone are strongly decreased.

5.4. Bile acids

BA are a group of compounds characterized by the steroid scaffolding with a carboxyl group located in the side chain. These compounds are the major catabolic products of cholesterol and facilitate either the excretion of bile lipids including cholesterol or the absorption of dietary lipids.

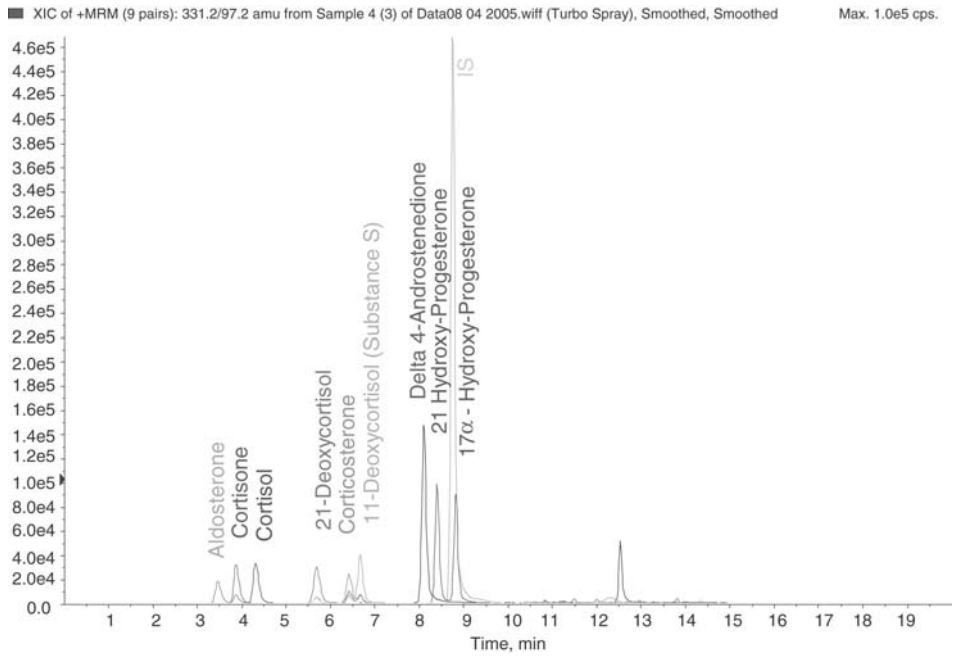


Fig. 22. Chromatographic trace obtained on an albumin serum solution spiked with a steroid mixture at 10 ng/mL.

The most prominent BA present in human are cholic acid (C), chenodeoxycholic acid (CDC), deoxycholic acid (DC), lithocholic acid (LC), and ursodeoxycholic acid (UDC), as derivatives of 5β-cholan-24-oic acid. Primarily they are present as glycine and taurine conjugates, with the conjugation occurring at carbon 24 of the structure. In addition to the above major BA, a wide array of minor components has been identified.

Hepato-biliary and intestinal diseases are marked by their increased concentration in plasma, urinary, and feces. Early diagnosis of many pathological conditions is often possible through individual separation and quantitation of BA.

The analysis of BA has been always challenging due to their wide variety, lack of any volatility, very low concentration in biological samples, and the small structural differences between them, with several cases of isomeric forms.

Outcome is that up to now some labor-consuming steps are required for a successful and comprehensive analysis of the BA range.

GC either alone or coupled with mass spectrometry (GC–MS) has been used for BA analysis in normal serum or urine since they provide high sensitivity and specificity. However, sample preparation represents the limiting factor: a preliminary separation of BA by class is needed, followed by hydrolysis and derivatization steps.

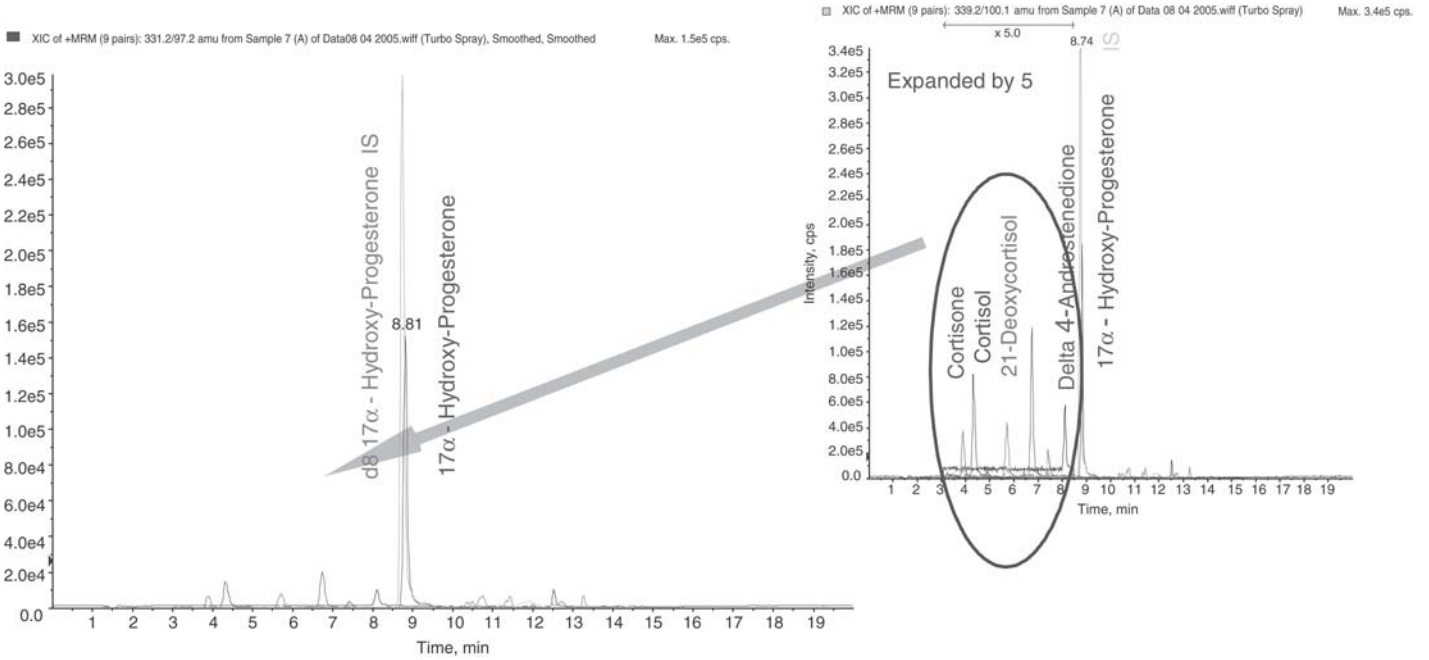


Fig. 23. Serum from a patient with 21-hydroxylase deficiency. A high concentration of 17-hydroxy-progesterone has been calculated at 28.9 ng/mL with the described methodology (left panel). Value obtained using immunoassay is 25 ng/mL. Normal value should be less than 5 ng/mL. Peaks corresponding to substance S, cortisol, cortisone, corticosterone, and aldosterone are strongly decreased (right panel).

Liquid chromatography coupled with UV detection has been widely applied in BA analysis. The main advantage over GC techniques is that the BA are determined without any derivatization step. The main drawback is the limited sensitivity, especially when applied to the dosage of BA in normal serum.

Due to the low volatility and the thermal instability of BA, and in order to address the persistent need for a rapid and sensitive means of BA screening in biological fluids, electrospray ionization–tandem mass spectrometry (ESI–MS/MS) techniques have been proposed for the analysis of these compounds. These methods allow direct analysis of the intact polar forms, either unconjugate or conjugate.

For differentiating BA isobaric forms, a prior chromatographic separation step is needed. The resulting LC–MS/MS demonstrates to be sensitive, robust, specific, easy, and sufficiently rapid, such as suitable for routine studies involving high number of specimens [28]. Fig. 24 shows what is achievable today by using LC–MS/MS in the BA analysis when chromatographic separation is implemented. By skipping the chromatography (just flow-injection analysis), analytical time is shortened to value compatible for a very-large-scale routine, as demonstrated by some researchers [29], but either resolution on isobaric BA (quite several ones) or good detectability for minor BA is lost.

6. Conclusions

The recent and continuing impact of LC–MSMS in NBS and other clinical chemistry applications is unquestionable. The chapter has focused the classical protocol (AA and AC) for detecting in a screening program the metabolic disorders in neonates. Several other advancements are now engaged. Besides the ones described above in the chapter and without the pretension to be exhaustive, it is worthy mentioning those related to purine and pyrimidine metabolism disorders [30], propionate metabolism disorder [31], some urea cycle disorders [32], galactose metabolism disorder [33], neuroendocrine disorders [34], folate and cobalamin deficiencies [35], and lysosomal storage disorders [36].

The excitement around this “quasi-universal” technology must be buffered by two important statements.

First, despite the efforts of the vendors in making the equipments simpler and easier to use, still LC–MS/MS cannot be viewed as a fully automated “black-box” like the other usual routine clinical instruments. Practical details described in the chapter for NBS prove that the entire analytical procedure chain is still far away from securing it in a blinded protocol.

Second important point is that classical NBS protocol has been a good example for having analytical speed married with a relevant number of provided information, the key point having been the FIA (no chromatography). Indeed, willing to extend the field to other clinical applications such as the ones mentioned in the

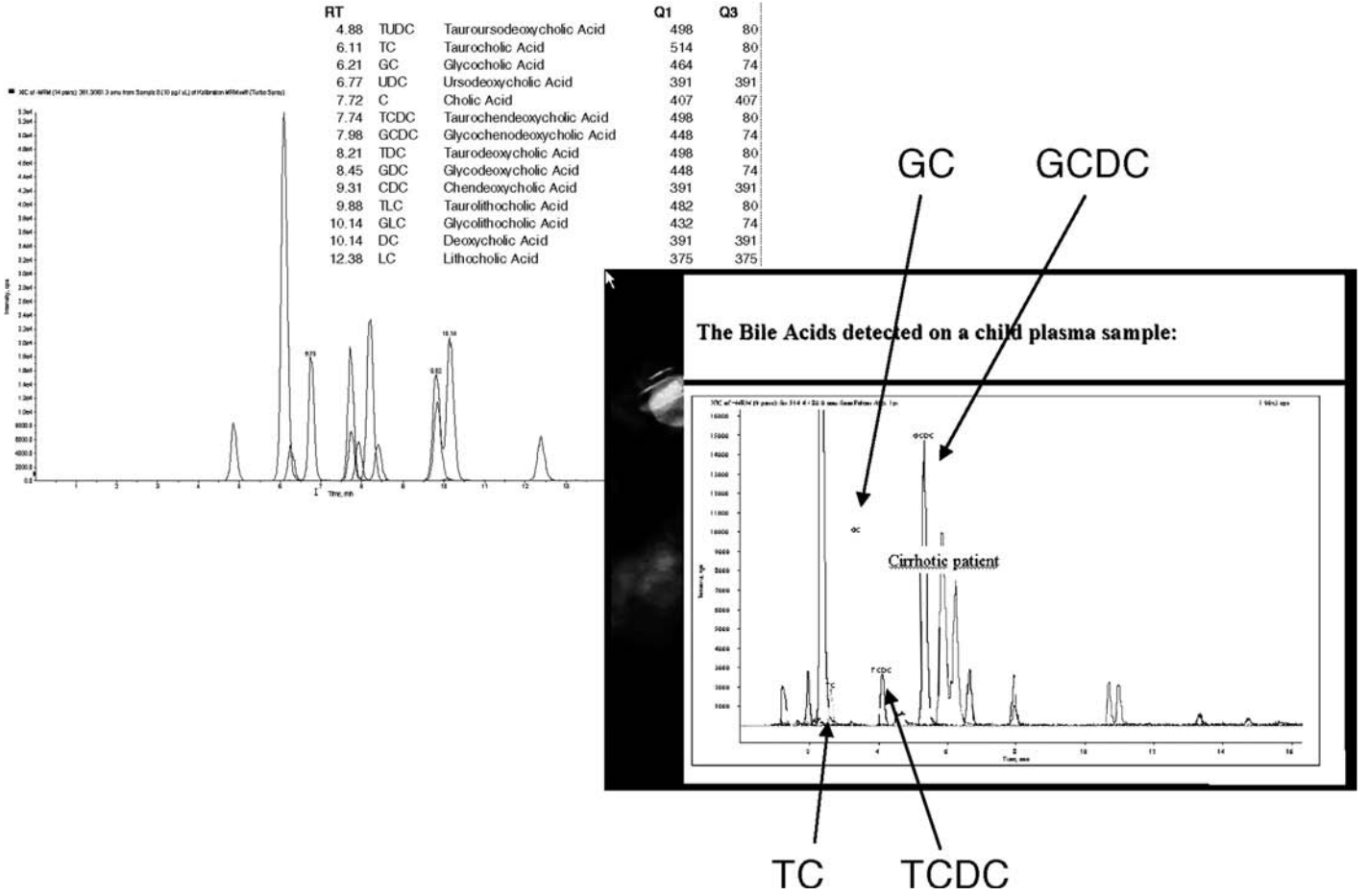


Fig. 24. Traces concerning the BA analysis by LC–MS/MS. (Left panel) A trace from the injection of a mixture at 10 ng/mL of the listed BA. (Right panel) A trace coming from a child plasma sample (qualified as cirrhotic). Strategical TC, TDC, GC, and GCDC bile acids are used for characterizing the type of disease.

whole chapter, the chromatographic separation becomes necessary for retrieving good and exhaustive information. And chromatography is synonymous of time.

In literature, some papers are presenting approaches skipping the chromatographic separation but the collected information is usually limited, just for a first-screening purpose.

Provided that anyway LC–MS/MS is an analytical technology deploying multi-parameters per reading, any newcomer must realize that the more and detailed information he is looking for, the more time must be paid for the analytical reading.

As for closing, bile acids could be a good example: their analysis is viable by FIA but the picture is very limited (no isobaric forms resolved low-abundance ones masked). Only with a chromatographic step (time demanding), all of them can be monitored (conjugated, unconjugated, isobaric forms, high- and low-abundance ones).

Acknowledgments

Part of the presented figures has been produced by exploiting in-house experiments. A small part has been edited by relying on the high-profiled experiments done with Dr. H.V. Van Thi and Prof. P. Bourdoux in the Newborn Screening Center, Laboratory of Pediatrics, Free University of Brussels (Belgium).

References

1. Chace, D.H., Kalas, T.A. and Naylor, W.N., Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns. *Clin. Chem.*, **49**, 1797–1817 (2003).
2. Bennett, M.J. and Rinaldo, P., The metabolic autopsy comes of age. *Clin. Chem.*, **47**, 1145–1146 (2001).
3. Guthrie, R. and Susi, A., A simple phenylalanine method for detecting phenylketonuria in large populations of newborn infants. *Pediatrics*, **32**, 338–343 (1963).
4. Chace, D.H., Millington, D.S., Terada, N., Kahler, S.G., Roe, C.R. and Hofman, L.F., Rapid diagnosis of phenylketonuria by quantitative analysis for phenylalanine and tyrosine in neonatal blood spots by tandem mass spectrometry. *Clin. Chem.*, **39**, 66–71 (1993).
5. Chace, D.H., Hilman, S.L., Millington, D.S., Kahler, S.G., Roe, C.R. and Naylor, E.W., Rapid diagnosis of maple syrup urine disease in blood spots from newborns by tandem mass spectrometry. *Clin. Chem.*, **41**, 62–68 (1995).
6. Chace, D.H., Hilman, S.L., Millington, D.S., Kahler, S.G., Adam, B.W. and Levy, H.L., Rapid diagnosis of homocystinuria and other hypermethioninemias from newborns' blood spots by tandem mass spectrometry. *Clin. Chem.*, **42**, 349–355 (1996).
7. Chace, D.H., Hilman, S.L., Van Hove, J.L.K. and Naylor, E.W., Rapid diagnosis of MCAD deficiency: quantitative analysis of octanoylcarnitine and other acylcarnitines in newborn blood spots by tandem mass spectrometry. *Clin. Chem.*, **43**, 2106–2113 (1997).
8. Rashed, M.S., Bucknall, M.P., Little, D., Awad, A., Jacob, M., Alamoudi, M., Alwattar, M. and Ozand, P.T., Screening blood spots for inborn errors of metabolism by electrospray tandem

mass spectrometry with a microplate batch process and a computer algorithm for automated flagging of abnormal profiles. *Clin. Chem.*, **43**, 1129–1141 (1997).
9. Roscher, A.A., Fingerhut, R., Liebl, B. and Olgemoeller, B., Erweiterung des neugeborenen-screenings durch tandemmassenspektrometrie. *Monatsschr. Kinderheilkd.*, **149**, 1297–1303 (2001).
10. Testa Cavedon, C., Bourdoux, P., Mertens, K., Van Thi, H.V., Herremans, N., De Laet, C. and Goyens, P., Age-related variations in acylcarnitine and free carnitine concentrations measured by tandem mass spectrometry. *Clin. Chem.*, **51**, 745–752 (2005).
11. Nagy, K., Takàts, Z., Pollreisz, F., Szabò, T. and Vékey, K., Direct tandem mass spectrometric analysis of amino acids in dried blood spots without chemical derivatization for neonatal screening. *Rapid Commun. Mass Spectrom.*, **17**, 983–990 (2003).
12. Pitt, J.P., Eggington, M. and Kahler, S.G., Comprehensive screening of urine samples for in-born errors of metabolism by electrospray tandem mass spectrometry. *Clin. Chem.*, **48**, 1970–1980 (2002).
13. Malvagia, S., la Marca, G., Casetta, B., Gasperini, S., Pasquini, E., Donati, M.A. and Zammarchi, E., Falsely elevated C 4-carnitine as expression of glutamate formiminotrans-ferase deficiency in tandem mass spectrometry newborn screening. *J. Mass Spectrom.*, **41**(2), 263–265 (2006).
14. Casetta, B., Tagliacozzi, D., Shushan, B. and Federici, G., Development of a method for rapid quantitation of amino acids by liquid chromatography-tandem mass spectrometry (LC-MSMS) in plasma. *Clin. Chem. Lab. Med.*, **38**, 391–401 (2000).
15. Piraud, M., Vianey-Saban, C., Petritis, K., Elfakir, C., Steghens, J.-P. and Bouchu, D., Ion-pairing reversed-phase liquid chromatography/electrospray ionization mass spectrometric analysis of 76 underivatized amino acids of biological interest: a new tool for the diagnosis of inherited disorders of amino acid metabolism. *Rapid Commun. Mass Spectrom.*, **19**, 1587–1602 (2005).
16. Casetta, B., Gadras, C., Belva-Besnet, H., Tessier, L., Simard, G. and Malthiery, Y., Research report: testing a rapid protocol for the quantitation by LC-MS/MS of 33 amino acids in biological fluids without derivatization. Proceedings of the 53rd ASMS Conference on Mass Spectrometry and Allied topics. San Antonio, TX, USA.
17. Rizzo, C., Boenzi, S., Wanders, R.J.A., Durain, M., Caruso, U. and Dionisi-Vici, C., Characteristic acylcarnitine profiles in inherited defects of peroxisome biogenesis: A novel tool for screening diagnosis using tandem mass spectrometry. *Pediatr. Res.*, **53**, 1013–1018 (2003).
18. Johnson, D.W., A rapid screening procedure for the diagnosis of peroxisomal disorders: Quantification of very long-chain fatty acids, as dimethylaminoethyl esters, in plasma and blood spots, by electrospray tandem mass spectrometry. *J. Inherit. Metab. Dis.*, **23**, 475–486 (2000).
19. Casetta, B., Gravier, M., Belva, H., Tavet, Y., Cuvelier, I. and Olichon, D., Quantitation by liquid chromatography-tandem mass spectrometry (LC-MSMS) of very long chain fatty acids (VLCFA) as a tool for the characterization of peroxisomal disorders. Poster D-130 at XIX International Congress of Clinical Chemistry/2005 AACC Annual Meeting.
20. Lacey, J.M., Minutti, C.Z., Magera, M.J., Tauscher, A.L., Casetta, B., McCann, M., Lymp, J., Houn Hahn, S., Rinaldo, P. and Matern, D., Improved specificity of newborn screening for congenital adrenal hyperplasia by second-tier steroid profiling using tandem mass spectrometry. *Clin. Chem.*, **50**, 621–625 (2004).
21. Taylor, R.L., Machacek, D. and Singh, R.J., Validation of a high-throughput liquid chromatography–tandem mass spectrometry method for urinary cortisol and cortisone. *Clin. Chem.*, **48**, 1511–1519 (2002).
22. Kushnir, M.M., Rockwood, A.L., Nelson, G.J., Terry, A.H. and Meikle, A.W., Liquid chromatography–tandem mass spectrometry analysis of urinary free cortisol. *Clin. Chem.*, **49**, 965 (2003).

23. Starcevic, B., DiStefano, E., Wang, C. and Catlin, D.H., *J. Chromatogr. B*, **792**, 197–204 (2003).
24. Reza Anari, M., Bakhtiar, R., Zhu, B., Huskey, S., Franklin, R.B. and Evans, D.C., Derivatization of ethinylestradiol with dansyl chloride to enhance electrospray ionization: application in trace analysis of ethinylestradiol in rhesus monkey plasma. *Anal. Chem.*, **74**, 4136–4144 (2002).
25. Nelson, R.E., Grebe, S.K., O'Kane, D.J. and Singh, R.J., Liquid chromatography–tandem mass spectrometry assay for simultaneous measurement of estradiol and estrone in human plasma. *Clin. Chem.*, **50**, 373–384 (2004).
26. Guo, T., Chan, M. and Soldin, S.J., Steroid profiles using liquid chromatography–tandem mass spectrometry with atmospheric pressure photoionization source. *Arch. Pathol. Lab. Med.*, **128**, 469–475 (2004).
27. Casetta, B., Javoy, V., Belva-Besnet, H., Malthiery, Y. and Gallois, Y., Routine Use of a Methodology for dosing 9 steroids in clinical research samples by LC-MSMS. Proceedings of the 53rd ASMS Conference on Mass Spectrometry and Allied topics. San Antonio, TX, USA.
28. Tagliacozzi, D., Mozzi, A.F., Casetta, B., Bertucci, P., Bernardini, S., Di Ilio, C., Urbani, A. and Federici, G., Quantitative analysis of bile acids in human plasma by liquid chromatography-electrospray tandem mass spectrometry: A simple and rapid one-step method. *Clin. Chem. Lab. Med.*, **41**, 1633–1641 (2003).
29. Johnson, D.W., ten Brink, H.J., Schuit, R.C. and Jakobs, C., Rapid and quantitative analysis of unconjugated C27 bile acids in plasma and blood samples by tandem mass spectrometry. *J. Lipid Res.*, **42**, 9–16 (2001).
30. Ito, T., van Kuilenburg, A.B.B., Bootsma, A.H., Haasnoot, A.J., van Crutchen, A., Wada, Y. and van Gennip, A.H., Rapid Screening of High-Risk Patients for Disorders of Purine and Pyrimidine Metabolism Using HPLC-Electrospray Tandem Mass Spectrometry of Liquid Urine or Urine-soaked Filter Paper Strips. *Clin. Chem.*, **46**, 445–452 (2000).
31. Magera, M.J., Helgeson, J.K., Matern, D. and Rinaldo, P., Methylmalonic acid measured in plasma and urine by stable-isotope dilution and electrospray tandem mass spectrometry. *Clin. Chem.*, **46**, 1804–1810 (2000).
32. la Marca, G., Casetta, B. and Zammarchi, E., Rapid determination of orotic acid in urine by a fast liquid chromatography/tandem mass spectrometric method. *Rapid Commun. Mass Spectrom.*, **17**, 788–793 (2003).
33. Jensen, U.G., Brandt, N.J., Chistensen, E., Skovoby, F., Norgaard-Pedersen, B. and Simonsen, H., Neonatal screening for galactosemia by quantitative analysis of hexose monophosphates using tandem mass spectrometry: A retrospective study. *Clin. Chem.*, **47**, 1364–1372 (2001).
34. Kushnir, M.M., Urry, F.M., Frank, E.L., Roberts, W.L. and Shushan, B., Analysis of Catecholamines in urine by positive-ion electrospray tandem mass spectrometry. *Clin. Chem.*, **48**, 323–331 (2002).
35. Gempel, K., Gerbitz, K.-D., Casetta, B. and Bauer, M.F., Rapid determination of total homocysteine in blood spots by liquid chromatography-electrospray ionization-tandem mass spectrometry. *Clin. Chem.*, **46**, 122–123 (2000).
36. Wang, D., Eadala, B., Sadilek, M., Chamoles, N.A., Turecek, F., Scott, R. and Gelb, M.H., Tandem mass spectrometric analysis of dried blood spots for screening of mucopolysaccharidosis I in newborns. *Clin. Chem.*, **51**, 898–900 (2005).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 17

Applications of mass spectrometry in oncology

JOHN ROBOZ*

Division of Hematology/Oncology, Department of Medicine, Mount Sinai School of Medicine, 1 Gustave Levy Place, New York, NY 10029, USA

*Tel.: 212-241-7382; Fax: 212-996-9801. E-mail: john.roboz@mssm.edu (J. Roboz).

1. Introduction

Cancer results from the multistep accumulation of somatic, and occasionally inherited, mutations that lead to clonal neoplastic cell transformation [1]. The associated genetic lesions include the activation of dominant oncogenes and the inactivation of tumor suppressor genes through mutation and loss of heterozygosity. Some 191 "cancer genes" have been reported, 90% of them exhibiting somatic mutations [2].

The current buzzword is "biomarkers," defined "as endogenous or injected molecules whose presence or metabolism correlates with important disease related physiological processes and/or disease outcomes" [3–6]. We know that tumors are always in a process of interaction with their immediate environment, resulting in the release, acquisition, or exchange of proteins, the nature and quantity of which is likely to change in the course of the growth of the tumor. The proteome can contain thousands of proteins, depending on cell or tissue type, health or disease state, and other factors. There are several distinct aspects of proteomic studies [7]. To decipher a proteome, the first objectives are the large-scale identification of proteins and their posttranslational modifications within a cell, tissue, or other biological sample, followed by structural characterization and the elucidation of the specific functions and interactions of targeted proteins, within and between cells. Although proteomics is undoubtedly the major area of ongoing cancer biomarker research, the glycoprotein-related aspects of glycomics [8,9] and the cancer-related sphingolipid (ceramid) aspects of lipidomics [10,11] are also areas of rapidly evolving importance in cancer research.

During the last decade, mass spectrometric techniques have been used successfully in all aspects of cancer medicine and research. These include environmental carcinogenesis, cancer biochemistry and molecular biology, immunology, all stages of chemotherapy from identification of natural products through all phases of the arduous drug development process (synthesis, cell culture experimentation,

Table 1
Scope of mass spectrometry in cancer medicine illustrating the range of applications from small molecules to biopolymers

Screening and diagnosis	• Opportunistic infections: detecting/ quantifying circulating microbial metabolites for diagnosis and monitoring antibacterial and antifungal chemotherapy
• Unique carcinogenic biomarkers and biomarker profiles	Biology
• Monitoring subjects at risk (dosimetry)	• Elucidation of cellular or structural changes leading to oncogenesis
• Detection of small genetic alterations in the background of normal genes	• Nature of relevant mutations and time of their occurrence
• Differentially expressed proteins in body fluids and tissues: upregulated, downregulated, or unique	• Identification and quantification of epitopes
Medical oncology and treatment monitoring	• Function-critical posttranslational modifications
• Toxic concentrations of drugs and/or metabolites in blood, urine, and tissues	• Changes in cellular proteins in apoptosis
• Pharmacokinetics; concentration × time curves	• Changes in cellular proteins in progression of tumors
• Pharmacodynamics; metabolism, protein binding, loading values	• Identification of critical protein–protein associations
• Pharmacogenetics	Chemoprevention
• Protein expression in the development of chemoresistance	

treatment of experimental animals) and, increasingly, medical oncology, including diagnosis [12] and Phase I–IV clinical trials, and treatment monitoring (reviewed in ref. [13]). Table 1 summarizes the scope of mass spectrometry in cancer medicine.

After a description of the surface enhanced laser desorption ionization (SELDI)–TOF technology, and a brief discussion of a few methodological challenges, there is a review of the diagnostic oncoproteomics in several malignancies, summarizing results and discussing advantages and shortcomings. Next, there is a review of representative applications (in no order of importance) in a variety of areas, aiming to illustrate the wide diversity of subjects of current interest in cancer research where mass spectrometry has been used successfully.

2. New methodology—SELDI–TOF–MS

2.1. Protein chips

In contrast to the MALDI technique, where the surface of the probe does not have an active role in the analytical process beyond holding the sample, in SELDI the probe surface plays an active role in a number of aspects of the processing of the analytes, e.g., extraction, structural modification, and amplification [14]. In the

ProteinChip™ technology [15] special array surfaces are used to selectively retain entire subsets of proteins directly, and in a single step from biological samples. Thus, in contrast to HPLC–MS, which combines elution chromatography with MS, SELDI–MS combines retention chromatography with MS.

The selectivity for the target protein(s) is based on biochemical characteristics including surfaces based on normal phase silica, strong and weak anion exchange, immobilized metal affinity capture (IMAC), various reactive moieties [16], and affinity technology [17]. Antibody-based chips have also been designed to "bait out" individual proteins from crude biological samples [18]. After a series of wash protocols, the captured proteins are mixed with appropriate matrices and released by MALDI for subsequent mass determinations by TOF–MS or MS/MS studies in Qq-TOF analyzers. Chips are now available where the energy absorbing molecules are already incorporated into the surface chemistry of the array. Practical aspects of the SELDI–TOF–MS have been reviewed [19,20].

2.2. Identification

Fig. 1 shows a typical SELDI mass spectrum, showing a number of peaks in both the normal and the pathologic samples and also a potential marker differentially expressed in the patient sample. Here it would be worthwhile to attempt identification, assuming that the unique expression is statistically significant. Often, however,

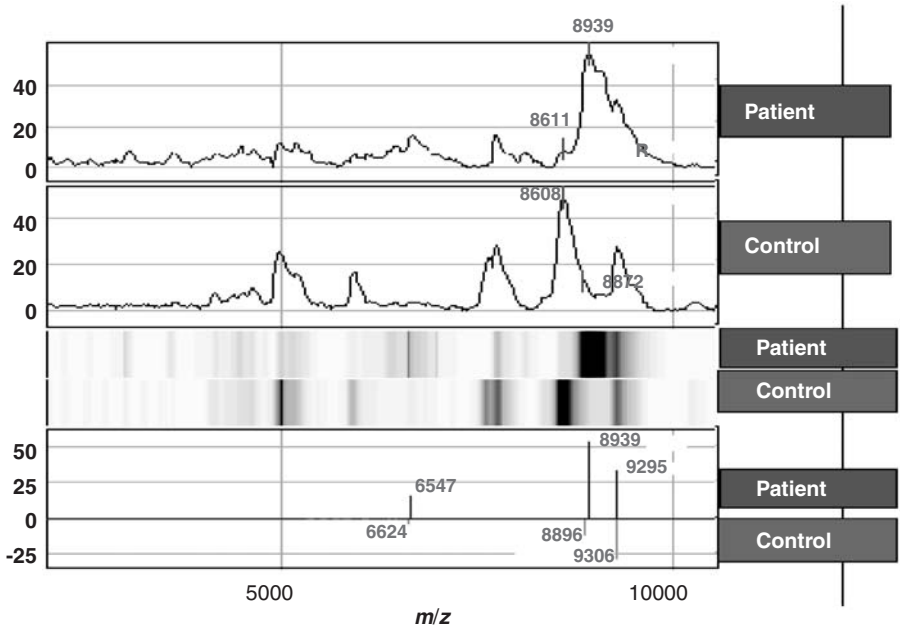


Fig. 1. Typical SELDI spectrum comparing serum samples from a patient and control. Upper two lines: conventional spectra. Middle: gel-form presentation. Lower lines: potential markers differentially expressed in the patient sample.

distinct markers are not found and one needs to use a proteomic pattern recognition algorithm to detect potentially important peaks within the protein profiles.

Putative biomarkers detected by SELDI may be characterized/identified by the judicious selection of any of several available strategies of protein identification, including conventional 2-DE separation followed by enzymatic digestion and the "bottom-up" approach, the sophisticated "top-down" strategy, or highly accurate mass determination for direct identification. There are SELDI accessories available for high-resolution tandem mass spectrometers. It is possible to identify proteins directly on the surface of protein chips with virtually no sample loss [21,22]. When the database searching is negative, *de novo* sequencing is indicated [23].

Whatever analytical approach is used for the characterization of proteins, the common last step is to use a computer algorithm for the evaluation of the data obtained [24]. The general role of bioinformatics in protein analysis, including database searches, sequence comparisons, and structural predictions, has been reviewed [25,26]. A concise review of the available software tools for database searching to interpret mass spectrometric data lists relevant original references [25].

2.3. Proteomic pattern diagnostics

The importance of the need for full identification of potential protein markers is controversial. Most papers on SELDI–TOF–MS for cancer diagnosis have been omitting the identification of potential markers beyond the determination of their approximate molecular mass. In the opinion of some investigators, one does not need individual identified markers as long as consistent protein profiles can be obtained [20]. In other words, the diagnosis of a disease should be considered as a prediction and should not be concerned about etiology.

When there are no individual peaks or groups of peaks with intensities significantly different between normal vs. pathologic samples, a bioinformatics algorithm is needed for diagnosis. There still are major unresolved challenges to the interpretation of SELDI data [27]. Most proposed algorithms use a supervised approach which is based on training datasets. Available programs are based on generic algorithms [28], classification and regression tree analysis [29–31], unified maximum separability algorithm [32,33], artificial neural networks providing association with disease grade [34], and an algorithm/*k*-nearest neighbors method, which was successfully applied to the published original ovarian cancer dataset [35]. A comprehensive pattern recognition procedure was designed to detect cancer-specific markers amid massive sets of mass spectral data; when applied to a published set of data on ovarian cancer, 100% specificity and 100% sensitivity were achieved, including early stage disease [36]. A novel statistical method, called *link test*, is based on the association between a specific mass spectrum marker and a microassay marker; this cross-platform approach was applied for finding prostate cancer (PC) biomarkers [37].

2.4. Problems and prospects

The initial media "hype" following the publication of the "original" article on applying SELDI–TOF–MS to ovarian cancer (see Section 4.1) was followed by severe criticism of several aspects of the methodology and interpretation techniques as well as applications to the diagnosis of ovarian and PCs. A number of remedies were suggested, including proposals to use more sophisticated mass spectrometers (e.g., Qq-TOF), change sample handling and experimental procedures, and apply advanced mass measurement, reproducibility, and validation approaches [38–42]. There is significant ongoing progress to eliminate the initial shortcomings, ranging from advances in chip technology to improve batch-to-batch reproducibility, through the use of proper mass calibration and internal standard techniques, to the design and testing of relevant algorithms to improve data handling and interpretation. There are efforts to adapt the entire process to robotic systems [43]. The SELDI technique establishes protein profiles easily and rapidly in body fluids. Based on current progress, it is reasonable to believe that SELDI–MS will prosper [44–46].

3. Other relevant methodological challenges

3.1. Analysis of cells

A novel technique, laser capture microdissection (LCM), has been developed to obtain very small populations of well-defined normal epithelial and adjacent tumor cells for subsequent lysing. In contrast to the ~50,000 LCM-procured cells needed for 2D-PAGE analysis, the SELDI process requires only 25–100 cells to obtain a usable protein profile [47,48]. The technique is also applicable to subsequent imaging analysis [49].

3.2. Direct tissue analysis and imaging MS

The principles and instrumentation of this exciting new technology are described in Chapter 23. This approach clearly has major potential in cancer research; indeed, several initial proofs of principle applications in tumor characterization, biomarker diagnosis for diagnosis, and even drug development have been described [50–52].

3.3. The problem of dynamic ranges

A well-known problem in the analysis of protein mixtures is the fact that the presence of components with high abundance may prevent even the detection (let alone quantification) of low-abundance proteins (which could ordinarily be

analyzed easily). For example, the concentration of plasma albumin is $>10^6$-fold higher than that of the tumor-derived cytokeratin. Ion suppression ("matrix effect") may occur when the quantitative response to an analyte is significantly reduced (possibly even eliminated) by the presence of a large quantity of another analyte or components of a buffer. This type of problems may often be handled effectively by using nano-flow rate separations in ESI–MS [53]. A highly efficient approach to improve dynamic range (by >10-fold, to zeptomole detection limits) in capillary separation–ESI–FTICR–MS involves the DREAMS technique where the acquisition of a normal mass spectrum is followed by another acquisition in which the most abundant ions detected in the first scan are not introduced into the FTCR trap because they are removed in a quadrupole accessory placed outside the magnetic field [53].

3.4. Low-abundance, low-molecular mass proteins or drugs in plasma/serum

Albumin, a major constituent in serum (60–80 mg/L), is known to act as a transport carrier for small proteins. Also, many antineoplastic drugs bind to albumin, often at the 80–95% level. Removal or depletion of albumin is a major problem: there are problems with ultrafiltration (e.g., membrane binding of small proteins and drugs) and other approaches, e.g., Cibacron dye columns and immunoaffinity-based protein subtraction chromatography. Albumin removal using acetonitrile may be a simple alternative [54]; however, there is no truly satisfying method at this time.

3.5. Quantification

Whatever the objective of a proteomic analysis (e.g., discovering diagnostic/prognostic markers, detecting new therapeutic targets), the confirmation of the presence (or absence) of a particular protein is not adequate. To carry out their functions within cells, proteins are continually synthesized or degraded; thus, knowledge of the quantity (relative or absolute) of the protein analyte is essential in most cases. It is ironic that, while mass spectrometry is often an excellent tool for quantification, the technology to quantify individual proteins in mixtures has been notoriously inadequate in both ESI and MALDI ionizations and in both MS/MS and SIM or SRM techniques. Strategies for the quantification of proteomes and subproteomes (based on posttranslational modifications) have been reviewed [55]. A novel strategy is based on "decomposing" spectra into peaks and baseline using so-called statistical finite mixture models [56].

Several techniques have been developed for quantification using stable isotope dilution [57]. There are special considerations for the global addition of stable isotope labels before or after protein digestion, and the metabolic labeling of proteins *in vivo*, e.g., growing cells or even whole animals in which all proteins have been

labeled biosynthetically [58]. In an alternative, label-free, approach, a fully automated technology has been developed for LC–MS/MS analysis of complex protein mixtures, based on the quantification (over a 32-fold range) of peptides directly after integrating ion current associated with each peptide peak [59].

Quantification using the intensities of protein peaks obtained by SELDI–TOF–MS is not yet a reliable approach; results should be used together with information from microassays which are more reliable [37].

4. Diagnostic oncoproteomics based on SELDI–TOF–MS

Oncoproteomics is the systematic application of proteomic technologies to oncology research [60]. Diagnostic proteomics concentrates on "differential display" comparisons of protein (peptide) concentrations in plasma or urine in health and disease with the following objectives: (i) early, rapid, and reliable diagnosis of cancer for timely therapeutic intervention based on identified specific markers; (ii) early diagnosis of relapse; (iii) early diagnosis for risk assessment to aid prevention.

Although in differential display proteomics one would hope to find unique protein tumor markers, perhaps resulting from posttranslational modifications in the neoplastic cells, it is more practical to search for proteins that are significantly up- or downregulated in the tumors. However, focus is shifting from methods that can analyze one marker at a time to pattern-matching approaches which allow the simultaneous measurement of a range of putative disease markers without the identification of specific tumor-associated proteins. A hybrid strategy has been suggested to retain the desirable attributes of high-information content MS patterns without giving up the capability to obtain identity [61].

There are only a few protein tumor markers used in clinical practice for diagnosis or prognosis (http://cis.nci.nih.gov/fact/5_18.htm). In 2002, the publication of a new, novel technology, SELDI–TOF–MS, for the early (Stage I) diagnosis of ovarian cancer was received with considerable enthusiasm by physicians, scientists, as well as the international news media (see below). Testing the diagnostic potential of SELDI–TOF–MS for cancer diagnosis has been burgeoning during the last few years [62]. There are a number of strategies to consider in clinical proteomics, from the definition of the clinical question, through data acquisition, pre- and postprocessing, to protein identification and method validation [63]. The importance of biomarkers may be appreciated even more by considering the increasing role of the Food and Drug Administration to predicate the "safe" and "effective" use of newly developed analytical approaches and the marketing of such for clinical applications [64]. It is noted that a number of tumor-related proteins have already been identified by various mass spectrometry techniques [13]. Here only representative SELDI-related diagnostic applications are reviewed.

4.1. Ovarian and endometrial cancer

The methodology in the "original" work on ovarian cancer aimed to recognize signature protein patterns. Thousands of peaks were analyzed by an iterative searching artificial intelligence algorithm. Using the results of a training set (50 each of pathologic and control sera), mass spectra were evaluated from 50 women with confirmed ovarian cancer and 66 controls. Diagnostic sensitivity was 100% and specificity was 95% [28]. The *Lancet* paper has been extensively criticized [65,66]. In a Point–Counterpoint exchange, a number of shortcomings were detailed, with respect to design, experimental techniques, and interpretation techniques; suggestions were made as to how to test and remedy the problems and validate the methods [67,68] (see also Section 2.4). Application of a new methodology, based on combinatorics and optimization-based logical analysis, to the original ovarian dataset provided several advantages leading to both sensitivity and specificity approaching 100% [69].

In a five-center, case–control study of hundreds of patients, three putative markers were identified by MS/MS using high-resolution SELDI–MS: apolipoprotein (downregulated), a truncated form of transthyretin (upregulated), and a fragment of inter-α-trypsin inhibitor heavy chain H4 (downregulated). With the combination of these biomarkers with CA125, both specificity and sensitivity were significantly improved with respect to the CA125 antigen, the "gold" standard [70]. In a follow-up study, an attempt was made to use these posttranslationally modified proteins for the classification of cancer types [71]. Another study compared low- and high-resolution platforms (Qq-TOF), both equipped with SELDI sources. As expected, the high-resolution platform yielded superior classification patterns [72].

A comparative study of malignant and normal endometrial tissues yielded a panel of proteins displaying differential expression in malignant tissues. A prominent putative marker was identified as chaperonin 10 by both MALDI–Qq-TOF and ESI–Qq-TOF–MS, confirmed by Western blot and immunohistochemistry [73]. A comparison of sera of patients with endometrial cancer with those of healthy females using SELDI–TOF–MS (weak cation exchange chips) yielded a number of putative biomarkers upon evaluations with three data mining tools (a tree classifier, Biomarker Wizard, and Biomarker Patterns System). The diagnostic pattern combined with 13 putative markers made it possible to differentiate patients with endometrial cancer from healthy subjects with specificity of 100% and sensitivity of 92.5% [74].

4.2. Breast cancer

The principles and potential clinical applications of SELDI–TOF–MS and microarray techniques have been reviewed with respect to screening, diagnosis, prediction of aggressiveness, response to treatment, and toxicity [75].

4.2.1. Serum

A 28.3 kDa protein, detected in 100% of invasive breast cancer samples, 80% with noninvasive disease ($n = 46$), and 4% of disease-free women ($n = 23$), was identified as belonging in the kallikrein protein family [76]. In a retrospective study (103 patients divided by staging, 25 patients with benign disease, 41 healthy women), using a panel of three unidentified biomarkers (4.3, 8.1, and 8.9 kDa) and bootstrap cross-validation, sensitivity was 93% for patients, and specificity was 91% for controls [32]. Comparable diagnostic results were found by other investigators using a different panel of diagnostic proteins and more advanced algorithms [77].

4.2.2. Nipple aspirate fluids (NAF)

Advantages of NAF include the noninvasive nature of sampling, an ability to sample both the diseased and the healthy contralateral breasts, the fact that NAF may be reflective of the microenvironment where the carcinoma originates, and that NAF is usually more concentrated than serum (ductal lavage may provide an even better account of the tumor as it represents the entire length of the duct). Various experimental aspects (including nipple aspiration, ductal lavage, endoscopy, cytopathology, as well as characterization of putative markers by SELDI–TOF–MS) of the intraductal approach to biomarker discovery have been reviewed [78].

In a prospective trial of 114 women, scheduled for diagnostic breast surgery, three putative markers (5200, 11,800, and 13,880 Da) were expressed differentially. Two other putative markers (5200 and 33,400 Da) differentiated between benign disease, ductal carcinoma *in situ,* and malignant tumor. Best results were obtained by combining clinical and proteomic data [79]. In another study, paired NAF samples from cancerous and noncancerous breasts were compared ($n = 23$) and 463 peaks were analyzed. Results included the recognition of two overexpressed and one underexpressed putative protein markers in tumor bearing breasts compared to disease-free subjects. Phenotypic proteomic NAF profiles differentiated between patients with early stage cancer and healthy women [80].

In a study aiming to establish quality control for NAF analysis, rigidly controlled experimental conditions were repeated 36 times. Algorithms were developed for the quantification of >700 analyte peaks ($\sim$18,000 time points) at low masses [81].

4.2.3. Breast tissues

A comparative study of tissues (obtained using LCM) from primary breast cancer with and without axillary lymph node metastasis was carried out with SELDI–TOF–MS and analyzed using ANOVA and multivariate logistic regression.

Two metal-binding polypeptides (4871 and 8596 Da) were identified as significant risk factors [82].

4.3. Prostate cancer

Testing for elevated levels of prostate-specific antigen (PSA) together with manual digital rectal examination is the accepted test for the early detection of PC. However, as benign prostatic hyperplasia (BPH) also causes elevated PSA, biopsy is still needed to confirm PC. SELDI–TOF–MS has been used to analyze free and complexed PSA, and prostate-specific membrane antigen (PSMA), in pure forms, cell lysates, sera, and seminal plasma samples. Analysis of cell lysates (~2000 cells obtained by laser capture microdissection) revealed free PSA in both normal and cancerous tissues, but not in stroma. PSMA was present only in cancer cells [83]. The general applicability of biomarkers for early diagnosis, and potential problems of SELDI–TOF–MS have been reviewed [84,85].

In a study to test a pattern-matching algorithm, serum protein profiles were obtained from 167 PC and 77 BPH patients, and 82 healthy men. Of some 63,000 peaks detected, 9 peaks (masses in the 4.4–9.5 kDa range) with high discriminatory power were selected to develop and train a decision tree classification algorithm. On testing by stratified, randomly selected samples, sensitivity was 83% and specificity was 97%. The predictive value was 94% for the study population and 91% for the general population [86]. In a similar study, using the boosted decision tree analysis approach, one of the two classifiers developed achieved 100% sensitivity and specificity but required 74 peaks and 500 base classifiers. A different evaluation yielded only 97% sensitivity and specificity for the test set; however, it required only 21 peaks and a combination of only 21 base classifiers [30]. In a third study, a classifier algorithm was established using seven masses (2–18.2 kDa range). PC was correctly predicted in 36/38 patients, while 177/228 subjects were correctly classified as BPH. The specificity for marginally elevated PSA (4–10 ng/mL, $n = 137$) was 71% [87]. With respect to methodologies, correlation and prediction confidence of the decision forest technique [88] and platform reproducibility were evaluated [89].

A different approach to diagnosis involved searching for individual markers. In one study, three potential markers were detected in PC but none in 12 controls. The 15.9 kDa marker appeared in 9/11 PC but was absent in 12 patients with BPH. The 15.2 kDa marker appeared in 9/11 PC and 4/12 BPH patients. The intensities of the 17.5 kDa marker were essentially the same in the PC and BPH groups. Because the 15.9 kDa protein was present in 82% of PC but was absent in all BPH, it was concluded that this protein may be a putative marker to differentiate PC from BPH [90].

In another study, a 50.8 kDa protein was detected in 96% of cancer patients ($n = 56$) but not in 70–80% of subjects with various benign prostate diseases ($n = 22$)

and 96% of controls ($n = 48$). Using the mass fingerprinting method, the protein was identified as being related to vitamin D-binding protein [91].

4.4. Pancreatic cancer

After using CA 19-9 as the only accepted diagnostic marker for pancreatic cancer for two decades [92], a bewildering number of potential biomarkers are currently under evaluation [93]. A peak (3334.7 Da) found by SELDI–TOF–MS in 5/15 pancreatic adenocarcinoma cell lines was identified by Qq-TOF–MS/MS as the COOH-terminal fragment of DMBT1, a putative tumor suppression protein intracellularly generated by limited prior proteolysis. Analyses of other cell lines suggested that the marker may be unique to pancreatic adenocarcinoma [94]. In another study, a differentially expressed peak (~16,570 Da) was detected in 10/15 samples from pancreatic adenocarcinoma in contrast to only 1/7 with other pancreatic diseases. The peak was identified as hepatocarcinoma–intestine–pancreas/pancreatitis-associated protein I (HIP–PAP I) by SELDI immunoassay [95].

Subsequently, protein profiles were obtained, after fractionating sera into six fractions, using various chips to remove albumin. A diagnostic algorithm revealed significant differences between patients with resectable adenocarcinoma, nonmalignant pancreatic diseases, and healthy subjects. The most discriminating peaks (all downregulated) were 3146 and 12,861 Da (fraction 1) and 3473, 5903, 8563, and 16,008 Da (fraction 6). Combinations of the markers performed better than the CA19-9 marker. Interestingly, the HIP/PAP protein of the previous study was not detected [33].

Another study of 245 plasma samples led to the selection of training cohort ($n = 71$ for both cancer patients and controls) for a vector machine learning algorithm. Four putative markers were recognized (in the 8.7–14.8 kDa mass range) yielding sensitivity of 97% and specificity of 94% in the training cohort. When applied to the entire validation cohort (in two institutions), both sensitivity and specificity were 91%. When combined with CA 19-9 results, 100% of tumors were diagnosed ($n = 29$) including early stages (Stages I and II) [96].

4.5. Bladder cancer

Aiming for a diagnostic urine test for transitional cell carcinoma (TCC, 95% of total cases), 94 samples and controls were analyzed. Of some 70 differentially expressed proteins and polypeptides in the 2–150 kDa mass range, 5 were preferentially expressed in TCC, at 3353 Da, 9495, 44.6, 100.120, and 133.190 kDa. The 3.3 kDa protein, also detected in microdissected bladder cells, was identified, by SELDI immunoassay and database search, as a member of the human defensin family. The diagnostic sensitivity of the combined markers was 78% compared to 33% of cytologic approaches [97].

Attempts to differentiate TCC from benign urogenital diseases led to a training sample set for a decision tree classification algorithm which, in turn, yielded a mass cluster pattern. In a blinded test set ($n = 38$) sensitivity was 96.3% and specificity was 87.0% [98]. In another study, a training set utilizing 5/187 mass peaks (from 104 urine samples) was used to establish a pattern for tree analysis. The pattern correctly predicted 49/68 test samples, 25/45 TCC samples, and 24/33 noncancerous samples [99].

An investigation of several methodological aspects of obtaining urinary protein profiles by SELDI–TOF–MS revealed that among the extrinsic factors instrument settings and matrix composition critically influenced peak detection and reproducibility, while freeze–thaw cycles had minimal effects. Intrinsic factors of significance included blood in urine, dilution, and first-void vs. midstream urine [100].

4.6. Head and neck cancer

A comparison of protein profiles from cell lines, derived from a primary tumor and a metastatic lymph node, revealed four differentially expressed proteins in the latter: two membrane-associated proteins (downregulated), annexin I and annexin II, glycolytic protein enolase-α (upregulated), and a calumenin precursor (downregulated). The identification of the upregulated proteins was validated using digestion with endoproteinase lysine-C [101].

Aiming to use serum to screen for differentially expressed proteins, head and neck squamous cell cancer samples ($n = 99$), "healthy" smokers ($n = 25$), and healthy controls ($n = 102$) were analyzed. An expected known biomarker, metallopanstimulin-1 (10,068 Da) was identified by SELDI immunoassay (rabbit polyclonal antibodies). A classification tree algorithm was used for the evaluation of numerous other protein peaks. The training set consisted of 75 samples from each group, and the rest of the samples constituted the test set. Discrimination of squamous cell cancer from controls and healthy smokers was accomplished with a sensitivity of 83.3% and specificity of 90% [102].

4.7. Miscellaneous malignancies

4.7.1. Colon and gastric cancer

The differential expression of a putative serum marker (8.9 kDa) was threefold higher in colon cancer patients ($n = 34$) than in controls ($n = 14$). The analyte was separated using C18-type Zip-Tip columns and, after a number of separation and purification steps, was digested in-gel with trypsin. The resulting peptides were analyzed using IT-MS/MS. The MS/MS data were compared with a sequence database and the protein identified as C3-desArg^{77}, a hydrolyzed component of human anaphylatoxin, complement component 3 precursor C3 [103].

In a subsequent study of comparable design, a classification algorithm identified a set of putative markers in the 4–8 kDa mass range. The model with the highest classification accuracy included two masses, at *m*/*z* 8132 and 4002. For an independent set of sera, the pattern could differentiate patients (different stages of colorectal cancer) from healthy subjects with both sensitivity and specificity of 95% [104].

Protein profiles were obtained using SELDI–TOF from patients with gastric cancer ($n = 127$), healthy controls ($n = 100$), and a small number of patients with other malignancies. Three masses selected as "fingerprints" (*m*/*z* 1468, 3935, and 7560) enabled a classifier algorithm to differentiate, in the training set, between cancer patients and controls with a sensitivity of 96% and specificity of 92%. In a blinded test set, sensitivity was 85% and specificity was 88%. The performance of this approach yielded better results than those based on the combined conventional carcinoembryonic antigen (CEA) and carbohydrate antigen (CA 19-9) tests [105].

4.7.2. Lung cancer

Cells ($\sim 3 \times 10^4$) were obtained from frozen sections of normal lung, atypical adenomatous hyperplasia, and malignant tumors using LCM. Six potential markers were present in tumor cells with significantly higher intensity and three peaks with significantly lower intensities, compared to normal cells; one peak (17,250 Da) was not detected in normals. A "malignant lung protein profile" made it possible to differentiate between tumor and premalignant pulmonary epithelium [106].

4.7.3. Melanoma

The concentrations of some putative markers in the 2.5–3.5 kDa range exhibited significant variations related to the clinical stages in the protein profiles of sera of patients with malignant cutaneous melanoma. No identifications were made [107]. Protein peak (3.3–30 kDa range) clustering and classification, followed by using supervised classification algorithm, generated a discriminating classification tree. Early stage melanoma recurrence was predicted with 72% sensitivity and 75% specificity [108].

4.7.4. Hepatocellular carcinoma

Aiming to develop a technique for the differentiation of carcinoma from chronic liver disease, serum protein profiles were obtained from 38 patients with carcinoma and 20 patients with chronic liver disease. Serum samples were fractionated into six fractions. Significant differences were observed in the 0.5–200 kDa mass range. Both two-way hierarchial clustering analysis and artificial neural network

algorithms were used to classify pooled serum samples. Specificity was 90% and sensitivity was 92% [109].

4.8. Other searches for biomarkers using SELDI–TOF–MS

Related investigations include: proteomic analysis of lymph [110], study of serum protein profiles in hemodialysis patients [111], a variety of applications in hematology [112–115], detection of multiple variants of serum amyloid alpha in renal cancer [116], search for biomarkers expressed by human pluripotent stem cells [117], study of the involvement of tumor necrosis superfamily members and a proliferation-inducing ligand in the resistance to apoptosis of B-CLL leukemic cells through an autocrine pathway [115] and protein profiling in neuroblastoma [118], and brain cancer [119,120]. Putative biomarkers were found in 39 human cancer cell lines [121] and the 60 human cell line panel of the NCI [122]. Development of SELDI affinity techniques is likely to be valuable for the proteomic evaluation of archival cytologic materials [123].

5. Representative other applications

5.1. Proteomic studies to uncover molecular mechanisms associated with malignancies

Most available MS technologies have been used to elucidate the proteomics of breast carcinoma. The degree of tissue heterogeneity of breast carcinomas, a serious problem obscuring quantitative comparative experiments, may be overcome by using LCM. Current emphasis is on infiltrating vs. *in situ* ductal carcinoma, aiming to uncover differential profiles for diagnosis as well as monitoring disease response to therapy [124]. The complexity of the breast cancer proteome may be simplified by concentrating on specific subcellular compartments. For example, MS-based approaches have been explored to study lysosomes, such as the aspartic protease cathepsin that has been shown to be involved in disease progression [125]. Despite the fact that removal of cells from their natural microenvironment may lead to gaining or losing certain characteristics, the *in vitro* study of cell cultures still has obvious advantages. Novel uses of MS include the measurement (GC/MS) of epithelial cell proliferation using $^{2}H_{2}O$ labeling for assessing the effects of antiproliferation chemopreventive and chemotherapeutic agents [126], and the 2D LC/MS analysis of similarities and differences between hundreds of membrane proteins in MCF7 and BT474 cell lines [127].

Several studies have been carried out to obtain proteomic profiles in human lung cancer cell lines. Proteomic signatures were obtained for different histological types of lung cancer. Hierarchial clustering analysis and principal component

analysis of separated (2D-DIGE) proteins revealed 32 proteins that were used to categorize cancer cell into distinct histological groups [128]. Investigations of the proteome of lung squamous carcinoma utilized MALDI–TOF–MS and several databases to identify some 76 differentially expressed protein spots obtained by electrophoresis [129].

Proteomic analyses of exosomes from malignant pleural effusions [130] and human mesothelioma cells [131] revealed several discrete sets of proteins involved in antigen presentation, signal transduction, migration, and adhesion, suggesting interactions between tumor cells and their environment. A large number of proteins were identified in a study of human pleural effusions including several that were suggested to play a role in the development and progression of the cancer phenotype [132].

A study of protein profiles in gastric adenocarcinoma revealed diverse alterations related to self-protection efforts of cells and changes during the malignant transformation. An 18 kDa antrum mucosa protein was significantly underexpressed in progressing tumors. It was concluded that the global consideration of the expressed profile alterations will provide insights into the pathogenesis of the tumor [133].

5.2. Proteomic profiles to provide predictors of drug-modulated targets and responses

Individuals with inherited familial adenomatous polyposis (FAP) develop numerous polyps, the premalignant precursors to colorectal carcinoma. A remarkable heterogeneity in patient response was observed in a clinical trial with a cyclooxygenase-2-inhibitor, celocoxib, which is known to be efficacious in FAP. SELDI proteomic profile revealed that a putative marker at 16,961.4 Da was a strong discriminator between response and nonresponse [134].

5.3. Profiles to identify proteins associated with disease progression

In cell line studies, an 11 kDa protein was identified by MALDI–MS and database search as S100C (calgizzarin) which is significantly downregulated in bladder cancer and is associated with poor survival; loss of S100C which was also significantly associated with poor survival in patients [135]. Two downregulated proteins, identified as isocitrate cytoplasmic and peroxiredoxin-II, were identified in both bladder cancer cell lines and human biopsies. Loss of these proteins marked the progression of malignancy [136]. In a study of tumor subsets in lung cancer, 15/1600 separated peaks provided a call-prediction model to distinguish primary tumors from metastasis and to distinguish between patients with resected nonsmall-cell lung cancer and poor prognosis from those with good prognosis [137].

5.4. Targeted biomarker detection via whole protein analysis

The developing technology of "top-down" protein identification does not rely on peptide ions for identification; thus, it avoids the dilution effect for small proteins associated with digestion. Targeted characterized proteins may be analyzed by readily available ion trap mass spectrometers. The concept was demonstrated by using top-down MS/MS for the identification of N-terminally acetylated thymosin β_4, that is expressed in certain lung adenocarcinoma cells, and is considered as a putative biomarker. This work is of interest because of the methodological details [138].

5.5. Sphingolipids in cancer pathogenesis and treatment

Ceramide, a major component of sphingolipid metabolism, functions as a tumor suppressor lipid, inducing antiproliferative and apoptotic responses in various malignant neoplastic cells. Conversely, sphingosine-1-phosphate (S1P) has been shown to be a tumor promoting lipid. Various exogenously supplied ceramides are now known to induce antiproliferative and other important cell function-related responses and thus represent a target for cancer therapy [139,140]. The development of a series of protocols for the high-throughput, structure-specific, and quantitative analysis of sphingolipids by HPLC–tandem MS (both triple quadrupole and ion trap) permits the investigation of this large and chemically complex group of compounds [141].

5.6. Quantification of antineoplastic drugs

There are literally hundreds of MS techniques described for the quantification of antineoplastic drugs in body fluids [13]. A representative publication describes the simultaneous determination of methotrexate and cyclophosphamide in urine by a validated LC–ESI–MS/MS method. The impressive lower limits of detection were 0.2 μg/L for methotrexate and 0.04 μg/L for cyclophosphamide [142]. The advantages of multiple reaction monitoring may be appreciated by reviewing a method developed for the quantification of the farnesyl transferase inhibitor lonafarnib in human plasma using HPLC coupled with tandem MS [143]. Attention is called to the increasing inclusion of validation experiments, using FDA guidelines, a common requirement for techniques used to obtain data for pharmacokinetic studies, particularly in Phase I clinical trials of new drugs (www.fda.gov/cder/guidance/4252fnl.html).

5.7. Helicobacter pylori

The stomachs of about half of the people in the world are colonized by *H. pylori,* a Gram-negative organism that is assigned as a Class 1 carcinogen. While most

colonized individuals are asymptomatic, a subpopulation of 10–20% develops peptic ulcers that may in turn evolve into adenocarcinoma; 30–90% of gastric cancers (a major health problem worldwide) are tied to this microorganism [144]. There are significant clinical and economic aspects of screening for, and diagnosis of, this infection [145].

5.7.1. Diagnosis

The "gold" standard of diagnosis is the ^{13}C-urea breath test which is based on the fact that while humans have no endogenous urease activity in the stomach, *Helicobacter* species have high urease activity. In the stomach *H. pylori* hydrolyzes ^{13}C-enriched urea to ^{13}C and NH_3; thus, the determination of the area ratios of the $^{13}CO_2$ to $^{12}CO_2$ peaks in expired air is diagnostic. Being a stable, nonradioactive isotope, ^{13}C can be administered safely to children and pregnant women. The excess ^{13}C in exhaled breath can be determined accurately with dual-inlet gas isotope-ratio mass spectrometers [146–148]. Bench-top GC/MS instruments (SIM mode) have also been evaluated for the determination of the $^{13}CO_2$ to $^{12}CO_2$ peak area ratios; both sensitivity and specificity values were in the 96–98% range [149].

5.7.2. Biomarkers of the bacterium

Lysates and extracts from six different *H. pylori* strains were analyzed by MALDI–TOF–MS. It was concluded that the strain-specific biomarkers identified might be used in a fingerprinting technique for strain typing [150]. In another MALDI–TOF–MS investigation, a potential biomarker of 58,268 Da could distinguish *H. pylori* from *H. mustalae* and *Campylobacter* species. It was concluded that, together with three strain-nonspecific markers, the technique is adequate for the rapid detection of these organisms in foods, beverages, or manufactured products [151]. Utilizing predictive information from the *H. pylori* genome, some 20 candidate proteins were identified by MALDI–TOF–MS in proteolytic digests of *H. pylori* lysates from blood samples of infected patients. It was concluded that this approach has potential for vaccine development [152]. In another approach to recognize antigenic proteins as candidates for vaccines, hundreds of proteins were separated by 2D electrophoresis and analyzed by MALDI–TOF–MS, revealing some 960 mass spectra leading to the confirmation of the presence of 24 previously unidentified proteins [153]. Use of similar methodology to study subproteomes of soluble and structure-bound *H. pylori* proteins led to the identification of several structure-bound proteins that may be candidates for diagnostic and/or vaccine investigations [154].

5.8. Molecular epidemiology for chemoprevention

All definitions of cancer chemoprevention (and there are many) include the use of chemical means for the inhibition, retardation, or reversal of the carcinogenetic process. In contrast to tumor biomarkers which are associated with established neoplasia or metastasis, there are two types of biomarkers relevant to cancer chemoprevention: risk biomarkers, referring to genetic predisposition, medical history, lifestyle, exposure, and cellular abnormalities (no detectable premalignant or malignant disease), and biomarkers of chemopreventive intervention, referring to such biological alterations of early/intermediate carcinogenesis which may be effected by chemopreventive agents [155].

An important group of relevant biomarkers are DNA adducts which originate from the chemical modification of bases in DNA or amino acids in proteins by toxic chemicals. Advantages of LC–ESI–MS methods, compared to ^{35}P-postlabeling (the "gold" standard) and GC–MS (with negative CI), include detection of highly polar compounds without derivatization, direct analyte identification, and accurate quantification using internal standards (preferably stable isotope labeling); current detection limit is $1/10^8$ nucleotides. Relevant applications include aflatoxin B_1 and hepatocellular carcinoma, oxidative DNA adducts of prostate carcinoma, and M_1G adducts in a variety of malignancies and in the predisposition to gastric adenocarcinoma resulting from infection by *H. pylori* [155].

5.9. Selenium

As a constituent of selenoproteins, selenium has several vital structural and enzymatic roles. Increased intake of selenium-enriched food has been shown to yield direct, inverse, or null associations with cancer risk [156,157]. Attempts have been made to correlate serum selenium levels with overall survival in non-Hodgkin's lymphoma [158], PC [159], and poor outcome in lung adenocarcinoma [160]. The most important selenium-containing organic compounds are selenomethionine (present in plants) and selenocysteine (present in animal proteins). A coupled capillary electrophoresis–inductively coupled mass spectrometry technique was developed for the speciation of two selenium species. The limits of detection for the two species studied in drinking water were 24 and 10 pg, respectively [161]. The predominant selenium species in both garlic (296 mg/(g Se)) and yeast (1922 mg/(g Se)) were γ-glutamyl-methylselenocysteine and selenomethionine. In rats, selenium from garlic was significantly more effective than selenium from yeast in suppressing the development of premalignant lesions and the formation of breast adenocarcinomas [162]. Several MS strategies for selenium speciation in dietary sources have been reviewed, with emphasis on validation and cancer preventive properties [163].

References

1. Hanahan, D. and Weinberg, R.A., The hallmarks of cancer. *Cell*, **100**, 57–70 (2000).
2. Futreal, P.A., A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183 (2005).
3. Ilyin, S.E., Belkowski, S.M. and Plata-Salaman, C.R., Biomarker discovery and validation: technologies and integrative approaches. *Trends Biotechnol.*, **22**, 411–416 (2004).
4. Whiteley, G.R., Proteomic patterns for cancer diagnosis—promise and challenges. *Mol. Biosyst.*, **2**, 358–363 (2006).
5. Aebersold, R., Perspective: a program to improve protein biomarker discovery for cancer. *J. Proteome. Res.*, **4**, 1104–1109 (2005).
6. deVera, I.E., Katz, J.E. and Agus, D.B., Clinical proteomics: the promises and challenges of mass spectrometry-based biomarker discovery. *Clin. Adv. Hematol. Oncol.*, **4**, 541–549 (2006).
7. Graves, P.R. and Haystead, T.A., Molecular biologist's guide to proteomics. *Microbiol. Mol. Biol. Rev.*, **66**, 39–63 (2002).
8. Hirabayashi, J. and Kasai, K., Separation technologies for glycomics. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, **771**, 67–87 (2002).
9. Sun, B., Ranish, J.A., Utleg, A.G., White, J.T., Yan, X., Lin, B. and Hood, L., Shotgun glycopeptide capture approach coupled with mass spectrometry for comprehensive glycoproteomics. *Mol. Cell Proteomic.*, **6**, 141–149 (2007).
10. Wenk, M.R., The emerging field of lipidomics. *Nat. Rev. Drug Discov.*, **4**, 594–610 (2005).
11. Yoo, H.H., Son, J., and Kim, D.H., Liquid chromatography–tandem mass spectrometric determination of ceramides and related lipid species in cellular extracts. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, **843**, 327–333 (2006).
12. van der Merwe, D.E., Oikonomopoulou, K., Marshall, J. and Diamandis, E.P., Mass spectrometry: uncovering the cancer proteome for diagnostics. *Adv. Cancer Res.*, **96**, 23–50 (2006).
13. Roboz, J., *Mass Spectrometry in Cancer Research*. CRC Press, Boca Raton, FL, 2002, pp. 1–554.
14. Howard, J.C., Heinemann, C., Thatcher, B.J., Martin, B., Gan, B.S. and Reid, G., Identification of collagen-binding proteins in *Lactobacillus* spp. with surface-enhanced laser desorption/ionization–time of flight ProteinChip technology. *Appl. Environ. Microbiol.*, **66**, 4396–4400 (2002).
15. Merchant, M. and Weinberger, S.R., Recent advancements in surface-enhanced laser desorption/ionization–time of flight–mass spectrometry. *Electrophoresis*, **21**, 1164–1177 (2000).
16. Fung, E.T., Thulasiraman, V., Weinberger, S.R., Delmasso, E.A., Protein biochips for differential profiling. *Current Opinion Biotechnol.*, **12**, 65–69 (2001).
17. Tang, N., Tornatore, P. and Weinberg, S.R., Current developments in SELDI affinity technology. *Mass Spectrom. Rev.*, **23**, 34–44 (2004).
18. Rubin, R.B. and Merchant, M., A rapid protein profiling system that speeds study of cancer and other diseases. *Am. Clin. Lab.*, **19**, 28–29 (2000).
19. Wiesner, A., Detection of tumor markers with protein ProteinChip technology. *Curr. Pharm. Biotechnol.*, **5**, 45–67 (2004).
20. Xiao, Z., Prieto, D., Conrads, T.P., Veenstra, T.D. and Issaq, H.J., Proteomic patterns: their potential for disease diagnosis. *Mol. Cell Endocrinol.*, **230**, 95–106 (2005).
21. Caputo, E., Moharram, R. and Martin, B.M., Methods for on-chip protein analysis. *Anal. Biochem.*, **321**, 116–124 (2003).
22. Lin, S., Tornatore, P., King, D., Orlando, R. and Weinberger, S.R., Limited acid hydrolysis as a means of fragmenting proteins isolated upon proteinchip array surfaces. *Proteomics*, **1**, 1172–1184 (2001).
23. Aebersold, R. and Goodlett, D.R., Mass spectrometry in proteomics. *Chem. Rev.*, **101**, 269–295 (2001).

24. Fenyo D., Identifying the proteome: software tools. *Curr. Opin. Biotechnol.*, **11**, 391–395 (2000).
25. Persson, B., Bioinformatics in protein analysis. *EXS*, **88**, 215–231 (2000).
26. Boguski, M.S. and McIntosh, M.W., Biomedical informatics for proteomics. *Nature*, **42**, 233–237 (2003).
27. Rodland, K.D., Proteomics and cancer diagnosis: the potential of mass spectrometry. *Clin. Biochem.*, **37**, 579–583 (2004).
28. Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C. and Liotta, L.A., Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**, 572–577 (2002).
29. Fung, E.T. and Enderwick, C., ProteinChip@ clinical proteomics: computational challenges and solutions. *Biotech. Suppl.*, **32**, S34–S41 (2002).
30. Qu, Y., Adam, B.-L., Yasui, Y., Ward, M.D., Cazares, L.H., Schellhammer, P.F., Feng, Z., Semmes, O.J. and Wright, G.L., Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from non-cancer patients. *Clin. Chem.*, **48**, 1835–1843 (2002).
31. Whittal, R.M. and Li, L., Time-lag focusing MALDI–TOF mass spectrometry. *Am. Lab.*, **24**, 30–36 (1997).
32. Li, J., Zhang, Z., Rosenzweig, J., Wang, Y.Y. and Chan, D.W., Proteomics and bioinformatics approach for identification of serum biomarkers to detect breast cancer. *Clin. Chem.*, **48**, 1296–1304 (2002).
33. Koopmann, J., Zhang, Z., White, N., Rosenzweig, J., Fedarko, N., Jagannath, S., Canto, M.I., Yeo, C.J., Chan, D.W. and Goggins, M., Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. *Clin. Cancer Res.*, **10**, 860–868 (2004).
34. Ball, G., Mian, S., Holding, F., Allibone, R.O., Lowe, J., Ali, S., Li, G., McCardle, S., Ellis, I.O., Creaser, C. and Rees, R.C., An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers. *Bioinformatics*, **18**, 395–404 (2002).
35. Li, L., Umbach, D.M., Terry, P. and Taylor, J.A., Application of the GA/KNN method to SELDI proteomics data. *Bioinformatics*, **20**, 1638–1640 (2004).
36. Zhu, W., Wang, X., Ma, Y., Roa, M., Glimm, J. and Kovach, J.S., Detection of cancer-specific markers amid massive mass spectral data. *PNAS*, **100**, 14666–14671 (2003).
37. Deng, X., Geng, H., Bastola, D.R. and Ali, H.H., Link test—a statistical method for finding prostate cancer biomarkers. *Comput. Biol. Chem.*, **30**, 425–433 (2006).
38. Pizzo, P., After empiric therapy: what to do until the granulocyte comes back. *Rev. Infect. Dis.*, **9**, 214–219 (1987).
39. Diamandis, E.P., Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J. Natl. Cancer Inst.*, **96**, 353–356 (2004).
40. Diamandis, E.P., Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell Proteomic.*, **3**, 367–378 (2004).
41. Baggerly, K.A., Morris, J.S. and Coombes, K.R., Reproducibility of SELDI–TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, **20**, 777–785 (2004).
42. Villanueva, J., Correcting common errors in identifying cancer-specific serum peptide signatures. *J. Proteome Res.*, **4**, 1060–1072 (2005).
43. Aivado, M., Spentzos, D., Alterovitz, G., Otu, H.H., Grall, F., Giagounidis, A.A., Wells, M., Cho, J.Y., Germing, U., Czibere, A., Prall, W.C., Porter, C., Ramoni, M.F. and Libermann, T.A., Optimization and evaluation of surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI–TOF MS) with reversed-phase protein arrays for protein profiling. *Clin. Chem. Lab. Med.*, **43**, 133–140 (2005).

44. Dijkstra, M., Vonk, R.J. and Jansen, R.C., SELDI–TOF mass spectra: a view on sources of variation. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci*., **847**, 12–23 (2006).
45. Diamandis, E.P., Serum proteomic profiling by matrix-assisted laser desorption–ionization time-of-flight mass spectrometry for cancer diagnosis: next steps. *Cancer Res.*, **66**, 5540–5541 (2006).
46. Engwegen, J.Y., Gast, M.C., Schellens, J.H. and Beijnen, J.H., Clinical proteomics: searching for better tumour markers with SELDI–TOF mass spectrometry. *Trends Pharmacol. Sci.*, **27**, 251–259 (2006).
47. Emmert-Buck, M.R., Bonner, R.F., Smith, P.D., Chuaqui, R.F., Zhuang, Z., Goldstein, S.R., Weiss, R.A. and Liotta, L.A., Laser capture microdissection. *Science*, **274**, 998–1001 (1996).
48. Craven, R.A. and Banks, R.E., Laser capture microdissection and proteomics: possibilities and limitation. *Proteomics*, **1**, 1200–1204 (2001).
49. Caldwell, R.L. and Caprioli, R.M., Tissue profiling by mass spectrometry: a review of methodology and applications. *Mol. Cell Proteomic.*, **4**, 394–401 (2005).
50. Reyzer, M.L., MALDI mass spectrometry for direct tissue analysis: a new tool for biomarker discovery. *J. Proteome Res.*, **4**, 1138–1142 (2005).
51. Chaurand, P., Schwartz, S.A., Reyzer, M.L. and Caprioli, R.M., Imaging mass spectrometry: principles and potentials. *Toxicol. Pathol.*, **33**, 92–101 (2005).
52. Brunelle, A., Biological tissue imaging with time-of-flight secondary ion mass spectrometry and cluster ion sources. *J. Mass Spectrom.*, **40**, 985–999 (2005).
53. Smith, R.D., Shen, Y. and Tang, K., Ultrasensitive and quantitative analyses from combined separations-mass spectrometry for the characterization of proteomes. *Acc. Chem. Res.*, **37**, 269–278 (2004).
54. Merrell, K., Southwick, K., Graves, S.W., Esplin, M.S., Lewis, N.E. and Thulin, C.D., Analysis of low-abundance, low-molecular-weight serum proteins using mass spectrometry. *J. Biomol. Tech.*, **15**, 238–248 (2004).
55. Zhang, H., Yan, W. and Aebersold, R., Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. *Curr. Opin. Chem. Biol.*, **8**, 66–75 (2004).
56. Dijkstra, M., Roelofsen, H., Vonk, R.J. and Jansen, R.C., Peak quantification in surface-enhanced laser desorption/ionization by using mixture models. *Proteomics*, **6**, 5106–5116 (2006).
57. Aebersold, R. and Mann, M., Mass spectrometry-based proteomics. *Nature*, **422**, 198–207 (2003).
58. MacCoss, M.J. and Matthews, D.E., Quantitative MS for proteomics: teaching a new dog old tricks. *Anal. Chem.*, **77**, 295A–302A (2005).
59. Higgs, R.E., Comprehensive label-free method for the relative quantification of proteins from biological samples. *J. Proteome Res.*, **4**, 1442–1450 (2005).
60. Jain, K.K., Recent advances in oncoproteomics. *Curr. Opin. Mol. Ther.*, **4**, 203–209 (2002).
61. Gillette, M., Place of pattern in proteomic biomarker discovery. *J. Proteome Res.*, **4**, 1143–1154 (2005).
62. Roboz, J., Mass spectrometry in diagnostic oncoproteomics. *Cancer Invest.*, **23**, 465–478 (2005).
63. Fung, E.T., Strategies in clinical proteomics: from discovery to assay. *Preclinica*, **2**, 253–258 (2004).
64. Hackett, J.L. and Gutman, S.I., Introduction to the food and drug administration (FDA) regulatory process. *J. Proteome Nat.*, **4**, 1110–1113 (2005).
65. Pritzker, K.P., Cancer biomarkers: easier said than done. *Clin. Chem.*, **48**, 1147–1150 (2002).
66. Diamandis, E.P., Proteomic patterns in serum and identification of ovarian cancer. *Lancet*, **360**, 170 (2002).
67. Diamandis, E.P., Point: proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin. Chem.*, **49**, 1272–1275 (2003).

68. Petricoin, E.F. and Liotta, L.A., Counterpoint: the vision for a new diagnostic paradigm. *Clin. Chem.*, **49**, 1276–1278 (2003).
69. Alexe, G., Alexe, S., Liotta, L.A., Petricoin, E.F., Reiss, M. and Hammer, P.L., Ovarian cancer detection by logical analysis of proteomic data. *Proteomics*, **4**, 766–783 (2004).
70. Zhang, Z., Bast, Jr., R.C., Yu, Y., Li, J., Sokoll, L.J., Rai, A.J., Rosenzweig, J.M., Cameron, B., Wang, Y.Y., Meng, X.Y., Berchuck, A., Van Haaften-Day, C., Hacker, N.F., de Bruijn, H.W., van der Zee, A.G., Jacobs, I.J., Fung, E.T. and Chan, D.W., Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res.*, **64**, 5882–5890 (2004).
71. Fung, E.T., Yip, T.T., Lomas, L., Wang, Z., Yip, C., Meng, X.Y., Lin, S., Zhang, F., Zhang, Z., Chan, D.W. and Weinberger, S.R., Classification of cancer types by measuring variants of host response proteins using SELDI serum assays. *Int. J. Cancer*, **115**, 783–789 (2005).
72. Conrads, T.P., Fusaro, V.A., Ross, S., Johann, D., Rajapakse, V., Hitt, B.A., Steinberg, S.M., Kohn, E.C., Fishman, D.A., Whitely, G., Barrett, J.C., Liotta, L.A., Petricoin, E.F., III and Veenstra, T.D., High-resolution serum proteomic features for ovarian cancer detection. *Endocr. Relat. Cancer*, **11**, 163–178 (2004).
73. Yang, E.C., Guo, J., Diehl, G., DeSouza, L., Rodrigues, M.J., Romaschin, A.D., Colgan, T.J., Siu, K.W., Protein expression profiling of endometrial malignancies reveals a new tumor marker: chaperonin 10. *J. Proteome Res.*, **3**, 636–643 (2004).
74. Zhu, L.R., Zhang, W.Y., Yu, L., Zheng, Y.H., Zhang, J.Z. and Liao, Q.P., Serum proteomic features for detection of endometrial cancer. *Int. J. Gynecol. Cancer*, **16**, 1374–1378 (2006).
75. Bertucci, F., Birnbaum, D. and Goncalves, A., Proteomics of breast cancer: principles and potential clinical applications. *Mol. Cell Proteomic.*, **5**, 1772–1786 (2006).
76. Liotta, L. and Petricoin, E.F., Molecular profiling of human cancer. *Nat. Rev. Genet.*, **1**, 48–56 (2000).
77. Hu, Y., SELDI–TOF–MS: the proteomics and bioinformatics approaches in the diagnosis of breast cancer. *Breast*, **14**, 250–255 (2005).
78. Dua, R.S., Isacke, C.M. and Gui, G.P., The intraductal approach to breast cancer biomarker discovery. *J. Clin. Oncol.*, **24**, 1209–1216 (2006).
79. Sauter, E.R., Shan, S., Hewett, J.E., Speckman, P., and Du Bois, G.C., Proteomic analysis of nipple aspirate fluid using SELDI–TOF–MS. *Int. J. Cancer*, **114**, 791–796 (2005).
80. Pawlik, T.M., Fritsche, H., Coombes, K.R., Xiao, L., Krishnamurthy, S., Hunt, K.K., Pusztai, L., Chen, J.N., Clarke, C.H., Arun, B., Hung, M.C., and Kuerer, H.M., Significant differences in nipple aspirate fluid protein expression between healthy women and those with breast cancer demonstrated by time-of-flight mass spectrometry. *Breast Cancer Res. Treat.*, **89**, 149–157 (2005).
81. Coombes, K.R., Fritsche, H.A., Clarke, C., Chen, J., Baggerly, K.A., Morris, J.S., Xiao, L., Hung, M. and Kuerer, H.M., Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin. Chem.*, **49**, 1615–1623 (2003).
82. Nakagawa, T., Huang, S.K., Martinez, S.R., Tran, A.N., Elashoff, D., Ye, X., Turner, R.R., Giuliano, A.E. and Hoon, D.S., Proteomic profiling of primary breast cancer predicts axillary lymph node metastasis. *Cancer Res.*, **66**, 11825–11830 (2006).
83. Wright, G.L., Cazares, L.H., Leung, S.-M., Nasim, S., Adam, B.-L., Yip, T.-T., Schellhammer, P.F., Gong, L. and Vlahou, A., Proteinchip surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer Prostatic Dis.*, **2**, 264–276 (1999).
84. Hessels, D., Verhaegh, G.W., Schalken, J.A. and Witjes, J.A., Applicability of biomarkers in the early diagnosis of prostate cancer. *Exp. Rev. Mol. Diagn.*, **4**, 513–526 (2004).
85. Diamandis, E.P., Re: diagnostic potential of serum proteomic patterns in prostate cancer. *J. Urol.*, **171**, 1244–1245 (2004).

86. Wagner, M., Naik, D.N., Pothen, A., Kasukurti, S., Devineni, R.R., Adam, B.L., Semmes, O.J. and Wright, Jr., G.L., Computational protein biomarker prediction: a case study for prostate cancer. *BMC Bioinformatics*, **5** (26) (2004).
87. Petricoin, E.F., Ornstein, K., Paweletz, C.P., Ardekani, A., Hackett, P.S., Hitt, B.A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C.B., Levine, P.J., Lienhan, W.M., Emmert-Buck, M.R., Steinberg, S.M., Kohn, E.C. and Liotta, L.A., Serum proteomic patterns for detection of prostate cancer. *J. Natl. Cancer Inst.*, **95**, 489–490 (2003).
88. Tong, W., Xie, Q., Hong, H., Shi, L., Fang, H., Perkins, R. and Petricoin, E.F., Using decision forest to classify prostate cancer samples on the basis of SELDI–TOF MS data: assessing chance correlation and prediction confidence. *Environ. Health Perspect.*, **112**, 1622–1627 (2004).
89. Semmes, O.J., Feng, Z., Adam, B.L., Banez, L.L., Bigbee, W.L., Campos, D., Cazares, L.H., Chan, D.W., Grizzle, W.E., Izbicka, E., Kagan, J., Malik, G., McLerran, D., Moul, J.W., Partin, A., Prasanna, P., Rosenzweig, J., Sokoll, L.J., Srivastava, S., Srivastava, S., Thompson, I., Welsh, M.J., White, N., Winget, M., Yasui, Y., Zhang, Z. and Zhu, L., Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. *Clin. Chem.*, **51**, 102–112 (2005).
90. Lehrer, S., Roboz, J., Ding, H., Zhao, S., Diamond, E.J., Holland, J.F., Stone, N.N., Droller, M.J. and Stock, R.G., Putative protein markers in the sera of men with prostatic neoplasms. *Br. J. Urol. Int.*, **92**, 223–225 (2003).
91. Hlavaty, J.J., Partin, A.W., Shue, M.J., Mangold, L.A., Derby, J., Javier, T., Kelley, S., Stieg, A., Briggman, J.V., Hass, G.M. and Wu, Y., Identification and preliminary clinical evaluation of a 50.8-kDa serum marker for prostate cancer. *Urology*, **61**, 1261–1265 (2003).
92. Okusaka, T., Yamada, T. and Maekawa, M., Serum tumor markers for pancreatic cancer: the dawn of new era? *JOP*, **7**, 332–336 (2006).
93. Garcea, G., Neal, C.P., Pattenden, C.J., Steward, W.P. and Berry, D.P., Molecular prognostic markers in pancreatic cancer: a systematic review. *Eur. J. Cancer*, **41**, 2213–2236 (2005).
94. Sasaki, K., Sato, K., Akiyama, Y., Yanagihara, K., Masaaki, O. and Yamauchi, K., Peptidomics-based approach reveals the secretion of the 29-residue COOH-terminal fragment of the putative tumor suppressor protein DMBT1 from pancreatic adenocarcinoma cell lines. *Cancer Res.*, **62**, 4894–4898 (2002).
95. Rosty, C., Christa, L., Kuzdzal, S., Baldwin, W.M., Zahurak, M.L., Carnot, F., Chan, D.W., Canto, M., Lillemoe, K.D., Cameron, J.L., Yeo, C.J., Hruban, R.H. and Goggins, M., Identification of hepatocarcinoma–intestine–pancreas/pancreatitis-associated protein I as a biomarker for pancreatic ductal adenocarcinoma by protein biochip technology. *Cancer Res.*, **62**, 1868–1875 (2002).
96. Honda, K., Hayashida, Y., Umaki, T., Okusaka, T., Kosuge, T., Kikuchi, S., Endo, M., Tsuchida, A., Aoki, T., Itoi, T., Moriyasu, F., Hirohashi, S. and Yamada, T., Possible detection of pancreatic cancer by plasma protein profiling. *Cancer Res.*, **65**, 10613–10622 (2005).
97. Vlahou, A., Schellhammer, P.F., Mendrinos, S., Patel, K., Kondylis, F.I., Gong, L., Nasim, S. and Wright, G.L.J., Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. *Am. J. Pathol.*, **158**, 1491–1502 (2001).
98. Zhang, Y.F., Wu, D.L., Guan, M., Liu, W.W., Wu, Z., Chen, Y.M., Zhang, W.Z. and Lu, Y., Tree analysis of mass spectral urine profiles discriminates transitional cell carcinoma of the bladder from noncancer patient. *Clin. Biochem.*, **37**, 772–779 (2004).
99. Liu, W., Guan, M., Wu, D., Zhang, Y., Wu, Z., Xu, M. and Lu, Y., Using tree analysis pattern and SELDI–TOF–MS to discriminate transitional cell carcinoma of the bladder cancer from noncancer patients. *Eur. Urol.*, **47**, 456–462 (2005).

100. Schaub, S., Wilkins, J., Weiler, T., Sangster, K., Rush, D. and Nickerson, P., Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney Int.*, **65**, 323–332 (2004).
101. Wu, W., Tang, X., Hu, W., Lotan, R., Hong, W.K. and Mao, L., Identification and validation of metastasis-associated proteins in head and neck cancer cell lines by two-dimensional electrophoresis and mass spectrometry. *Clin. Exp. Metastasis*, **19**, 319–326 (2002).
102. Wadsworth, J.T., Somers, K.D., Stack, B.C.J., Cazares, L., Malik, G., Adam, B.L., Wright, G.L. and Semmes, O.J., Identification of patients with head and neck cancer using serum protein profiles. *Arch. Otolaryngol. Head Neck Surg.*, **130**, 98–104 (2004).
103. Dolios, G., Roboz, J. and Wang, R., *Identification of Colon Cancer associating Protein in Plasma using MALDI–TOF Mass Spectrometry. 51st Annual Conference on Mass Spectrometry*, Montereal, Quebec, Canada, 2003.
104. Liu, X.P., Shen, J., Li, Z.F., Yan, L. and Gu, J., A serum proteomic pattern for the detection of colorectal adenocarcinoma using surface enhanced laser desorption and ionization mass spectrometry. *Cancer Invest.*, **24**, 747–753 (2006).
105. Su, Y., Shen, J., Qian, H., Ma, H., Ji, J., Ma, H., Ma, L., Zhang, W., Meng, L., Li, Z., Wu, J., Jin, G., Zhang, J. and Shou, C., Diagnosis of gastric cancer using decision tree classification of mass spectral data. *Cancer Sci.*, **98**, 37–43 (2006).
106. Zhukov, T.A., Johanson, R.A., Cantor, A.B., Clark, R.A. and Tockman, M.S., Discovery of distinct protein profiles specific for lung tumors and premalignant lung lesions by SELDI mass spectrometry. *Lung Cancer*, **40**, 267–279 (2003).
107. Ferrari, L., Seraglia, R., Rossi, C.R., Bertazzo, A., Lise, M., Allegri, G. and Traldi, P., Protein profiles in sera of patients with malignant cutaneous melanoma. *Rapid Commun. Mass Spectrom.*, **14**, 1149–1154 (2000).
108. Wilson, L.L., Tran, L., Morton, D.L. and Hoon, D.S., Detection of differentially expressed proteins in early-stage melanoma patients using SELDI–TOF mass spectrometry. *Ann. N. Y. Acad. Sci.*, **1022**, 317–322 (2004).
109. Poon, T.C.W., Yip, T., Chan, A.T.C., Yip, C., Yip, V., Mok, T.S.K., Lee, C.C.Y., Leung, T.W.T., Ho, S.K.W. and Johnson, P.J., Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. *Clin. Chem.*, **49**, 752–760 (2003).
110. Leak, L.V., Liotta, L.A., Krutzsch, H., Jones, M., Fusaroa, V.A., Ross, S.J., Zhao, Y. and Petricoin, E.F.3., Proteomic analysis of lymph. *Proteomics*, **4**, 753–765 (2004).
111. Langlois, R.G., Trebes, J.E., Dalmasso, E.A., Ying, Y., Davies, R.W., Curzi, M.P., Colston, B.W., Turteltaub, K.W., Perkins, J., Chromy, B.A., Choi, M.W., Murphy, G.A., Fitch, J.P. and McCutchen-Maloney, S.L., Serum protein profile alterations in hemodialysis patients. *Am. J. Nephrol.*, **24**, 268–274 (2004).
112. Miguet, L., Bogumil, R., Decloquement, P. et al., Discovery and identification of potential biomarkers in a prospective study of chronic lymphoid malignancies using SELDI–TOF–MS. *J. Proteome Res.*, **5**, 2258–2269 (2006).
113. Rees-Unwin, K.S., Morgan, G.J. and Davies, F.E., Proteomics and the haematologist. *Clin. Lab. Haematol.*, **26**, 77–86 (2004).
114. Albitar, M., Potts, S.J., Giles, F.J., O'Brien, S., Keating, M., Thomas, D., Clarke, C., Jilani, I., Aguilar, C., Estey, E. and Kantarjian, H., Proteomic-based prediction of clinical behavior in adult acute lymphoblastic leukemia. *Cancer*, **106**, 1587–1594 (2006).
115. Kern, C., Cornuel, J.F., Billard, C., Tang, R., Rouillard, D., Stenou, V., Defrance, T., Ajchembaum-Cymbalista, F., Simonin, P.Y., Feldblum, S. and Kolb, J.P., Involvement of BAFF and APRIL in the resistance to apoptosis of B-CLL through an autocrine pathway. *Blood*, **103**, 679–688 (2004).
116. Tolson, J., Bogumil, R., Brunst, E., Beck, H., Elsner, R., Humeny, A., Kratzin, H., Deeg, M., Kuczyk, M., Mueller, G.A., Mueller, C.A. and Flad, T., Serum protein profiling by SELDI

mass spectrometry: detection of multiple variants of serum amyloid alpha in renal cancer patients. *Lab. Invest.*, **84**, 845–856 (2004).
117. Hayman, M.W. and Przyborski, S.A., Proteomic identification of biomarkers expressed by human pluripotent stem cells. *Biochem. Biophys. Res. Commun.*, **316**, 918–923 (2004).
118. Combaret, V., Bergeron, C., Brejon, S., Iacono, I., Perol, D., Negrier, S. and Puisieux, A., Protein chip array profiling analysis of sera from neuroblastoma patients. *Cancer Lett*., **228**, 91–96 (2005).
119. Schwartz, S.A., Weil, R.J., Johnson, M.D., Toms, S.A. and Caprioli, R.M., Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression. *Clin. Cancer Res.*, **10**, 981–987 (2004).
120. Liu, J., Zheng, S., Yu, J.K., Zhang, J.M. and Chen, Z., Serum protein fingerprinting coupled with artificial neural network distinguishes glioma from healthy population or brain benign tumor. *J. Zhejiang Univ. Sci. B*, **6**, 4–10 (2005).
121. Shiwa, M., Nishimura, Y., Wakatabe, R., Fukawa, A., Arikuni, H., Ota, H., Kato, Y. and Yamori, T., Rapid discovery and identification of a tissue-specific tumor biomarker from 39 human cancer cell lines using the SELDI proteinchip platform. *Biochem. Biophys. Res. Commun.*, **309**, 18–25 (2003).
122. Nishizuka, S., Charboneau, L., Young, L., Major, S., Reinhold, W.C., Waltham, M., Kouros-Mehr, H., Bussey, K.J., Lee, J.K., Espina, V., Munson, P.J., Petricoin, E.F.3., Liotta, L.A. and Weinstein, J.N., Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 14229–14234 (2003).
123. Fetsch, P.A., Simone, N.L., Bryand-Greenwood, P.K., Marincola, F.M., Filie, A.C., Petricoin, E.F., Liotta, L.A. and Abati, A., Proteomics evaluation of archival cytologic material using SELDI affinity mass spectrometry. *Am. J. Clin. Pathol.*, **118**, 870–876 (2002).
124. Somiari, R.I., Somiari, S., Russell, S. and Shriver, C.D., Proteomics of breast carcinoma. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, **815**, 215–225 (2005).
125. Journet, A. and Ferro, M., The potentials of MS-based subproteomic approaches in medical science: the case of lysosomes and breast cancer. *Mass Spectrom. Rev.*, **23**, 393–442 (2004).
126. Misell, L.M., Hwang, E.S., Au, A., Esserman, L. and Hellerstein, M.K., Development of a novel method for measuring in vivo breast epithelial cell proliferation in humans. *Breast Cancer Res. Treat.*, **89**, 257–264 (2005).
127. Xiang, R., Shi, Y., Dillon, D.A., Negin, B., Horvath, C. and Wilkins, J.A., 2D LC/MS analysis of membrane proteins from breast cancer cell lines MCF7 and BT474. *J. Proteome Res.*, **3**, 1278–1283 (2004).
128. Seike, M., Kondo, T., Fujii, K., Okano, T., Yamada, T., Matsuno, Y., Gemma, A., Kudoh, S. and Hirohashi, S., Proteomic signatures for histological types of lung cancer. *Proteomics*, **5**, 2939–2948 (2005).
129. Li, C., Xiao, Z., Chen, Z., Zhang, X., Li, J., Wu, X., Li, X., Yi, H., Li, M., Zhu, G. and Liang, S., Proteome analysis of human lung squamous carcinoma. *Proteomics*, **6**, 547–558 (2006).
130. Bard, M.P., Hegmans, J.P., Hemmes, A., Luider, T.M., Willemsen, R., Severijnen, L.A., van Meerbeeck, J.P., Burgers, S.A., Hoogsteden, H.C. and Lambrecht, B.N., Proteomic analysis of exosomes isolated from human malignant pleural effusions. *Am. J. Respir. Cell Mol. Biol.*, **31**, 114–121 (2004).
131. Hegmans, J.P., Bard, M.P., Hemmes, A., Luider, T.M., Kleijmeer, M.J., Prins, J.B., Zitvogel, L., Burgers, S.A., Hoogsteden, H.C. and Lambrecht, B.N., Proteomic analysis of exosomes secreted by human mesothelioma cells. *Am. J. Pathol.*, **164**, 1807–1815 (2004).
132. Tyan, Y.C., Proteomic profiling of human pleural effusion using two-dimensional nano liquid chromatography tandem mass spectrometry. *J. Proteome Res.*, **4**, 1274–1286 (2005).
133. He, Q.Y., Cheung, Y.H., Leung, S.Y., Yuen, S.T., Chu, K.M. and Chiu, J.F., Diverse proteomic alterations in gastric adenocarcinoma. *Proteomics*, **4**, 3276–3287 (2004).

134. Xiao, Z., Serum proteomic profiles suggest celecoxib-modulated targets and response predictors. *Cancer Res.*, **64**, 2904–2909 (2005).
135. Memon, A.A., Sorensen, B.S., Meldgaard, P., Fokdal, L., Thykjaer, T. and Nexo, E., Down-regulation of S100C is associated with bladder cancer progression and poor survival. *Clin. Cancer Res.*, **11**, 606–611 (2005).
136. Memon, A.A., Chang, J.W., Oh, B.R. and Yoo, Y.J., Identification of differentially expressed proteins during human urinary bladder cancer progression. *Cancer Detect. Prev.*, **29**, 249–255 (2005).
137. Yanagisawa, K., Shyr, Y., Xu, B.J., Massion, P.P., Larsen, P.H., White, B.C., Roberts, J.R., Edgerton, M., Gonzalez, A., Nadaf, S., Moore, J.H., Caprioli, R.M. and Carbone, D.P., Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet*, **362**, 433–439 (2003).
138. Yan, F., He, M., Hogan, J.M., Rossie, S.S. and McLuckey, S.A., Targeted biomarker detection via whole protein ion trap tandem mass spectrometry: thymosin beta(4) in a human lung cancer cell line. *J. Mass Spectrom.*, **40**, 444–451 (2005).
139. Reynolds, C.P., Ceramide synthesis and metabolism as a target for cancer therapy. *Cancer Lett.*, **206**, 169–180 (2004).
140. Ogretmen, B., Biologically active sphingolipids in cancer pathogenesis and treatment. *Nat. Rev. Cancer*, **4**, 604–616 (2005).
141. Merrill Jr., A.H., Sullards, M.C., Allegood, J.C., Kelly, S. and Wang, E., Sphingolipidomics: high-throughput, structure-specific, and quantitative analysis of sphingolipids by liquid chromatography tandem mass spectrometry. *Methods*, **36**, 207–224 (2005).
142. Barbieri, A., Sabatini, L., Accorsi, A., Roda, A. and Violante, F.S., Simultaneous determination of *t*,*t*-muconic, *S*-phenylmercapturic and *S*-benzylmercapturic acids in urine by a rapid and sensitive liquid chromatography/electrospray tandem mass spectrometry method. *Rapid Commun. Mass Spectrom.*, **18**, 1983–1988 (2004).
143. Appels, N.M., Quantitative analysis of the farnesyl transferase inhibitor lonafarnib (Sarasar, SCH66336) in human plasma using high-performance liquid chromatography coupled with tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **19**, 2187–2192 (2005).
144. Crowe, S.E., *Helicobacter* infection, chronic inflammation, and the development of malignancy. *Curr. Opin. Gastroenterol.*, **21**, 32–38 (2005).
145. Fendrick, A.M., Chernew, M.E., Hirth, R.A., Bloom, B.S., Bandekar, R.R. and Scheiman, J.M., Clinical and economic effects of population-based *Helicobacter pylori* screening to prevent gastric cancer. *Arch. Intern. Med.*, **159**, 142–148 (1999).
146. Lotterer, E., Ludtke, F.E., Tegeler, R., Lepsien, G. and Bauer, F.E., The ^{13}C-urea breath test—detection of *Helicobacter pylori* infection in patients with partial gastrectomy. *Z. Gastroenterol.*, **31**, 115–119 (1993).
147. Cutler, A.F. and Toskes, P., Comparison of ^{13}C-urea blood test to ^{13}C-urea breath test for the diagnosis of *Helicobacter pylori*. *Am. J. Gastroenterol.*, **94**, 959–961 (1999).
148. Chey, W.D., Murthy, U., Toskes, P., Carpenter, S. and Laine, L., The ^{13}C-urea blood test accurately detects active *Helicobacter pylori* infection: a United States, multicenter trial. *Am. J. Gastroenterol.*, **94**, 1522–1524 (1999).
149. Lee, H.S., Gwee, K.A., Teng, L.Y., Kang, J.Y., Yeoh, K.G., Wee, A. and Chua, B.C., Validation of ^{13}C-urea breath test for *Helicobacter pylori* using a simple gas chromatograph-mass selective detector. *Eur. J. Gastroenterol. Hepatol.*, **10**, 569–572 (1998).
150. Nilsson, C.L., Fingerprinting of *Helicobacter pylori* strains by matrix-assisted laser desorption/ionization mass spectrometric analysis. *Rapid Commun. Mass Spectrom.*, **13**, 1067–1071 (1999).
151. Winkler, M.A., Uher, J. and Cepa, S., Direct analysis and identification of *Helicobacter* and *Campylobacter* species by MALDI–TOF mass spectrometry. *Anal. Chem.*, **71**, 3416–3419 (1999).

152. McAtee, C.P., Fry, K.E. and Berg, D.E., Identification of potential diagnostic and vaccine candidates of *Helicobacter pylori* by "proteome" technologies. *Helicobacter*, **3**, 163–169 (1998).
153. Krah, A., Schmidt, F., Becher, D., Schmid, M., Albrecht, D., Rack, A., Buttner, K. and Jungblut, P.R., Analysis of automatically generated peptide mass fingerprints of cellular proteins and antigens from *Helicobacter pylori* 26695 separated by two-dimensional electrophoresis. *Mol. Cell Proteomic.*, **2**, 1271–1283 (2003).
154. Backert, S., Kwok, T., Schmid, M., Selbach, M., Moese, S., Peek, Jr., R.M., Konig, W., Meyer, T.F. and Jungblut, P.R., Subproteomes of soluble and structure-bound *Helicobacter pylori* proteins analyzed by two-dimensional gel electrophoresis and mass spectrometry. *Proteomics*, **5**, 1331–1345 (2005).
155. Sharma, R.A. and Farmer, P.B., Biological relevance of adduct detection to the chemoprevention of cancer. *Clin. Cancer Res.*, **10**, 4901–4912 (2004).
156. Rayman, M.P., The importance of selenium to human health. *Lancet*, **356**, 233–241 (2001).
157. Alaejos, M.S., Diaz Romero, F.J. and Diaz, R.C., Selenium and cancer: some nutritional aspects. *Nutrition*, **16**, 376–383 (2001).
158. Last, K.W., Cornelius, V., Delves, T., Sieniawska, C., Fitzgibbon, J., Norton, A., Amess, J., Wilson, A., Rohatiner, A.Z. and Lister, T.A., Presentation serum selenium predicts for overall survival, dose delivery, and first treatment response in aggressive non-Hodgkin's lymphoma. *J. Clin. Oncol.*, **21**, 2335–2341 (2003).
159. Lipsky, K., Zigeuner, R., Zischka, M., Schips, L., Pummer, K., Rehak, P. and Hubmer, G., Selenium levels of patients with newly diagnosed prostate cancer compared with control group. *Urology*, **63**, 912–916 (2004).
160. Chen, G., Wang, H., Miller, C.T., Thomas, D.G., Gharib, T.G., Misek, D.E., Giordano, T.J., Orringer, M.B. Hanash, S.M. and Beer, D.G., Reduced selenium-binding protein 1 expression is associated with poor outcome in lung adenocarcinomas. *J. Pathol.*, **202**, 321–329 (2004).
161. Magnuson, M.L., Creed, J.T. and Brockhoff, C.A., Speciation of selenium and arsenic compounds by capillary electrophoresis with hydrodynamically modified electroosmotic flow and on-line reduction of selenium(VI) to selenium(IV) with hydride generation inductively coupled plasma mass spectrometric detection. *Analyst*, **122**, 1057–1061 (1997).
162. Ip, C., Birringer, M., Block, E., Kotrebai, M., Tyson, J.F., Uden, P.C. and Lisk, D.J., Chemical speciation influences comparative activity of selenium-enriched garlic and yeast in mammary cancer prevention. *J. Agric. Food Chem.*, **48**, 2062–2070 (2001).
163. Infante, H.G., Hearn, R. and Catterick, T., Current mass spectrometry strategies for selenium speciation in dietary sources of high-selenium. *Anal. Bioanal. Chem.*, **382**, 957–967 (2005).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 18

Application of mass spectrometry in brain research: neurotransmitters, neuropeptides, and brain proteins

LASZLO PROKAI*

Department of Molecular Biology & Immunology, University of North Texas Health Science Center, Fort Worth, TX 76107, USA

1. Introduction

The anatomical and cellular complexity of the mammalian central nervous system (CNS) with its vast number of synapses and associated intricate biochemical processes often presents technical challenges to the application of mass spectrometry. The neurotransmitter acetylcholine, for example, is rapidly inactivated by the

*Tel.: +1 (817) 735-2206; Fax: +1 (817) 735-2118. E-mail: lprokai@hsc.unt.edu (L. Prokai).

enzyme acetylcholinesterase, and therefore an appropriate sampling method is necessary to determine its brain concentration in different neurophysiologic states or upon different treatments. Neuropeptides that fulfill many important functions in the CNS also present obstacles to their exploration by mass spectrometry because of their low tissue levels, limited biostability, and degradation background from brain proteins, when their detection is pursued. Potential solutions briefly discussed here range from sophisticated *in vivo* sampling techniques such as microdialysis to microwave tissue irradiation to inhibit or minimize postmortem enzymatic degradation for the practice of neuropeptidomics. An additional subject discussed in this chapter is mass spectrometry-based neuroproteomics. Critical issues that require attention to realize the enormous potential of neuroproteomics include appropriate tissue preparation methods to focus on a relevant subproteome, combination with separation techniques to simplify complex mixtures and enrich desired brain proteins or peptides obtained through their proteolytic degradation, and development of methods that allow for quantification. Expression profiling of synaptic plasma-membrane proteins and potential exploitation of a quantitative proteomics approach to study neurodegenerative conditions are discussed as representative examples to show the power of mass spectrometry-based methods in this field.

Although the brain is considered to be a single organ, its internal structure and organization is extraordinarily complex. The human (mammalian) brain contains about 10^{11} neurons, each neuron contains an axon (a long process leading from the cell body to another cell to propagate the action potential, an electrical activity common to all neurons), and most axons make functional connections with other neurons at junctions called synapses [1]. Synapses are means of interneuronal communication by making direct electrical contacts and, through converting action potential to chemical signals, by the movement of molecules such as neurotransmitters between cells. Neurons are arranged in many groupings, called nuclei, related by function. In addition, there are large numbers of other cells called glia (astrocytes, oligodendrocytes, and microglia) in the brain. Anatomical, cellular, and biochemical complexities of the CNS often present challenges for analytical methods to adequately reflect constituents and processes associated with its biochemistry, physiology, pathology, and pharmacotherapy. This chapter focuses on selected areas of brain research that exemplify these challenges, and discusses potential solutions to address them with the power of mass spectrometry.

2. Methodology

2.1. Neurotransmitters

Many neurochemicals are stored usually in synaptic vesicles of the synapses. They may be released into the so-called synaptic clefts by an action potential [2]. In the synaptic cleft, transmitter molecules diffuse across the extracellular space

into the postsynaptic membrane to interact with their specific receptors to effect cellular changes that are specific for these receptors in the postsynaptic cell. Neurotransmitters may be small molecules such as amino acid transmitters, biogenic amines, and acetylcholine, or they may also be various neuropeptides.

Small-molecule neurotransmitters are usually found in large concentration in the brain, and numerous methods that do not utilize mass spectrometric detection have been developed to analyze them [3–5]. Acetylcholine [ACh, $(CH_3)_3N^+CH_2CH_2OCOCH_3$] has, however, presented a challenge. The extracellular concentration of ACh in the mammalian brain is typically very low due to the rapid hydrolysis of ACh to choline [$(CH_3)_3N^+CH_2CH_2OH$] by acetylcholinesterase. In addition, postmortem degradation of ACh has been an additional issue in studies that relied on the removal of the brain of animals sacrificed for the determination of this neurotransmitter. Early *in vivo* experiments, such as those using push–pull cannulae [6], also suffered from degradation of ACh in the sample prior to its detection. The introduction of microdialysis as the least invasive way of monitoring transmitter release *in vivo* [7] has provided a solution to this problem.

2.1.1. In vivo *microdialysis*

Microdialysis employs a semipermeable hollow-fiber membrane implanted in the tissue. It allows for the sampling of chemicals from the extracellular space of the brain, when the implanted probe is perfused at low flow rates and usually with a solution mimicking the composition of the cerebrospinal fluid (CSF). Collection times of about 5–30 min are typically used depending on the time resolution desired, concentration of the analyte, and detection limit of the assay. The scheme of a typical experimental setup to perform *in vivo* microdialysis from conscious, freely moving animals (usually rats) is shown in Fig. 1. Upon using membranes with appropriately sized pores, microdialysis will exclude proteins from the sample while allowing smaller molecules such as neurotransmitters to pass through. The protein-free samples are then analyzed. Although low limits of detection can be achieved for most small-molecule neurotransmitters without the use of mass spectrometry [8,9], determination of basal ACh levels in rat brain has often been a challenge for neurochemists. However, reversed-phase ion-pair liquid chromatography (LC) coupled with positive-ion electrospray ionization (ESI) tandem mass spectrometry (MS/MS) has been shown to detect ACh with low limit of detection (1.4 fmol) and, thus, to measure this neurotransmitter and related endogenous compounds in rat brain microdialysates [10]. Further improvements in assay performance have been reported recently [11,12].

2.2. Neuropeptides

Numerous challenges have emerged when analyzing samples from brain tissue by chromatographic fractionation followed by ESI-MS/MS and matrix-assisted laser desorption/ionization (MALDI) mass spectrometry [13]. The principal challenge

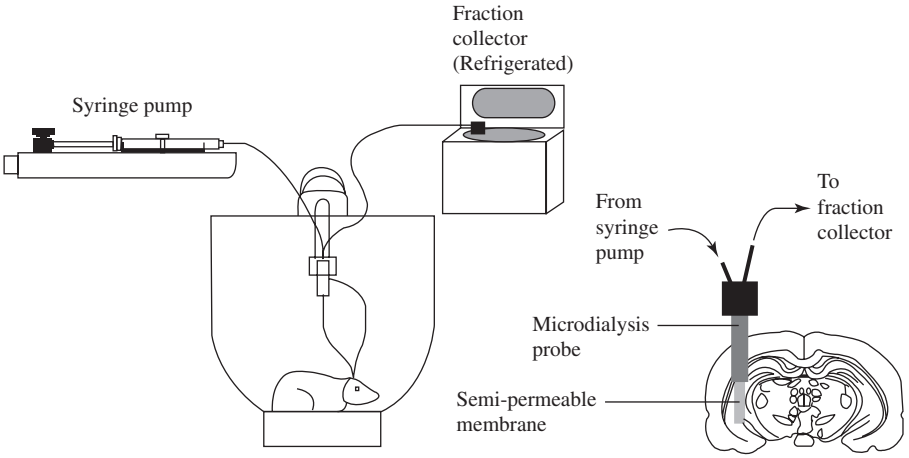


Fig. 1. System to collect microdialysates from the brain of experimental animals (rats).

has been the low levels of neuropeptides in the brain relative to the high levels of peptides that result from postmortem protein degradation. To block such protein degradation, rats and mice can be sacrificed by focused microwave irradiation, which inactivates enzymes in the tissue within seconds and permits the detection of numerous neuropeptides by mass spectrometry [14]. Because microwave devices capable of focused irradiation are not widely available due to their high price, postmortem degradation of proteins can also be reduced very substantially by sacrificing the animals using the standard decapitation method followed by an immediate irradiation of the head in a conventional microwave oven [15].

Alternatively, *in vivo* microdialysis sampling similar to the method shown schematically in Fig. 1 and described in Section 2.1.1 can also be used for neuropeptide discovery or screening. This technique essentially circumvents protein degradation associated with the methods discussed in the previous section. Neuropeptides collected by microdialysis can be preconcentrated and desalted by reversed-phase LC [16], and subsequently supplied directly onto a micro- or nanoflow-LC column for gradient elution and ESI-MS as well as MS/MS analysis [17]. Microdialysis is suitable not only for an *in vivo* sampling of the extracellular space of the brain but also to simultaneously introduce exogenous agents such as neuropeptides into the tissue (which is often called "retrodialysis") to investigate the effect or fate of these agents in the brain. As an example, Fig. 2 demonstrates the combined use of *in vivo* microdialysis and LC/ESI-MS to study kyotorphin-induced Met-enkephalin release in the brain [18]. Briefly, the animals were stereotaxically implanted with guide cannulae that reached the globus pallidus of the brain—a site with high enkephalin-like immunoreactivity. After their recovery from surgery, a microdialysis probe (CMA/12, CMA/Microdialysis, Acton, MA, USA) was inserted into the guide cannula, and the dipeptide Tyr-Arg (kyotorphin) was delivered by "retrodialysis" while microdialysates were collected simultaneously. (The term

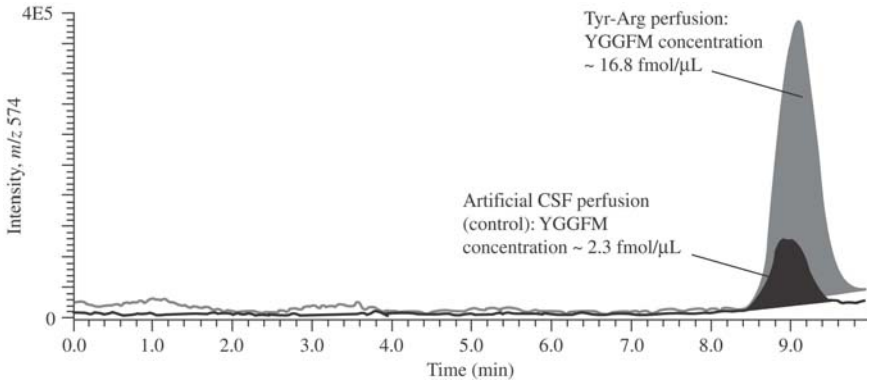


Fig. 2. LC/ESI-MS analysis of the opioid peptide Met-enkephalin (Tyr-Gly-Gly-Phe-Met = YGGFM; SIM, *m/z* 574) in microdialysates collected from the globus pallidus region of the rat brain after the perfusion of the probe (CMA/12) at 2 µL/min with artificial CSF (black trace/area) and with 5 nmol/µL of kyotorphin (Tyr-Arg) in artificial CSF (gray trace/area).

retrodialysis refers to the technique where an agent is dissolved in the perfusion fluid for delivery into the tissue during a microdialysis experiment.) In the control experiment, the probe was only perfused with artificial CSF. The collected microdialysates were analyzed by gradient reversed-phase capillary LC/ESI-MS (conditions given in ref. [19]). On the basis of selected ion monitoring (SIM) chromatograms for protonated Met-enkephalin, *m/z* 574, and external calibration with solutions of known concentration of the peptide, microdialysates collected from the animal that received kyotorphin showed more than sevenfold increase in the concentration of Met-enkephalin compared to that of the control animal.

2.3. Brain proteins (neuroproteomics)

Proteins of the mammalian brain are also of great interest to neuroscientists. A large number of proteins and their complex networks covering diverse biological functions can be studied via the emerging methods of neuroproteomics. Although mapping of all proteins and their intricate interplay in prototype unicellular eukaryotes are being pursued extensively by various techniques, analysis of protein constituents in organelles and specifically isolated subcellular fractions or protein complexes appears to be a viable ("subproteome") approach that reduces complexity and allows for a meaningful application of the technique to brain research [21]. For example, Fig. 3 shows the scheme of a routine procedure to obtain synaptosomes (isolated synapses) and synaptic plasma-membrane fraction by subcellular fractionation of brain tissue [22,23]. Methods to obtain synaptic junctions and, with their further fractionation, presynaptic active zone and postsynaptic density fractions from synaptosomes have been developed [24].

Nevertheless, large numbers of proteins are present in brain-derived samples and they need to be separated for identification and quantification. The separation

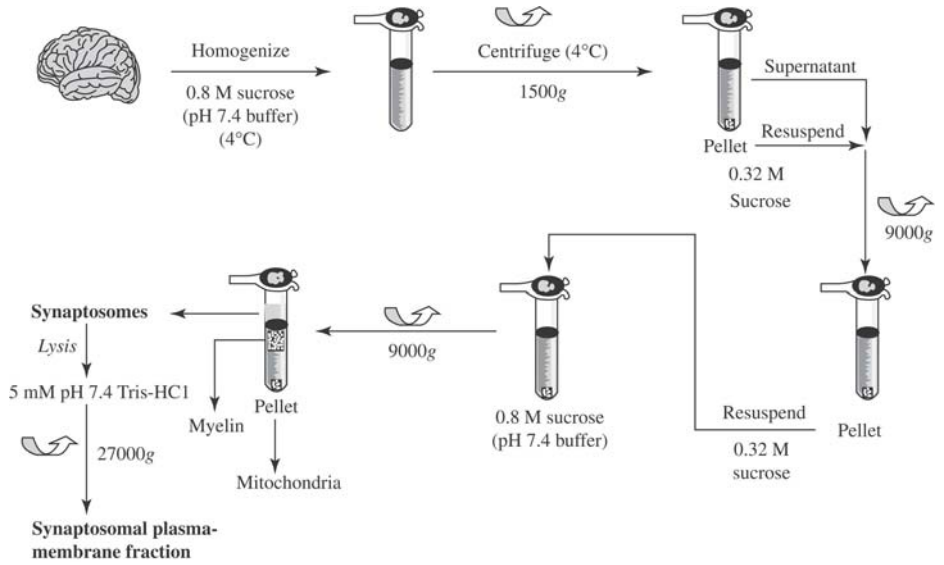


Fig. 3. A scheme of a procedure to obtain synaptosomes [22] (isolated synapses) and synaptic plasma membrane fraction [23] by subcellular fractionation of brain tissue via sucrose gradient.

and visualization of complex protein mixtures are commonly performed by two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). 2D-PAGE followed by in-gel protease (trypsin) digestion, MALDI/time-of-flight (TOF) mass spectrometry, and sequence database searching is the technique most frequently used in today's neuroproteomics studies [25,26]. Although sensitivity and robustness of MALDI-TOF/MS generally allow for the rapid identification of proteins, implementation of ESI is advantageous because ionization selectivity changes may be exploited for peptides present in a proteolytic digest sample [27]. Moreover, detection of a large proportion of the peptide ions can be accomplished by nanoflow-ESI in combination with online LC techniques when signal suppression would have otherwise occurred in MALDI analysis of the peptide mixture. MS/MS used in combination with LC/ESI-MS can also provide sequence tags that greatly reduce the amount of information necessary for an unambiguous match to proteins when using protein database-searching tools. To overcome limitations of the 2D-PAGE (not applicable to hydrophobic proteins, proteins with extreme isoelectric point and low-abundance proteins), one-dimensional sodium dodecylsulfate (SDS)-PAGE may be employed. However, given the complexity of protein samples usually subjected to neuroproteomics studies, comigration of multiple proteins may still result in complex peptide mixtures after the gel is cut and in-gel proteolytic digestions are performed to permit peptide-based protein identification. Therefore, the addition of reverse-phase LC, as a second dimension of separation, combined with online ESI-MS/MS analyses has become popular, when SDS-PAGE is employed. Fig. 4 shows the application of this gel-enhanced LC/MS/MS

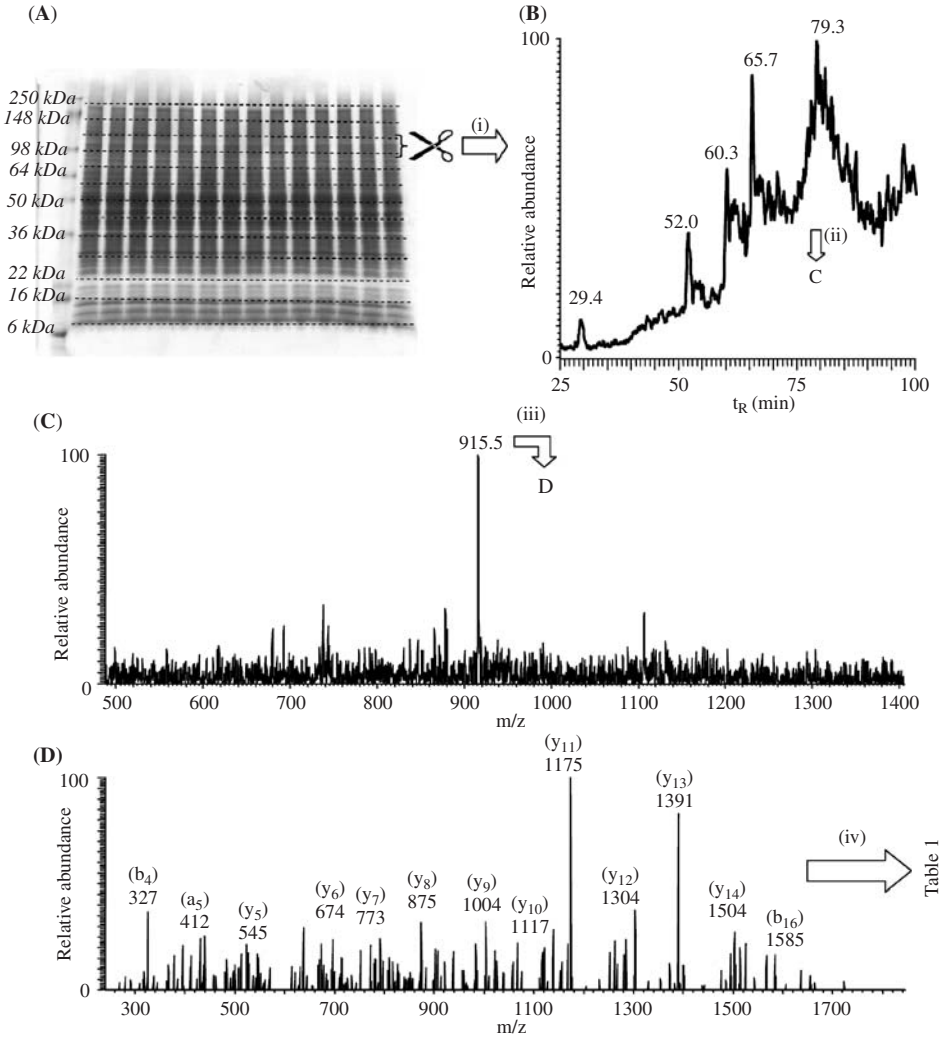


Fig. 4. Illustration of proteomic analysis of the rat synaptic plasma-membrane fraction by combination of SDS-PAGE and gradient reversed-phase LC/ESI-MS/MS (GeLC/MS/MS) on a quadrupole ion-trap instrument [21]. (A) The developed gel is cut into bands and the bands are (i) destained, digested (trypsin), and the sample is desalted for injection into the column. (B) Base-peak chromatogram obtained from the tryptic digest of the 98–120 kDa. Data-dependent acquisition is employed, where in one acquisition cycle (ii) a full-scan mass spectrum is acquired (C), followed by (iii) CID-MS/MS of the most intense ion (*m/z* 915.5) in this mass spectrum (D). (However, MS/MS is not initiated, when ion intensity in the full-scan mass spectrum is below a preset threshold.) Database search (iv) matches the MS/MS to the protein(s) (Table 1, major sequence ions of the peptide are indicated in chart D).

(GeLC/MS/MS) strategy to neuroproteomic analysis—using the synaptic plasma-membrane fraction isolated from rat forebrain according to the procedure in Fig. 3 as an example [21]. The particular tryptic peptide (GVGIISEGNETVEDIAAR; amino acid residues abbreviated with one-letter codes) identified based on the presented MS/MS (Fig. 4D) and the proteins (isoforms of the α-subunit of Na^+/K^+-transporting ATPase in the rat species *Rattus norvegicus*) matched to this particular sequence by database search are listed in Table 1. By conducting a database search based on the entire set of mass spectra and tandem mass spectra recorded from the in-gel tryptic digest of the 98–120 kDa band in the SDS-PAGE (Fig. 4A–B), additional peptide sequences that belonged to these proteins were also found (Table 2). Three out of the six peptides matched specifically to the α_3-isoform of Na^+/K^+-transporting ATPase.

Fractionation of the digested sample by strong cation-exchange chromatography prior to reversed-phase LC/ESI-MS/MS analysis can also be employed, which affords 2D separation without the use of gel electrophoresis [21]. A multidimensional LC technique that integrates strong cation-exchange resin and reversed-phase resin in a biphasic column for coupling with online ESI/MS/MS is also available and serves as the basis of an automated "shotgun" proteomics method called multidimensional protein identification technology (MudPIT) [28].

In quantitative neuroproteomics based on 2D-PAGE, intact proteins are separated and the abundance of a protein is determined by measuring stain intensity of the

Table 1

Peptide identified from CID-MS/MS shown in Fig. 4D (major sequence-related ions of the peptide are indicated in the spectrum) by BioWorks (version 3.2, Thermo Fisher, San Jose, CA, USA) using the National Center for Biotechnology Information (NCBI) nonredundant (nr) protein database with rat (*R. norvegicus*) selected in the species option of the program

Peptide[a]	Protein[b]	$M_{protein}$[c]	Peptide positions[d]
GVGIISEGNETVEDIAAR	ATPase, Na^+/K^+ transporting, α_1 polypeptide (6978543)	113055	630–647
GVGIISEGNETVEDIAAR	ATPase, Na^+/K^+ transporting, α_2 polypeptide (6978545)	112218	627–644
GVGIISEGNETVEDIAAR	Na^+/K^+ ATPase α_3-subunit (6978547)	111737	620–637

[a] Amino acid residues are abbreviated with one-letter codes.

[b] NCBI accession number given in parentheses.

[c] Molecular mass (Da).

[d] Starting and end points in protein's listed sequence (numbering from 1 to *n* in the amino- to carboxy-terminal direction).

Table 2

Tryptic peptides matched by BioWorks (version 3.2) to the Na^+/K^+-transporting ATPase α-subunit in the 98–120 kDa SDS-PAGE band of the rat synaptic plasma-membrane fraction (BioWorks version 3.2) after in-gel digestion and data-dependent LC/ESI-MS/MS

Peptide	Precursor ion (*m*/*z*)	*P*[a]	Protein (NCBI number)
VDNSSLTGESEPQTR	$[M+2H]^{2+}$ (810.3)	1.14×10^{-4}	ATPase, Na^+/K^+ transporting, α[b]
EAFQNAYLELGGLGER	$[M+2H]^{2+}$ (883.9)	3.21×10^{-7}	Na^+/K^+-transporting ATPase α_3-subunit (6978547)
QGAIVAVTGDGVNDSPALK	$[M+2H]^{2+}$ (906.5)	2.16×10^{-9}	ATPase, Na^+/K^+ transporting, α[b]
GVGIISEGNETVEDIAAR	$[M+2H]^{2+}$ (915.5)	4.11×10^{-8}	ATPase, Na^+/K^+ transporting, α[b]
YQLSIHETEDPNDNR	$[M+2H]^{2+}$ (915.9)	1.19×10^{-7}	Na^+/K^+-transporting ATPase α_3-subunit (6978547)
IISAHGCKVDNSSLTGESEPQTR	$[M+3H]^{3+}$ (810.4)	2.58×10^{-4}	Na^+/K^+-transporting ATPase α_3-subunit (6978547)

The NCBI nr protein database was used with *R. norvegicus* selected in the species option of the program.

[a] Probability of an incorrect match (false positive).

[b] Tryptic fragment shared by multiple isoforms of the protein.

protein spot on the gel [29]. The alternative LC–MS/MS-based approach often uses stable-isotope labeling techniques, e.g., long-term metabolic labeling of animals with a diet enriched in heavy nitrogen (^{15}N) [30], or labeling and fractionating the protein sample with isotope-coded affinity tag (ICAT) methodology [31]. The latter techniques afford relative quantification (see description of the ICAT and related methods earlier in this book [32]). For example, ICAT-labeled tryptic peptides of proteins present in the cerebral cortex synaptic plasma-membrane fraction of morphine-naïve and 7-day morphine-treated rats were successfully identified using capillary LC-ESI/MS/MS in conjunction with protein database searching [33]. Fig. 5 shows a pair of ions at *m*/*z* 737.3 and 741.2 with a significant difference in ion intensities, when averaged over the 40- to 42-min retention-time window of the gradient reversed-phase LC separation. The $\Delta = 4$ Th difference indicated that they

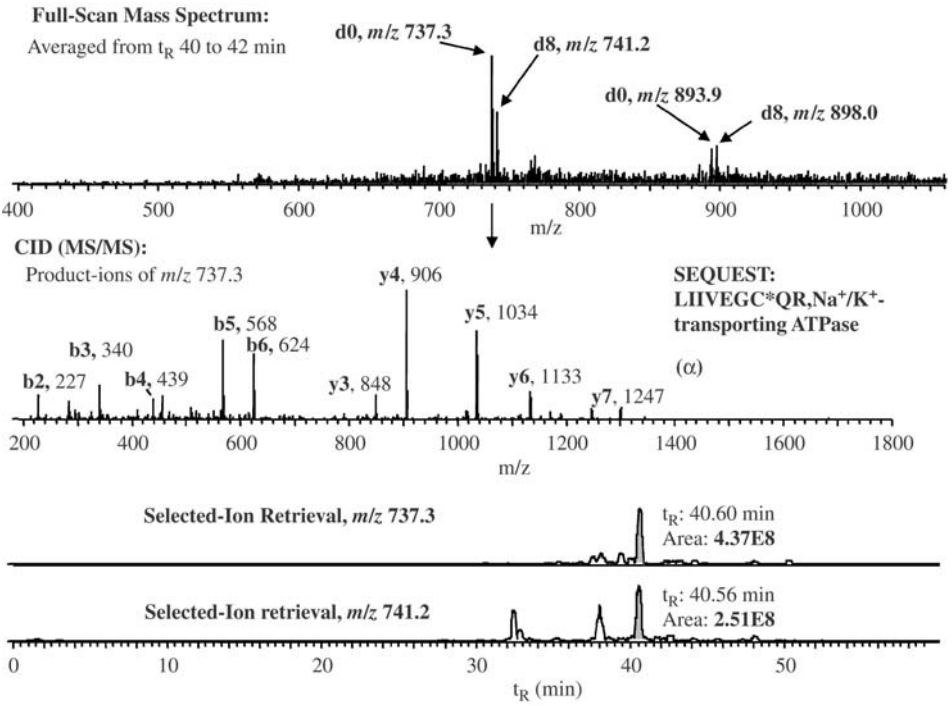


Fig. 5. ICAT-labeled tryptic peptides [*m/z* 737.3 and 741.2 (doubly charged positive ions)] from synaptic plasma-membrane fraction detected by gradient reversed-phase LC/ESI-MS and MS/MS data-dependent acquisition in the retention-time range of 40–42 min. Following protein database search, the sequence was determined to be LIIVEGC*QR, an ICAT-labeled tryptic peptide of Na^+/K^+-transporting ATPase α-subunit (asterisk indicates the ICAT label on cysteine, C). The matched tryptic peptide of the protein is not isoform specific (refer to text in this chapter explaining results from experiments summarized in Fig. 4 and Tables 1 and 2). Upon calculation of the peak area ratios (d8/d0) obtained from the selected ion chromatograms, this particular protein was found to be present in 42% lower abundance in the rat subjected to chronic morphine exposure (d8 label), compared to its morphine-naïve control (d0 label). An additional ICAT pair [*m/z* 893.9 and 898.0 (doubly charged positive ions)] in this retention-time range belongs to the tryptic peptide YQVDPDAC*FSAK of the voltage-dependent anion channel 1 (NCBI accession number 6755963), which did not show difference in protein abundance between the morphine-treated and morphine-naïve animals. (Reproduced with permission from: Prokai, L., Zharikova, A.D. and Stevens Jr., S.M., *J. Mass Spectrom.*, **40**, 169–175 (2005). ©2005 Copyright John Wiley & Sons Limited.)

were doubly charged ($[M+2H]^{2+}$) ions showing the attributes of a successful labeling with the light (nondeuterated, d0) and heavy (octadeuterated, d8) affinity labels. Upon profiling these molecular ions by selected-ion retrieval, the area under the peaks showed a decrease of 43% in the abundance of the corresponding protein in the cortical synaptic membrane of one animal that received chronic subcutaneous morphine administration. Again, MS/MS product-ion spectra obtained

through collision-induced dissociation (CID) and data-dependent acquisition method in conjunction with LC/ESI-MS provided sequence tag for the positive protein identification. Tryptic fragment LIIVEGCQR (with the appropriate modification by the biotin-carrying tag at Cys, C; amino acid residues were abbreviated with the one-letter symbols) matched unequivocally to the Na^+/K^+ ATPase α-subunit of the rat. After performing the ICAT experiment in triplicate, the abundance of this integral membrane protein was found to decrease by 39 ± 2% after exposure of the rat to morphine for 7 days.

3. Discussion

3.1. Neurotransmitters

The measurement of neurotransmitters in brain tissue and extracellular fluid has been used to develop diagnosis and effective treatment strategies for neuropsychiatric and neurodegenerative diseases. For example, ACh is associated with learning and memory [34], and its involvement is well recognized in several dysfunctions of the brain such as Alzheimer's disease [35], Parkinson's disease [36], and dementia [37]. The measurement of ACh in the brain is important in animal models of these conditions that either disrupt the production of ACh or stimulates the overproduction of acetylcholinesterase, an enzyme that catalyzes the rapid hydrolysis of ACh to choline once it has performed its function. Because of this rapid hydrolysis, ACh levels become meaningless when the animal is killed and decapitated, and then the brain is removed for the extraction of the analyte. Therefore, *in vivo* microdialysis [7] has been the method of choice for obtaining samples to reflect the extracellular ACh concentrations in the brain. Because of the low physiological ACh levels, many investigators have had to use an appropriate enzyme inhibitor (e.g., neostigmine added to the perfusion medium of the probe) to limit degradation of the neurotransmitter by acetylcholinesterase. ACh could then be determined from the collected microdialysates by, e.g., LC with electrochemical detection [38]. Esterase inhibitors may, however, affect physiology of the system and, thus, may interfere with data interpretation [39]. The use of LC/ESI-MS/MS furnishes low limits of detection and is therefore suitable for the measurement of ACh in rat brain microdialysates without the use of an acetylcholinesterase inhibitor [10].

3.2. Neuropeptides

Peptides perform many important functions in the CNS as neurotransmitters, neuromodulators, or neurohormones [40]. Neuropeptides are involved in a wide variety

of systems including pain, memory, reproduction, reward mechanisms, food and water intake, circadian rhythms, and many others. The extent of those systems that are crucially affected by neuropeptides is vast and their interactions are often very complex. Over 100 mammalian neuropeptides have been found and many have been postulated but remained to be isolated and identified. The etiology of numerous brain maladies involves neuropeptides through their hypo- or hypersecretion, alterations in storage, release, catabolism, and modifications by posttranslational processing. For example, the release of dynorphin A may be associated with spontaneous pain according to a recent study in a mouse model for neuropathic cancer pain [41]. Beyond their role in CNS physiology, neuropeptides are therefore considered key sources of drug discovery, diagnostics, and therapeutics. The construction of a neuropeptidome, a fact database for endogenous neuropeptides, has been proposed to aid these efforts [13]. However, low tissue levels of neuropeptides, their limited biostability, and degradation background from brain proteins are obstacles for the practice of neuropeptidomics. Therefore, appropriate measures, such as the use of microwave irradiation that inactivates enzymes in the tissue within seconds, are required before one obtains neuropeptide fractions for mass spectrometric analysis [14,15]. Alternatively, microdialysis can be used for *in vivo* sampling of neuropeptides from the extracellular fluid of the brain, which removes the target analytes from the tissue and, thus, they escape enzymatic degradation by neuropeptidases. In addition to the exploration of the neuropeptidome, *in vivo* microdialysis combined with mass spectrometry can be used for probing the effect of various compounds (conveniently introduced by retrodialysis) on the secretion of selected neuropeptides [18], to study neuropeptide metabolism [42,43], and for other related *in vivo* experiments involving neuropeptides [44] to help neuroscientists understand brain physiology or pathology and propose new methods of medical diagnosis, prognosis, and treatment. For example, kyotorphin is a neuropeptide physiologically synthesized in the brain by a specific enzyme, kyotorphin synthetase [45], from L-Tyr, L-Arg, and ATP in the presence of Mg^{2+}. In a clinical trial, administration of L-Arg solution has shown potential benefits for treating various pain conditions due to the presumed kyotorphin synthesis in the brain and spinal cord [46]. Kyotorphin produces opioid analgesia indirectly via the release of Met-enkephalin—based on *in vitro* studies [47]. The use of excised tissue and analytical techniques with poor molecular specificity may be of concern regarding the validity of findings in these experiments *in vivo*. Through the combined use of brain microdialysis in a living animal and LC/ESI-MS as a high-specificity assay method, the study summarized in Fig. 2 has provided unequivocal evidence that kyotorphin functions as a Met-enkephalin releaser. This *in vivo* testing method should also be very valuable, e.g., during the development of novel, brain-targeted kyotorphin analogs [48] intended to overcome shortcomings of the endogenous neuropeptide as a drug candidate such as poor blood–brain barrier penetration and inadequate biostability [49].

3.3. Neuroproteomics

Exploration of the brain by the methods of proteomics has apparent advantages. Simply identifying the presence of proteins in key compartments within neurons and glia will provide an essential framework for understanding their function [20]. One of the distinct features of neuroproteomic analysis, which is not attainable with RNA expression data, is the ability to fractionate brain proteins into various subpopulations [29]. Nearly one-third of the proteome is believed to consist of integral membrane proteins of biological membrane bilayers that compartmentalize living cells and have been identified as important drug targets [50]. Because by far the most attention has been historically focused on the electrical properties of neurons and their connections at synapses, focus on synaptic membrane-protein fractions (Fig. 3) is well justified. By using multidimensional proteomics approaches that employ complementary ionization and mass spectrometric methods in combination with orthogonal separation techniques, synaptic plasma-membrane proteins from rat forebrain can be successfully identified [21]. Given the complexity of the sample, the number of proteins identified is affected by the number of gel-based or chromatographic separation stages performed before mass spectrometric analysis. The advantages of SDS-PAGE as the first dimension of separation have been that the method is simple and widely used by neuroscientists, it can be applied to membrane proteins, and samples can be separated side-by-side and stained/destained simultaneously in order to increase the amount of low-abundance proteins available for subsequent analysis. Routine MALDI-TOF/MS is limited to less complex digest mixtures due to ion suppression effects, and reliable protein identification requires the detection of several tryptic peptides from each protein with the moderate mass accuracy of the method ($<$20 ppm). However, the use of GeLC/MS/MS has allowed for unambiguous identification of tryptic peptides and corresponding proteins via fragment-ion tag database searching, as demonstrated in Fig. 4.

Proteomes of cells are dynamic and are directly affected by environmental factors such as stress, aging, diseases, and drug treatment. Many changes in synaptic activity following physiological variations induced by such factors can be explained by determining protein-expression-level differences between control (“healthy”) and perturbed physiological states. Since protein expression analysis allows us to identify those protein(s) actively involved, for example, in the progression of a particular disease, strategies for molecular intervention can now be formulated with further knowledge of the biological pathways associated with the disease. Similarly, by employing quantitative protein analysis methods, the effect(s) of existing drugs or potential drug candidates can be thoroughly characterized and compared to allow for a detailed interrogation of the mechanism of action. Conventional protein expression methods involving immunohistochemical techniques are generally limited to proteins specific to a particular antibody. Quantitation by 2D-PAGE is usually inadequate for low-abundance and/or membrane proteins present in the sample of

interest. Therefore, the method of choice in proteomics has been shifting to stable-isotope labeling or tagging followed by mass spectrometry, when accurate quantification is desired [51]. The value of these methods for neuroproteomics has been particularly well demonstrated by the study focusing on the effect of chronic morphine exposure on the proteins of the synaptic plasma-membrane fraction in rats, where the ICAT strategy coupled with nanoflow reversed-phase LC/ESI-MS and MS/MS was employed [33]. Out of the 80 proteins covered simultaneously by the method in a single assay without protein-specific antibodies, the expression of several important synaptic plasma-membrane proteins has been shown to change significantly as a result of chronic morphine exposure *in vivo*. The underlying mechanisms and their biological significance are yet to be elucidated for the majority of such changes measured. However, the downregulation of Na^+/K^+ ATPase pumps (Fig. 5) can be implicated in the neurobiology of opioid tolerance and dependence. These very important synaptic membrane proteins maintain an ionic concentration difference (electric potential) across the membrane that allows for the propagation of electrical impulses along nerve cells and across synaptic clefts between nerve cells [52]. Their downregulation causes a decrease in electrogenic Na^+/K^+ pumping, which would explain the observed subsensitivity of neurons to opiates upon developing non-specific (heterologous) tolerance to these drugs [53]. Therefore, continued applications of neuroproteomics are likely to contribute to a better understanding of the mechanisms involved in maladies affecting the CNS such as morphine withdrawal, dependence, and tolerance. The results of the studies may also lead to the development of new strategies for the management of neurological diseases.

4. Future trends

Studies reported on the application of *in vivo* microdialysis and mass spectrometry to brain research have employed rats as experimental animals. Although LC/MS/MS methods provide adequate sensitivity, selectivity, and precision to allow for measurements of small-molecule neurotransmitters in brain tissue and microdialysis samples, there is an increasing demand to minimize sample volume and improve throughput and robustness. The driving force of this demand is in part the proliferation of genetically altered (mutant, overexpressed, and knockout) mice models created to understand the CNS and model its diseases. Over the last several years, cerebral microdialysis has also become one of the new methods established in state-of-the-art brain monitoring upon neurointensive care [54]. Further refinement of sampling and assay technologies both at the preclinical level and in neurointensive medicine will provide enormous potential for revealing the role of small-molecule neurotransmitters in normal and pathological processes [3].

Neuropeptidomics and neuroproteomics are in their infancy. Improved methods to obtain neuropeptide-rich fractions from the brain for subsequent interrogations by mass spectrometry are needed [55], along with the development of

neuropeptide-directed analytical separation, mass spectrometric data acquisition, and processing methods. In the field of neuroproteomics, one of the disadvantages of gel-based approaches has been that they are difficult to automate. Therefore, other multidimensional separation methods [28] are expected to gain increased acceptance for the identification of membrane and/or important but low-abundance proteins in complex samples from the mammalian brain. Information on the presence of proteins in various subcellular fractions allows for the design of studies probing their specific functions in health and disease. Explorations of posttranslational modifications in the brain proteome, which was not covered in this chapter, have also been promising [56,57]. Quantitative proteomics studies are likely to contribute to a better understanding of diseases affecting the CNS. There are vast areas of translational investigations expected to reveal potential biomarkers for and correlate proteins with brain disorders [29], which may also lead to the development of new strategies to manage these conditions in humans.

5. Conclusions

This chapter covered selected applications of mass spectrometry and highlighted its power to support diverse studies focused on the mammalian brain. Much remains to be developed in methodology before mass spectrometry becomes a widely accepted and a routinely employed technique in the neurosciences and impacts the diagnosis, prognosis, and therapy of neurodegenerative/neuropsychiatric diseases. However, progress has been steady, which clearly warrants continued exploration and development of methods for these applications.

Acknowledgments

The author thanks Drs. Alevtina D. Zharikova, Stanley M. Stevens, Jr., and Ho-Seung Kim for contributing to the work described in this chapter, and Drs. Jerry W. Simecka and Katalin Prokai-Tatrai for their helpful comments on the manuscript. The studies on neuropeptides benefited from the author's collaboration with Drs. John Roboz and William J. Simonsick, Jr. The research presented here has been supported by grants (MH59360, NS44765, AG025384 and RR12023) from the National Institutes of Health (Bethesda, MD, USA) and the University of Florida Research Foundation. Laszlo Prokai is currently the Robert A. Welch Professor at the University of North Texas Health Science Center.

References

1. Wild, G.C. and Benzel, E.C., Central nervous system cell types. In: *Essentials of Neurochemistry*. Jones and Bartlett, Boston, 1994, pp. 3–15 (Chapter 1, ISBN: 0-86720-862-7).
2. Wild, G.C. and Benzel, E.C., The synapse. In: *Essentials of Neurochemistry*. Jones and Bartlett, Boston, 1994, pp. 33–43 (Chapter 3, ISBN: 0-86720-862-7).

3. Shah, A.J., Crespi, F. and Heidbreder, C., Amino acid neurotransmitters: separation approaches and diagnostic value. *J. Chromatogr. B*, **781**, 151–163 (2002).
4. Cheng, F.C. and Kuo, J.S., High-performance liquid-chromatographic analysis with electrochemical detection of biogenic-amines using microbore columns. *J. Chromatogr. B*, **665**, 1–13 (1995).
5. Powell, P.R. and Ewing, A.G., Recent advances in the application of capillary electrophoresis to neuroscience. *Anal. Bioanal. Chem.*, **382**, 581–591 (2005).
6. Bartholini, G., Stadler, H., Ciria, M.G. and Lloyd, K.G., Use of push-pull cannula to estimate dynamics of acetylcholine and catecholamines within various brain areas. *Neuropharmacology*, **15**, 515–519 (1976).
7. Ungerstedt, U., Herrera-Marschitz, M., Jungnelius, U., Ståhle, L., Tossman, U. and Zetterström, T., Dopamine synaptic mechanisms reflected in studies combining behavioral recordings and brain dialysis. *Adv. Biosci.*, **37**, 219–231 (1982).
8. Smolders, I., Sarre, S., Michotte, Y. and Ebinger, G., The analysis of excitatory, inhibitory and other amino-acids in rat-brain microdialysates using microbore liquid-chromatography. *J. Neurosci. Meth.*, **57**, 47–53 (1995).
9. Parrot, S., Bert, L., Mouly-Badina, L., Sauvinet, V., Colussi-Mas, J., Lambas-Senas, L., Robert, F., Bouilloux, J.P., Suaud-Chagny, M.F., Denoroy, L. and Renaud B., Microdialysis monitoring of catecholamines and excitatory amino acids in the rat and mouse brain: recent developments based on capillary electrophoresis with laser-induced fluorescence detection—a mini-review. *Mol. Cell. Neurobiol.*, **23**, 793–804 (2003).
10. Zhu, J., Wong, P.S.H., Cregor, M., Gitzen, J.F., Coury, L.A. and Kissinger, P.T., *In vivo* microdialysis and reverse phase ion pair liquid chromatography/tandem mass spectrometry for the determination and identification of acetylcholine and related compounds in rat brain. *Rapid Commun. Mass Spectrom.*, **14**, 1695–1700 (2000).
11. Guo, Y., Analysis of quaternary amine compounds by hydrophilic interaction chromatography/mass spectrometry (HILIC/MS). *J. Liq. Chromatogr. Rel. Tech.*, **28**, 497–512 (2005).
12. Uutela, P., Reinila, R., Piepponen, P., Ketola, R.A. and Kostiainen, R., Analysis of acetylcholine and choline in microdialysis samples by liquid chromatography/tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **19**, 2950–2956 (2005).
13. Minamino, N., Tanaka, J., Kuwahara, H., Kihara, T., Satomi, Y., Matsubae, M. and Takao, T., Determination of endogenous peptides in the porcine brain: possible construction of Peptidome, a fact database for endogenous peptides. *J. Chromatogr. B*, **792**, 33–48 (2003).
14. Svensson, M., Skold, K., Svenningsson, P. and Andrén, P.E., Peptidomic-based discovery of novel neuropeptides. *J. Proteome Res.*, **2**, 213–219 (2003).
15. Che, F.Y., Lim, J., Pan, H., Biswas, R. and Fricker, L.D., Quantitative neuropeptidomics of microwave-irradiated mouse brain and pituitary. *Mol. Cell. Proteomics*, **4**, 1391–1405 (2005).
16. Emmett, M.R., Andren, P.E. and Caprioli, R.M., Specific molecular mass detection of endogenously released neuropeptides using *in vivo* microdialysis mass spectrometry. *J. Neurosci. Meth.*, **62**, 141–147 (1995).
17. Haskins, W.E., Watson, C.J., Cellar, N.A., Powell, D.H. and Kennedy, R.T., Discovery and neurochemical screening of peptides in brain extracellular fluid by chemical analysis of *in vivo* microdialysis samples. *Anal. Chem.*, **76**, 5523–5533 (2004).
18. Prokai, L., Roboz, J., Kim, H.-S., Ma, L. and Wang, T., *In vivo* microdialysis and LC/ESI-MS studies on kyotorphin-induced Met-enkephalin release in the brain. *Proceedings of the 44th ASMS Conference on Mass Spectrometry and Allied Topics,* Santa Fe. American Society for Mass Spectrometry, 1996, p. 945.
19. Roboz, J., Yo, Q., Mang, A. and Van Soest, R., On-line buffer removal and fraction selection in gradient capillary high-performance liquid-chromatography prior to electrospray mass-spectrometry of peptides and proteins. *Rapid Commun. Mass Spectrom.*, **8**, 621–626 (1994).

20. Grant, S.G.N. and Blackstock, W.P., Proteomics in neuroscience: from protein to network. *J. Neurosci.*, **21**, 8315–8318 (2001).
21. Stevens Jr., S.M., Zharikova, A.D. and Prokai, L., Proteomic analysis of the synaptic plasma membrane fraction isolated from rat forebrain. *Mol. Brain Res.*, **117**, 116–128 (2003).
22. Hajos, F., Improved method for preparation of synaptosomal fractions of high purity. *Brain Res.*, **93**, 485–489 (1975).
23. Kanner, B.I., Active-transport of γ-aminobutyric acid by membrane-vesicles isolated from rat brain. *Biochemistry*, **17**, 1207–1211 (1978).
24. Phillips, G.R., Huang, J.K., Wang, Y., Tanaka, H., Shapiro, L., Zhang, W.D., Shan, W.S., Arndt, K., Frank, M., Gordon, R.E., Gawinovicz, M.A., Zhao, Y.M. and Colman, D.R., The presynaptic particle web: ultrastructure, composition, dissolution, and reconstitution. *Neuron*, **32**, 63–77 (2001).
25. Fountoulakis, M., Schuller, E., Hardmeier, R., Berndt, P. and Lubec, G., Rat brain proteins: two-dimensional protein database and variations in the expression level. *Electrophoresis*, **20**, 3572–3579 (1999).
26. Yuan, X.L., Russell, T. and Wood, G., Analysis of the human lumbar cerebrospinal fluid proteome. *Electrophoresis*, **23**, 1185–1196 (2002).
27. Stevens Jr., S.M., Kem, W.R. and Prokai, L., Investigation of cytolysin variants by peptide mapping: enhanced protein characterization using complementary ionization and mass spectrometric techniques. *Rapid Commun. Mass Spectrom.*, **16**, 2094–2101 (2002).
28. Wolters, D.A., Washburn, M.P. and Yates, J.R. III, An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.*, **73**, 5683–5690 (2001).
29. Kim, S.I., Voshol, H., van Oostrum, J., Hastings, T.G., Cascio, M. and Glucksman, M.J., Neuroproteomics: expression profiling of the brain's proteomes in health and disease. *Neurochem. Res.*, **29**, 1317–1331 (2004).
30. Wu, C.C., MacCoss, M.J., Howell, K.E., Mathews, D.E. and Yates, J.R. III, Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal. Chem.*, **76**, 4951–4959 (2005).
31. Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotech.*, **17**, 994–999 (1999).
32. Vertes, A., Mass spectrometry in proteomics. In: Vékey, K., Telekes, A. and Vertes, A. (Eds.), *Medical Applications of Mass Spectrometry*. Elsevier, Amsterdam, 2008 (Chapter 8, section 2.6, this book).
33. Prokai, L., Zharikova, A.D. and Stevens Jr., S.M., Effect of chronic morphine exposure on the synaptic plasma-membrane subproteome of rats: a quantitative protein profiling study based on isotope-coded affinity tags and liquid chromatography/mass spectrometry. *J. Mass Spectrom.*, **40**, 169–175 (2005).
34. Hasselmo, M.E. and Bower, J.M., Acetylcholine and memory. *Trends Neurosci.*, **16**, 218–222 (1993).
35. Giacobini, E., Cholinergic function and Alzheimer's disease. *Int. J. Geriatr. Psychiatry*, **18**, S1–S5 (2003).
36. Bosboom, J.L.W., Stoffers, D. and Wolters, E.C., The role of acetylcholine and dopamine in dementia and psychosis in Parkinson's disease. *J. Neur. Transm., Suppl.*, **65**, 185–195 (2003).
37. Lane, R.M., Potkin, S.G. and Enz, A., Targeting acetylcholinesterase and butyrylcholinesterase in dementia. *Int. J. Neurophsychopharm.*, **9**, 101–124 (2006).
38. Prokai, L. and Zharikova, A.D., Neuropharmacodynamic evaluation of a centrally active thyrotropin-releasing hormone analogue [Leu^2]TRH and its chemical brain-targeting system. *Brain Res.*, **952**, 268–274 (2002).

39. Ichikawa, J., Dai, J. and Meltzer, H.Y., Acetylcholinesterase inhibitors are neither necessary nor desirable for microdialysis studies of brain acetylcholine. *Curr. Sep.*, **19**, 37–43 (2000).
40. Strand, F.L., Neuropeptides: general characteristics and neuropharmaceutical potential in treating CNS disorders. In: Prokai, L. and Prokai-Tatrai, K. (Eds.), *Peptide Transport and Delivery into the Central Nervous System, Progress of Drug Research* (Vol. 61). Birkhäuser, Basel, 2003, pp. 1–37 (Chapter 1, ISBN 3-7643-6716-4).
41. Shimoyama, M., Tatsuoka, H., Ohtori, S., Tanaka, K. and Shimoyama, N., Change of dorsal horn neurochemistry in a mouse model of neuropathic cancer pain. *Pain*, **114**, 221–230 (2005).
42. Andrén, P.E. and Caprioli, R.M., *In vivo* metabolism of substance-P in rat striatum utilizing *in vivo* microdialysis liquid chromatography—micro-electrospray mass spectrometry. *J. Mass Spectrom.*, **30**, 814–824 (1995).
43. Prokai, L., Kim, H.-S., Zharikova, A., Roboz, J., Ma, L., Deng, L. and Simonsick Jr., W.J., Electrospray ionization mass spectrometric and LC/MS studies on the metabolism of synthetic dynorphin A peptides in brain tissue *in vitro* and *in vivo*. *J. Chromatogr. A*, **800**, 59–68 (1998).
44. Klintenberg, R. and Andrén, P.E., Altered extracellular striatal *in vivo* biotransformation of the opioid neuropeptide dynorphin A(1–17) in the unilateral 6-OHDA rat model of Parkinson's disease. *J. Mass Spectrom.*, **40**, 261–270 (2005).
45. Ueda, H., Yoshihara, Y., Fukushima, N., Shiomi, H., Nakamura, A. and Takagi, H., Kyotorphin (tyrosine-arginine) synthetase in rat-brain synaptosomes. *J. Biol. Chem.*, **262**, 8165–8173 (1987).
46. Takagi, H., Harima, A. and Shimizu, H., A novel clinical treatment of persistent pain with L-arginine. *Eur. J. Pharmacol.*, **183**, 1443–1443 (1990).
47. Takagi, H., Shiomi, H., Ueda, H. and Amano, H., Novel analgesic dipeptide from bovine brain is a possible Met-enkephalin releaser. *Nature*, **282**, 410–412 (1979).
48. Chen, P., Bodor, N., Wu, W.-M. and Prokai, L., Strategies to target kyotorphin analogues to the brain. *J. Med. Chem.*, **41**, 3773–3781 (1998).
49. Prokai, L., Peptide drug delivery into the central nervous system. In: Jucker, E.M. (Ed.), *Progress in Drug Research* (Vol. 51). Birkhäuser, Basel, 1998, pp. 95–131 (ISBN 376435822X).
50. Cascio, M. and Rapaka, R.S., Structural biology and structural genomics/proteomics. *J. Pept. Res.*, **60**, 307–311 (2002).
51. Abersold, R. and Mann, M., Mass spectrometry-based proteomics. *Nature*, **422**, 198–207 (2003).
52. Boyer, R.F., 2006 www.wiley.com/legacy/college/boyer/0470003790/animations/membrane_transport/membrane_transport.swf.
53. Biser, P.S., Thayne, K.A., Fleming, W.W. and Taylor, D.A., Na^+, K^+ ATPase alpha-subunit isoform distribution and abundance in guinea-pig longitudinal muscle/myenteric plexus after exposure to morphine. *Brain Res.*, **931**, 186–193 (2002).
54. Ungerstedt, U. and Rostami, E., Microdialysis in neurointensive care. *Curr. Pharm. Des.*, **10**, 2145–2152 (2004).
55. Pettersson, A., Amirkhani, A., Arvidsson, B., Markides, K. and Bergquist, J., A feasibility study of solid supported enhanced microdialysis. *Anal. Chem.*, **76**, 1678–1682 (2004).
56. Kim, S.Y., Chudapongse, N., Lee, S.M., Levin, M.C., Oh, J.T., Park, H.J. and Ho, I.K., Proteomic analysis of phosphotyrosyl proteins in morphine-dependent rat brains. *Mol. Brain Res.*, **113**, 58–70 (2005).
57. Choi, J., Forster, M.J., McDonald, S.R., Weintraub, S.T., Carroll, C.A. and Gracy, R.W., Proteomic identification of specific oxidized proteins in ApoE-knockout mice: relevance to Alzheimer's disease. *Free Rad. Biol. Med.*, **36**, 1155–1162 (2004).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 19

The human pituitary proteome: clinical applications

XIANQUAN ZHAN[a,b], HAROLD SACKS[c], and DOMINIC M. DESIDERIO[a,b,*]

[a]*Charles B. Stout Neuroscience Mass Spectrometry Laboratory, University of Tennessee Health Science Center, 847 Monroe Avenue, Memphis, TN 38163, USA*
[b]*Department of Neurology, University of Tennessee Health Science Center, 847 Monroe Avenue, Memphis, TN 38163, USA*
[c]*Division of Endocrinology, University of Tennessee Health Science Center, 847 Monroe Avenue, Memphis, TN 38163, USA*

*Corresponding author. Tel.: 901-448-5488; Fax: 901-448-7842. E-mail: ddesiderio@utmem.edu (D.M. Desiderio).

1. Introduction

The pituitary is the master regulatory gland, is the most protected organ in the body, and secretes several different pituitary hormones that regulate important hypothalamic–pituitary–target organ axes in the body. Any defect in those regulatory systems has an associated pathology. A pituitary adenoma is a common pathological change of the pituitary, and is a common and critical endocrine tumor. The alteration of protein composition is a crucial factor in the pathogenesis of pituitary adenomas. Mass spectrometry (MS)-based proteomics plays an important role to clarify those protein alterations, to elucidate the basic molecular mechanisms in the formation of a pituitary adenoma, and to detect tumor-specific proteins and potential biomarkers.

1.1. Proteomics, functional proteomics, and comparative proteomics

Proteomics is an important component of functional genomics. In contrast to structural genomics, which studies the human genome sequence [1], functional genomics focuses on two levels—mRNA and protein, which play important roles in an understanding of the regulation of biological systems. The genome, transcriptome, and proteome are highly complementary systems, and correspond, respectively, to the DNA, mRNA, and protein in a cell, tissue, or organ. Fig. 1 demonstrates the relationships among a gene, the genome, and genomics; mRNA, the transcriptome, and transcriptomics; and a protein, the proteome, and proteomics. For a given organism, organ, tissue, or cell, the genome is relatively stable, whereas the transcriptome and proteome are dynamic, and change with time and conditions (for example, different psychological stages, disease states, and experimental conditions) [2,3]. Because the proteome is much more complex than the transcriptome due to the diversity in the efficiencies of translation, a very wide array exists for post-translation modifications (PTMs), protein translocations, protein interactions (with DNA, ligands, and other proteins), protein regulation, etc., in the processes that lead from mRNA to protein [4,5]. Proteomics directly reveals those important protein compositions and modifications that are associated with a given condition.

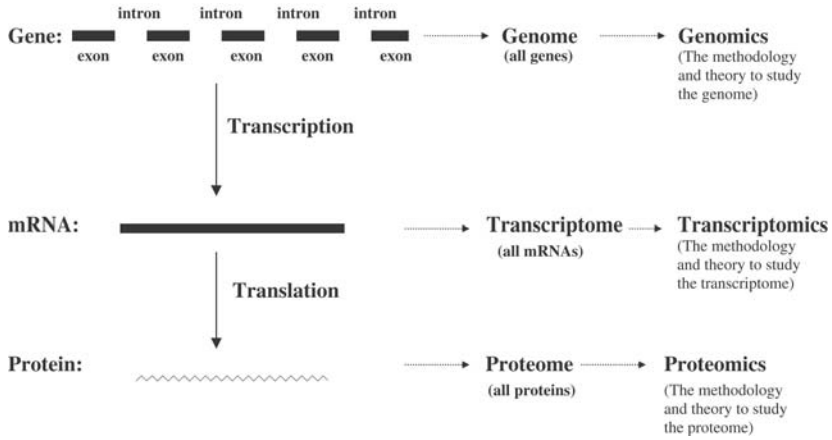


Fig. 1. A flowchart to demonstrate the relationship among the genome, transcriptome, and proteome. An intron is a nucleic acid sequence that does not code any protein, and an exon is a nucleic acid sequence that codes protein. mRNA is "reverse-transcripted" by the enzyme reverse transcriptase into complementary DNA (cDNA). An mRNA sequence has a one-to-one correspondence to a cDNA sequence. cDNA only includes the sequence of exons, and not introns. Reproduced from Zhan and Desiderio [14], with permission from Wiley-VCH, copyright 2005.

The theoretical objective of proteomics is to array and study "all" proteins in a proteome (the full complement of proteins produced by a particular genome), and to provide a systematic and detailed analysis of the protein population in a whole organism, organ, tissue, cell, or subcellular compartment. Actually, that goal is virtually impossible to achieve because of the many different experimental factors and the complex physicochemical nature of proteins. The current experimental systems probably access only ca. 10% of the proteome. An important goal of proteomics is to understand the cellular function at the protein level by means of the dynamic proteome under any given condition; that goal is functional proteomics [6,7]. Functional proteomics will focus only on those proteins that are associated with a unique condition—for example, a disease, different development stages, different pathology stages, different drug-treated conditions, etc. The bulk of the differentially expressed proteins (DEPs) that are associated with a unique condition are usually detected by comparative proteomics. The technology employed in comparative proteomics includes gel-based comparative proteomics and stable isotope-labeling quantitative proteomics. Gel-based comparative proteomics commonly includes two-dimensional gel electrophoresis (2DGE), 2D gel image analysis, MS characterization of proteins, and bioinformatics. Stable isotope-labeling quantitative proteomics usually includes the stable isotope [for example, isotope-coded affinity tag (ICAT)] labeling of a sample, liquid chromatography (LC), tandem mass spectrometry (MS/MS), quantification of the separated peptides, and bioinformatics.

1.2. The pathophysiological basis of pituitary adenoma comparative proteomics

The human pituitary anterior gland includes five highly differentiated cell types that originate from the neural epithelium—a primordial cell in early embryogenesis. Each cell type produces a specific hormone that participates in multiple regulatory hypothalamic–anterior pituitary–target organ axes (Fig. 2) [8]. Those five systems perform a range of very important physiological functions in the human body. Those five types of cells that constitute the anterior lobe of the pituitary gland are the corticotrophs (secrete ACTH), somatotrophs (secrete growth hormone (GH)), lactotrophs (secrete PRL), thyrotrophs (secrete TSH), and gonadotrophs (secrete FSH and LH).

Pituitary tumors could arise from any one of those different cell types, and the tumor's secretion products depend on the cell of origin (Fig. 2). Mixed tumors (the co-secretion of GH with PRL, TSH, or ACTH) may also arise from a single cell type. Molecular genetic studies have also indicated that a pituitary adenoma is monoclonal in origin [9,10]. ACTH oversecretion results in Cushing's disease, with features of hypercortisolism; GH hypersecretion leads to acral overgrowth and a metabolic dysfunction associated with acromegaly; and PRL oversecretion leads to gonadal failure, secondary infertility, and galactorrhea. More rarely, TSH hypersecretion leads to hyperthyroxinemia and goiter, and hypersecreted GH (or its respective protein subunits) leads to gonadal dysfunction. In contrast, tumors that arise from gonadotroph cells do not efficiently secrete their gene products, and they are usually clinically silent [8].

The formation of a pituitary adenoma is thought to be due to either a constant supply of hypothalamic-releasing hormones within those adenomas or a genetic defect within the pituitary [11], which involves different proteins or protein systems. Each type of pituitary adenoma may involve the corresponding tumor-related and specific proteins. Comparative proteomics is an excellent method to provide a systems-level approach to detect and identify those DEPs between pituitary adenomas and normal cells. Those data could: clarify the basic molecular mechanisms of pituitary tumorigenesis; classify tumors on a molecular level; identify cancer biomarkers and detect novel drug-targets; and provide an "early stage" diagnosis, potential therapy, and accurate prognosis. In turn, a pituitary adenoma is an excellent biomodel to be studied by comparative proteomics.

1.3. Basic techniques used for studying proteomics

The state-of-the-art biological mass spectrometric "soft-ionization" technologies (matrix-assisted laser desorption/ionization (MALDI); electrospray ionization (ESI)) facilitate the routine characterization of proteins. Those two ionization methods are integrated with several different ion analyzers to form a variety of mass spectrometers,

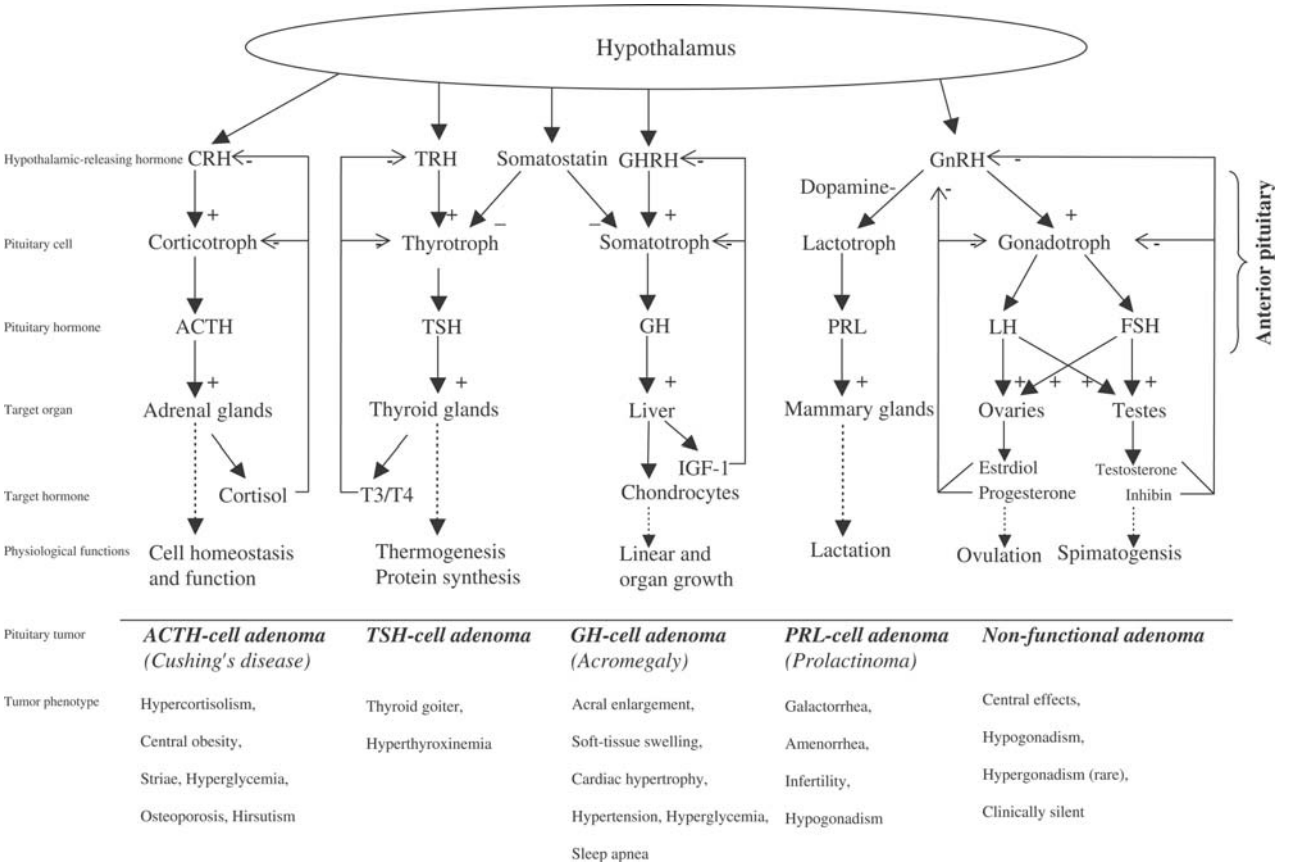


Fig. 2. A scheme of the hypothalamic–anterior pituitary–target organ axis systems and pituitary adenoma pathogenesis (Melmed [8]). (+) Stimulatory regulation; (−) inhibitory regulation. Reproduced from Zhan and Desiderio [14], with permission from Wiley-VCH, copyright 2005, and modified from Melmed [8], with permission from Copyright Clearance Center Inc. (Re: *Journal of Clinical Investigation*), copyright 2003.

including MALDI–time of flight (MALDI−TOF), ESI–quadrupole ion trap (ESI−qIT), MALDI/ESI–linear ion trap (MALDI/ESI−LTQ), MALDI/ESI−quadrupole-TOF (MALDI/ESI−Q-TOF), and MALDI−TOF−TOF instruments. MALDI–TOF is commonly used to produce peptide mass fingerprinting (PMF) data; qIT, LTQ, and Q-TOF produce MS/MS, or amino acid sequence-determining data [12]. PMF and MS/MS data can be obtained with a TOF–TOF instrument [13].

Genome sequences [1] and bioinformatics. Generate protein databases that are used to characterize proteins and peptides in a biological sample. Proteomics is a global experimental approach to analyze a proteome, which includes all of the proteins in a tissue, cell, or body fluid at any given time [14]. Protein-separation technology includes 2DGE and multiple-dimensional LC coupled with different stable isotope-labeling strategies, and protein databases include Swiss-Prot and NCBInr. Proteins are characterized with PMF and MS/MS data [12].

Search engines have been developed to compare PMF and MS/MS data to the protein databases [14]. PMF data search engines include PeptIdent (http://us.expasy.org/tools/peptident.html), Mascot (http://www.matrixscience.com/search_form_select.html), MS-Fit (http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm), and Pro-Found (http://129.85.19.192/ profound_bin/WebProFound.exe). MS/MS data search engines include SEQUEST for the ESI–qIT, PROTEINLYNX 3.5 software for the Q-TOF, Mascot software (http://www.matrixscience.com/search_form_select.html), and the Global Proteome Machine (GPM) software (http://h003.thegpm.org/tandem/thegpm_tandem.html).

MS has been used to characterize 2DGE- and LC-separated human pituitary proteins and DEPs that are related to human pituitary adenomas [2,12,15–17]. PTMs of proteins are also important to study. MS/MS effectively characterizes the PTMs and determines the modified sites. LC–ESI–qIT MS/MS has been used to identify the phosphorylation [18] and nitration sites [19–21] of proteins in the human pituitary proteome. A human pituitary proteome reference database has been established in our laboratory [12,15], many tumor-related proteins were identified [2,17]. And the primary structure and PTMs of pituitary proteins were MS-characterized [2,12,15–21]. Those data could contribute to the basic and clinical research studies of human pituitary tumors. Moreover, novel protein chip-based MS technology [22] might eventually be used for the screening, diagnostics, and therapeutics of human pituitary diseases.

2. The pituitary gland and mass spectrometry: an endocrinologist's perspective

In 1970, Guillemin and co-workers used MS to determine the amino acid sequence of the ovine thyrotrophin-releasing hormone (TRH), the first of the hypothalamic hypophysiotrophic peptides to be discovered [23]. Ten years earlier, the

development of the insulin immunoassay by Yalow and Berson [24] and its application to the measurement of other peptides [25], including pituitary hormones, triggered a rapid expansion of knowledge in pituitary physiology and disease. Before that time, pituitary hormones were measured by insensitive, laborious, and time-consuming bioassays. Significant developments and improvements in MS instrumentation, ionization methods, and computer methods now permit the separation, structural identification, and study of a large number of known and new proteins (proteomics) in the pituitary. Proteomics, genomics (the study of DNA), and transcriptomics (the study of messenger RNA and its transcription) are complementary methodologies of the modern era that are likely to increase and propel our understanding of pituitary physiology, pathophysiology, and therapeutics to new vistas.

Physiologically, an example of this new methodology is the recent identification by MS of several forms of phosphorylated GH in normal (cadaveric) pituitary glands [18]. Phosphorylation of tyrosine, serine, and threonine residues in cellular proteins such as receptors, receptor substrates, and kinases is a critical step in the signaling pathways that connect the extracellular hormones or cytokines with the ultimate biological response of cells. These data suggest the hypothesis that phosphorylated GH might function as a signaling protein within the somatotroph (an intracrine effect) in contradistinction to the GH that is secreted into blood (an endocrine effect). Other questions arise. Is the phosphorylation specific for GH, or does that PTM also occur in the other pituitary cell types? What happens to GH phosphorylation in GH-secreting pituitary adenomas?

There are other examples of PTMs in normal and adenomatous pituitary tissue, such as the nitration and nitrosylation of proteins [19–21]. Experimentally, nitric oxide (NO), an intracellular messenger molecule, activates the release of several anterior pituitary hormones. NO can combine with superoxide (O_2^-) to form peroxynitrite ($OONO^-$), which is a highly reactive anion that nitrates tyrosine residues. Likewise, NO can combine with thiol ($-SH$) groups to lead to the nitrosylation of cysteine residues. The role and significance of nitration and nitrosylation in pituicyte signaling and growth under physiological and pathological conditions require further investigation; e.g., what might be the effects of nitration and nitrosylation overproduction, or of their pharmacologic blockade?

From a pathological perspective, the detailed MS study of pituitary tumors can be used to determine whether proteins that regulate cell growth or hormone secretion are over- or under-expressed in neoplastic tissue, and whether they can be used as blood biomarkers of tumor mass and activity. For example, secretagogin, which is important for pancreatic β-cell insulin secretion, is a newly discovered protein in the normal adenohypophysis; it and its mRNA are underexpressed in null cell/gonadotroph pituitary adenomas [17]. Further studies are needed to determine whether secretagogin plays a physiological role in pituitary hormone secretion.

MS technology, coupled with a variety of powerful separation techniques, can examine the expression of more than 1000 proteins [26]. Likewise, of the thousands of genes in the human genome, the technique of gene-expression microarray can identify those genes whose functions are altered by neoplastic transformation. This method has become an important investigative tool in pituitary oncology [8,27] and other areas of clinical oncology [28]. A thumb-nail size photolithography chip is prepared that contains thousands of short DNA sequence probes of known genes that are arrayed and bound to glass surfaces. Messenger RNA is extracted from tumor samples, labeled with a fluorescent dye, and applied to the chip. The excess labeled mRNA is washed off. Spots remain only where the tumor mRNA and complementary gene-specific DNA probes have hybridized. Multiple gene analysis for up- and down-regulation is performed with Prediction Analysis of Microarrays (PAM) software, and the data are compared to normal tissue samples [28].

From a therapeutic viewpoint, new drugs are needed to treat patients—especially those patients whose pituitary tumors have not been cured by surgical extirpation and radiotherapeutic ablation, and in whom ongoing tumor growth and/or hormone hypersecretion cannot be controlled by any of the current treatment modalities. Hopefully, genomic and proteomic advances in pituitary oncology, driven in large part by MS, will lead to the realization of that clinically important goal.

3. Methodology

Proteomics is a multidisciplinary study that includes protein chemistry, MS, chromatography, bioinformatics, etc. The basic proteomics techniques are grouped into two types: protein separation and protein identification. 2DGE and LC are the main protein- and peptide-separation techniques, respectively. MS coupled with bioinformatics analysis identifies proteins. Two types of proteomics systems (LC, gel) are used to analyze human pituitary tissues to obtain tumor-related proteins and PTMs.

A brief procedure of gel-based comparative proteomics [2] is: the extracted pituitary proteins are separated with 2DGE and visualized (for example, silver-stained); digitized 2DGE images are compared to obtain differential spots between pituitary adenomas and controls; the protein in each differential spot is digested in-gel with trypsin; the purified tryptic peptides are analyzed with mass spectrometry (PMF; or MS/MS); and the MS data are used to search a protein database for protein identification. 2DGE coupled with Western blotting [19,21] is often used to detect PTMs (for example, phosphorylation and nitration) and protein isoforms. The LC procedure of stable isotope-labeling quantitative proteomics [18] is: the extracted proteins are digested with trypsin; the tryptic peptides from samples and controls are labeled with a different isotope, respectively; the labeled samples are mixed (1:1) and analyzed with LC−MS/MS; the area under the curve (AUC) of the

LC-separated peptides is used to determine the DEPs; and the MS/MS data are used to search a protein database for protein identification.

3.1. Gel-based comparative proteomics of human pituitary adenoma tissues

3.1.1. Pituitary protein preparation

Sample preparation is a very critical step for the efficient separation of proteins with 2DGE and for their subsequent MS-characterization. Moreover, good sample preparation must maximize the number of proteins that are extracted from a pituitary tissue, extract all proteins in a quantitative manner, and avoid any proteolytic protein degradation. Actually, the key to good sample preparation is an efficient protein solubilization with a minimum of handling. For pituitary protein preparation, a combination of chemical and physical methods is a very effective strategy to extract proteins from pituitary tissues by homogenization, lyophilization, chemical resolving, repeated sonication, and centrifugation [12,15].

Several critical features for the efficient extraction of proteins from a pituitary include: (i) the pituitary control and adenoma tissues should be washed thoroughly with sodium chloride (0.9%) to remove any blood on the surface of the tissue; (ii) due to the limited amount of the pituitary tissue (control and tumor; ca. 0.5 g control; ca. 15–80 mg adenoma), homogenization and lyophilization minimize any protein loss; (iii) the pharmalyte in the protein solubilization buffer improved protein solubilization, stabilized the *p*I of proteins, and improved isoelectric focusing (IEF) with an IPG strip; (iv) a combination of urea and thiourea improved protein solubilization and IEF [29]; (v) repeated sonication improved protein solubilization; (vi) any tube or pipette tip that contacts a protein sample must be siliconized; (vii) sufficient centrifugation (20 min, 13,000 $\times$ *g*) removed any undissolved material before IEF; (viii) high concentrations of acetic acid (2 M) and urea inhibited the activity of endogenous proteases (thus, protein inhibitors are not needed); and (ix) keratin from skin and hair was avoided by wearing latex gloves and cap.

3.1.2. Between-gel reproducibility and protein-loading capacity

For gel-based comparative proteomics, the levels of between-gel and between-sample reproducibility [30], and the protein-loading capacity [31] are the crucial experimental factors to accurately discover any DEPs. 2DGE and 2DGE-analysis software were used to optimize the quality and reproducibility of the 2D gels obtained from human pituitary tissues. First-dimension IEF was performed with an immobilized pH-gradient dry gel-strip (IPG strip). Two second-dimension SDS–PAGE systems were used: the horizontal single-gel system (Multiphor II) that can analyze 1 gel at a time on a pre-cast gradient gel and the vertical multigel

system (Dodeca) that can analyze up to 12 gels at a time. The spatial and quantitative reproducibilities of protein spots on gels obtained with those two systems were evaluated for the separation of the complex human pituitary tissue proteome. Our between-gel reproducibility was >98%. For the Dodeca gel system, the between-gel reproducibility and the linear separation capability were both superior compared to the Multiphor II gel system.

3.1.3. Principle of gel-based comparative proteomics

2DGE-based comparative proteomics was established in our laboratory to analyze human pituitary adenomas [2]. The proteomes from one human pituitary macroadenoma tissue and one control tissue were compared [2]. Four different types of differential protein spots were found, and the representative differential spots between pituitary adenomas and controls are shown in Fig. 3. An "increased" spot meant that the spot must exist in each adenoma-gel and in at least one control-gel, and that the ratio was >3 of the average normalized spot volume in the adenoma-gels to the average normalized spot volume in the control-gels. A "decreased" spot meant that the spot must exist in each control-gel and in at least one adenoma-gel, and that the ratio was >3 of the average normalized spot volume in the control-gels to the average normalized spot volume in the adenoma-gels. A "new" spot meant that the spot must exist in each adenoma-gel, but not in any control-gel. A "lost" spot meant that the spot must exist in each control-gel, but not in any adenoma-gel. With an improvement in the 2DGE sensitivity, the "lost" spot might turn into a "decreased" spot, and the "new" spot might turn into "increased" spot. Strictly speaking, the "increased" and "new" spots should belong to an expression-up-regulated protein, and the "decreased" and "lost" spots should belong to a down-regulated protein. Mass spectrometry [PMF (Fig. 4) and MS/MS (Fig. 5)] and bioinformatics have been extensively used to characterize the 2DGE-seperated pituitary proteins.

3.1.4. Heterogeneity of human pituitary proteomes

The heterogeneity of human pituitary proteomes is a crucial factor that must be studied in order to accurately validate DEPs between pituitary adenomas and controls because the pituitary adenoma and control tissues are not from the same patient; those different sources result in some uncontrolled experimental factors (gender, age, race) that occur within any comparative proteomics study of human pituitary adenomas. Therefore, the heterogeneity of a human pituitary proteome was studied as a function of three important factors—gender, age, and race [16].

A total of 30 high-resolution 2DGE gels from eight pituitary control tissues were used for a comparative analysis. A total of ca. 1000 protein spots were detected in each 2DGE map. Fifty-one differential spots (7 spots due to gender, 17 to age, 15 to race, and 12 to the co-effects of age and race) were found when pituitary

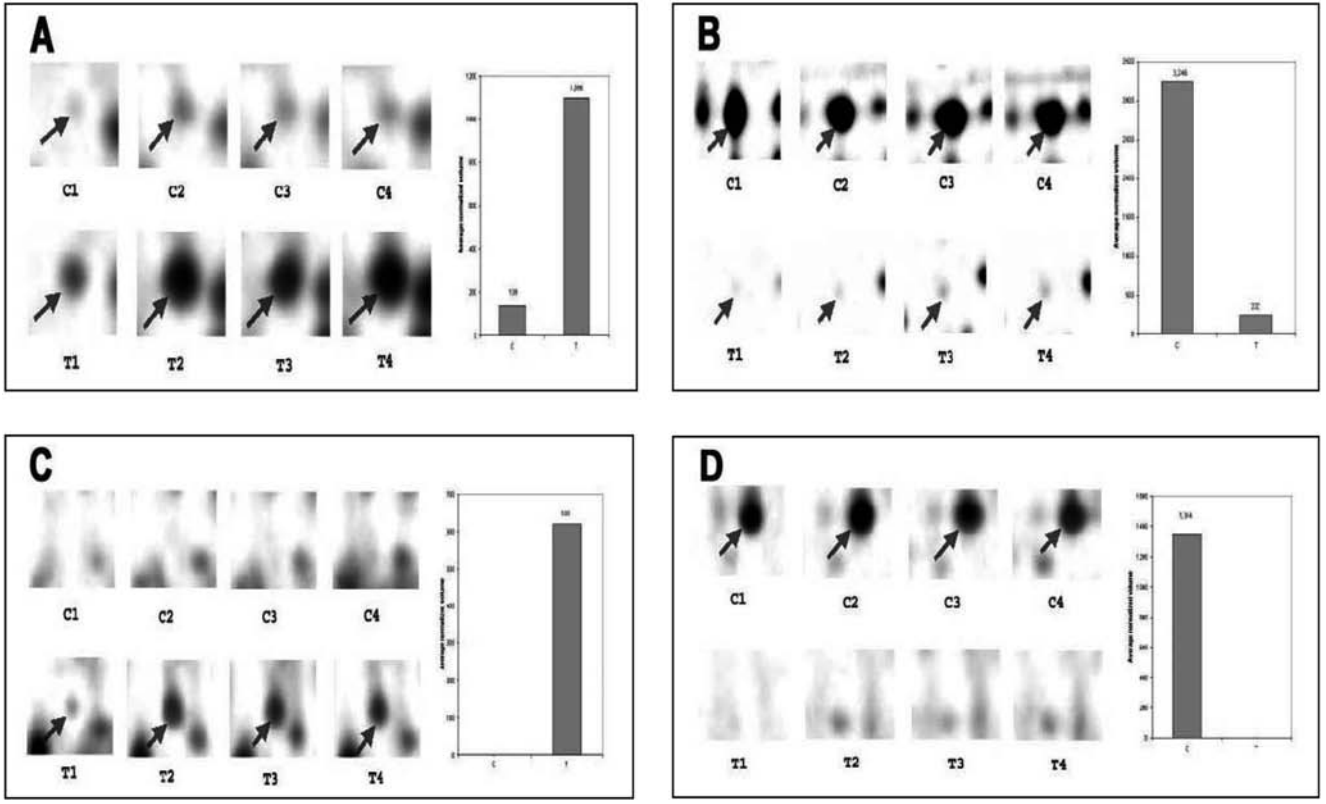


Fig. 3. Four types of representative differentially expressed protein spot images in a human pituitary adenoma compared to a pituitary control (Desiderio and Zhan [2]). (A) An "increased" protein spot in an adenoma (vs. control); (B) a "decreased" protein spot in an adenoma (vs. control); (C) a "new" protein spot in an adenoma (in adenoma, not in control); (D) a "lost" protein spot in an adenoma (in control, not in adenoma). C1, C2, C3, and C4 are pituitary controls; T1, T2, T3, and T4 are pituitary adenomas. Reproduced from Desiderio and Zhan [2], with permission from *Cellular and Molecular Biology*, copyright 2003.

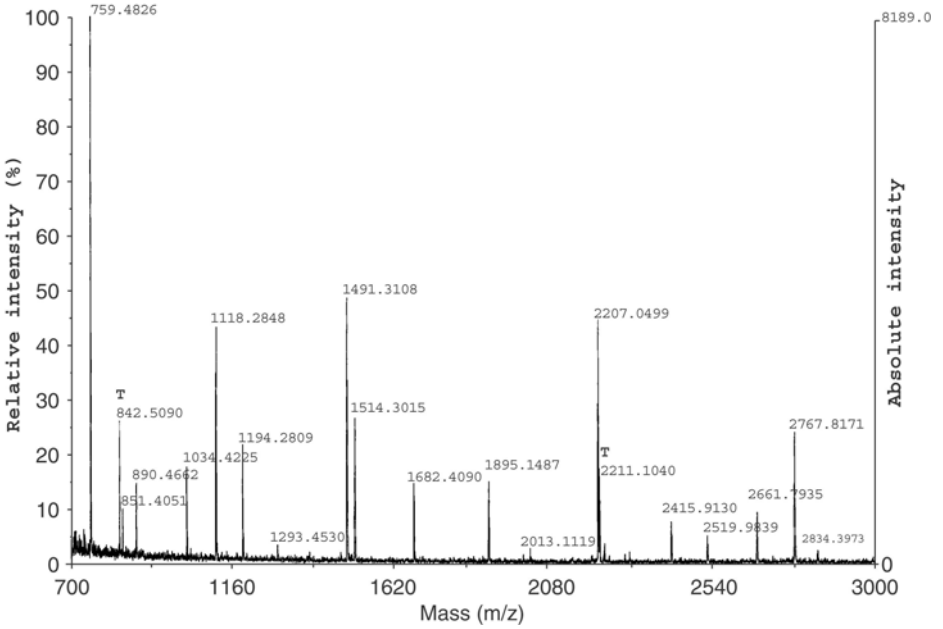


Fig. 4. A representative MALDI–TOF mass spectrum (Desiderio and Zhan [2]). T: a trypsin auto-digestion fragment. The gel spots were cut from a silver-stained 2D gel, digested in-gel with trypsin (37°C, ca. 18 h). The tryptic peptides were purified with a ZipTip C18 tip (Millipore), and eluted directly from the microcolumn onto the MALDI plate with 2 μl of an α-cyano-4-hydroxycinnamic acid (CHCA) solution (12.5 mg/ml in 50%, v/v, acetonitrile/0.1%, v/v, TFA). The matrix was dried in ambient air. The mass spectrum was obtained in the delayed extraction, reflectron, positive-ion mode on a Perseptive Biosystems MALDI–TOF Voyager DE-RP mass spectrometer (Framingham, MA, USA). The mass spectrum was internally mass-calibrated with two fragment-ion masses of the trypsin auto-digestion products ($[M+H]^+ = 842.509$ and 2211.104 Da). Those masses that result from trypsin, matrix, keratins from skin and hair, and other unknown contaminants were removed manually from the mass spectrum. The corrected list was the PMF data that were used to search those protein databases. The protein was identified as fibrinogen gamma chain (Swiss-Prot No. P02679). Reproduced from Desiderio and Zhan [2], with permission from *Cellular and Molecular Biology*, copyright 2003.

proteomes were compared according to gender, age, and race. For those 51 differential spots, 33 DEPs (6 spots due to gender, 9 to age, 8 to race, and 10 to the co-effect of age and race) were MS-characterized. A functional analysis of those DEPs showed that prolactin was expressed higher in the female than in the male, and that somatotropin was related to gender, age, and race. Some proteins associated with hormone regulation (for example, follistatin, thyroid hormone receptor beta-2, adenylate cyclase-inhibiting G alpha protein) were related to age and race. The DEPs that were related to age were mainly those proteins that are associated with cell growth, proliferation, differentiation, apoptosis, and death; those proteins did not show any difference with gender and race. Those differential spots were not

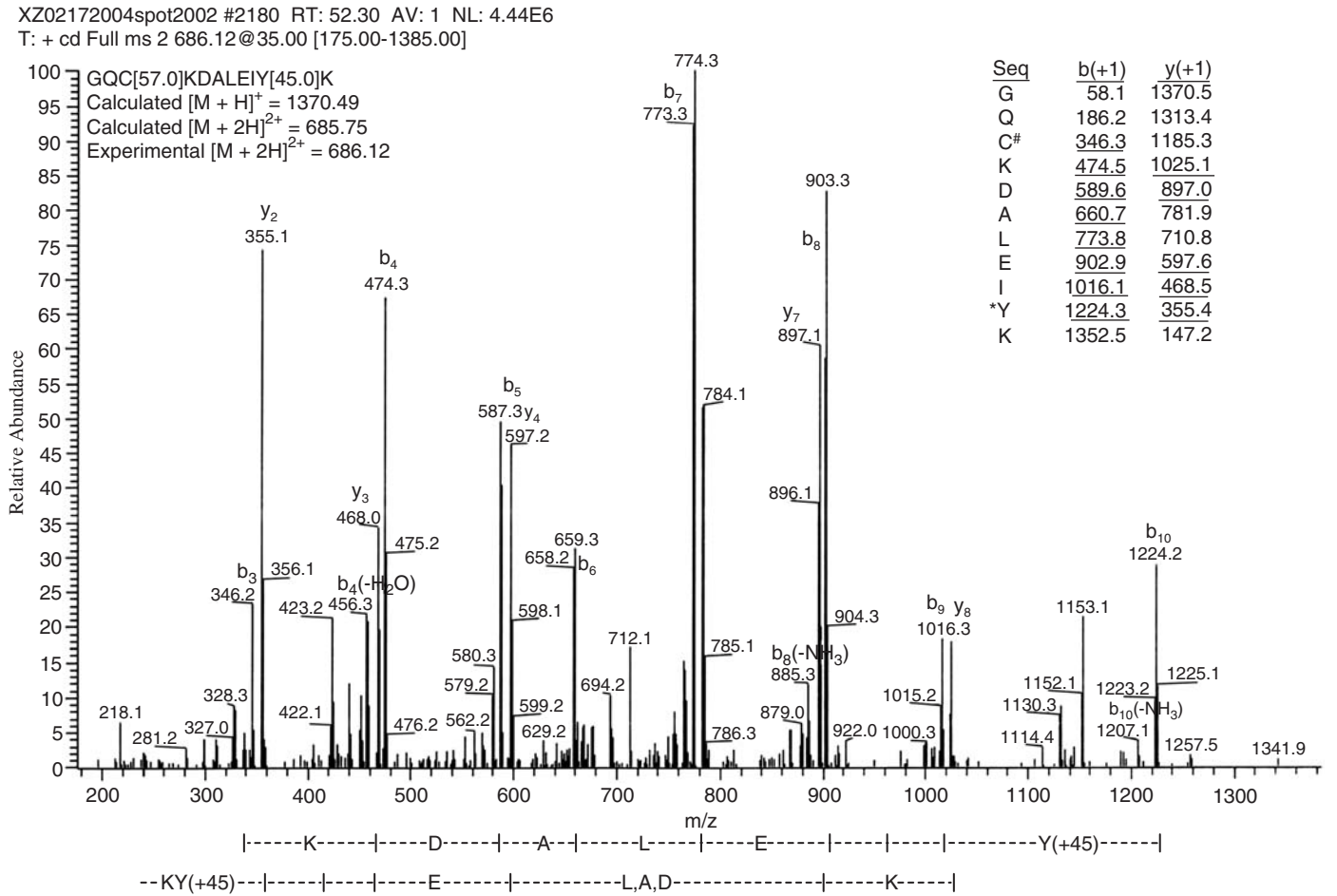


Fig. 5. SEQUEST (top right) and *de novo* (bottom) analysis of an MS^2 spectrum of the precursor ion ($[M+2H]^{2+}$, at m/z = 686.12, RT = 52.30 min, and scan number 2180) for a nitrotyrosyl peptide (Tyr-237) ^{228}GQC$^{\#}$KDALEI*YK^{238} that contained 11 amino acids, and that was derived from synaptosomal-associated protein (spot 1). *Y: nitrotyrosine. C$^{\#}$: carbamidomethyl-Cys. Reproduced from Zhan and Desiderio [19], with permission from Elsevier Science (USA), copyright 2004.

found in our subsequent large-scale comparison from pituitary adenoma tissues and pituitary control tissues [17].

3.1.5. Proteomic profiles of human pituitary adenomas

3.1.5.1. Reference 2DGE map. One control pituitary and one adenoma tissue were analyzed with 2DGE and MS. For the control pituitary proteome [15], the 2D map contained 1094 protein spots; a total of 62 spots, corresponding to 38 different proteins, were MS-characterized. For the pituitary adenoma tissue proteome [12], the 2D map contained ca. 1000 protein spots; 135 protein spots that represented 111 proteins were MS-characterized. Those proteins correlated to different functional groups. The protein identification data were used to construct a Web-based reference database of the human pituitary (www.utmem.edu/proteomics).

3.1.5.2. Differentially expressed proteins. A large-scale comparative proteomics study was performed on a set of human pituitary samples: controls ($n = 8$, gels = 30) vs. several different cell types of non-functional (NF) pituitary adenomas (NF^-, $n = 3$, gels = 9; LH^+, $n = 3$, gels = 9; FSH^+, $n = 3$, gels = 9; $FSH^+ + LH^+$, $n = 3$, gels = 9; unknown cell type, $n = 3$, gels = 3) [32] and prolactinomas ($n = 4$, gels = 12) (Evans et al., in preparation). A total of 251 differential spots were found, among which 93 differential protein spots (65 decreased spot volumes, 28 increased) were subjected to in-gel trypsin digestion and MS-characterization. Seventy-two spots (50 decreased, 22 increased), representing 56 DEPs (34 down-regulated, 22 up-regulated), were characterized with MS and database analysis. The functional roles that are involved in those multiple protein systems are summarized in Fig. 6.

Results indicated that: (i) neuroendocrine-related proteins (somatotropin, secretagogin, and μ-crystallin homolog) were down-regulated in NF pituitary adenomas and the prolactinomas; (ii) prolactin existed in six isoforms that were down-regulated in NF adenomas, and were not changed in the prolactinomas; (iii) somatotropin existed in at least 17 isoforms that were down-regulated in NF adenomas and the prolactinomas; (iv) cell proliferation, differentiation, and apoptosis-related proteins were down-regulated in the NF adenomas and the prolactinomas; (v) immunologic regulation proteins and tumor-related antigen (immunoglobulin, tumor rejection antigen-1) were down-regulated in NF adenomas; (vi) some cell-defense and stress-resistance proteins (phospholipid hydroperoxide glutathione peroxidase, CD59 glycoprotein, and heat shock 27 kDa protein) were down-regulated in the pituitary adenomas; (vii) some metabolic enzyme-related proteins (for example, isocitrate dehydrogenase [NADP] cytoplasmic, tryptophan 5-hydroxylase 2, matrix metalloproteinase-9, aldose reductase, lactoylglutathione lyase, acyl-CoA-binding protein, etc.) were up-regulated in the pituitary adenomas; and (viii) for cell-signal proteins, some were down-regulated and some were up-regulated in adenomas;

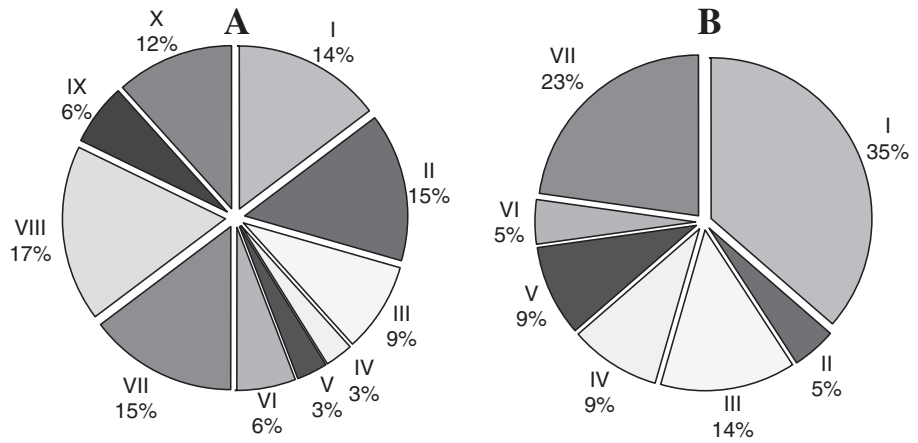


Fig. 6. Functional categories of the 56 MS-characterized differentially expressed proteins. (A) Down-regulated proteins in pituitary adenoma ($n = 34$): (I) neuro-endocrine and hormones; (II) cytokine and cellular signal-related proteins; (III) cellular defense and stress resistance; (IV) mRNA splicing, transport or translation-related enzyme; (V) DNA-binding proteins; (VI) metabolic enzymes; (VII) immunologic regulation proteins and tumor-related antigen; (VIII) transport proteins; (IX) cell proliferation, differentiation, apoptosis-related proteins; (X) others. (B) Up-regulated proteins in pituitary adenomas ($n = 22$): (I) metabolic enzyme-related proteins; (II) energy metabolism; (III) cellular signal proteins; (IV) cell cycle, cell growth and proliferation proteins; (V) cellular defense response; (VI) protein folding-related protein; (VII) others. Reproduced from Zhan and Desiderio [14], with permission from Wiley-VCH, copyright 2005.

those cell signals are involved in the complex biological roles in the cell growth, proliferation, differentiation, apoptosis, and death cycles.

3.1.5.3. Comparative proteomics data vs. comparative transcriptomics data. Those same pituitary tumor samples [NF adenomas (NF^-, $n = 3$; LH^+, $n = 3$; FSH^+, $n = 3$; $FSH^+ + LH^+$, $n = 3$; unknown cell type, $n = 3$) and prolactinomas ($n = 4$)] were also analyzed with a GeneChip microarray to detect the differentially expressed genes (DEGs) at the mRNA level in those human pituitary adenomas compared to controls. A total of 374 DEGs were found (215 down-regulated, 159 up-regulated) in NF adenomas with a change-fold of >2, and 213 genes (153 down-regulated, 60 up-regulated) in the prolactinomas with a change-fold of >2. Those comparative proteomics data (56 DEPs derived from 72 differential spots) were compared to those comparative transcriptomics data to determine any consistent or similar results at the protein and mRNA expression levels ([32], Evans et al., in preparation). Nine genes—somatotropin, prolactin, secretagogin, tissue transglutaminase, isocitrate dehydrogenase [NADP] cytoplasmic, cellular retinoic acid-binding protein II, G-protein beta, calreticulin, and hemoglobin beta chain—indicated a consistent change in the protein and mRNA expression levels in adenomas relative to controls (Table 1).

Table 1
Comparative proteomics vs. comparative transcriptomics in human pituitary adenomas

Differentially expressed proteins				Differentially expressed genes			
Swiss-Prot	Protein name	NF	PRL	GenBank	Gene name	NF	PRL
P01241	Somatotropin (GH1)	−	−	NM_000515	GH1: growth hormone 1	−	−
				NM_002059	GH2: growth hormone 2	−	−
				NM_000823	GHRHR: growth hormone-releasing hormone receptor	−	−
P01236	Prolactin (PRL)	−	+/−	NM_000948	PRL: prolactin	−	+/−
O76038	Secretagogin (SCGN)	−	− (weak)	NM_006998	SCGN: secretagogin	−	
P21980	Tissue transglutaminase (TGM2)	−	−	NM_004613	TGM2: transglutaminase 2	−	−
O75874	Isocitrate dehydrogenase [NADP] cytoplasmic (IDH1)	+	+ (weak)	NM_005896	IDH1: isocitrate dehydrogenase ($NADP^+$), soluble	+	
P29373	Cellular retinoic acid-binding protein II	+		M97815	Cellular retinoic acid-binding protein 2	+	
P16520	Guanine nucleotide-binding protein G(I)/G(S)/G(T) beta subunit 3	+		M31328	Guanine nucleotide-binding protein (G protein), beta polypeptide	+	
P27797	Calreticulin precursor	+		AI807225	KDEL endoplasmic reticulum (ER) protein	+	
gi1066765	Hemoglobin, beta chain (HBB)	−	−	NM_000518	HBB: hemoglobin, beta	−	−
				NM_000519	HBD: hemoglobin, delta	−	

NF: non-functional pituitary adenoma; PRL: hyperprolactinoma. The bracket after the protein name refers to its corresponding gene name. (+) Up-regulated in pituitary adenomas relative to controls; (−) down-regulated in pituitary adenomas relative to controls. Reproduced from Zhan and Desiderio [14], with permission from Wiley-VCH, copyright 2005.

3.2. The proteomics of PTM proteins in human pituitary adenomas

3.2.1. Two-dimensional Western blotting to study PTM proteins

2DGE-based Western blotting against anti-3-nitrotyrosine antibodies, and also against anti-pSer, pThr, pTyr antibodies, is an effective method to detect any nitrotyrosine proteins and phosphorylated proteins, respectively. A Western blotting image was digitized and PDQuest 2D image analysis was performed to locate the positive spots between the Western blot image and negative controls, and the positive Western blot spots were matched to the corresponding silver-stained 2D gel image. That approach was used to detect nitroproteins in human pituitary postmortem tissues (Fig. 7; see below) [19,21]. Moreover, two-dimensional Western blotting against individual protein antibodies could be used to array the isoforms of each protein.

3.2.2. Determination of phosphorylation sites

An off-line immobilized metal affinity column (IMAC, Ga^{3+}) preferentially enriched the phosphopeptides that were present in a complex whole-digest mixture from a pituitary control tissue, and the phosphopeptide-enriched samples were analyzed with LC–ESI–Q-IT MS/MS under experimental conditions that optimized the neutral loss of a molecule of phosphoric acid (H_3PO_4 = 98 Da) from a phosphopeptide [18]. Six phosphorylated proteins (GH, chromogranin A, secretogranin I, 60S ribosomal protein P1 and/or P2, DnaJ homolog subfamily C member 5, and galanin) were identified (Table 2), and each phosphorylation site was determined with MS/MS and MS^3. The structure and function of the six identified phosphoproteins were analyzed in detail. For several of these proteins (e.g., GH), Giorgianni et al. were the first to describe their phosphorylation in the human, and those findings are now listed in the Swiss-Prot annotations and in the Phosphosite knowledge base. Since that study, 81 phosphopeptides that contained *50* different phosphorylation sites have been found; those phosphopeptides map to *26* different phosphoproteins [33].

3.2.3. Determination of nitration sites

2DGE-based Western blotting was used to detect, and LC–MS/MS to characterize, nitroproteins in the human pituitary control tissues [19]. Proteins from 2D gel spots that corresponded to the strongly positive anti-nitrotyrosine Western blot spots (Fig. 7) were subjected to in-gel trypsin digestion and LC–MS/MS analysis. MS/MS determined the nitration site of each nitrated peptide. Each amino acid sequence was first determined by the accurate *de novo* sequence method, and secondly by SEQUEST analysis (Fig. 5). *De novo* sequencing

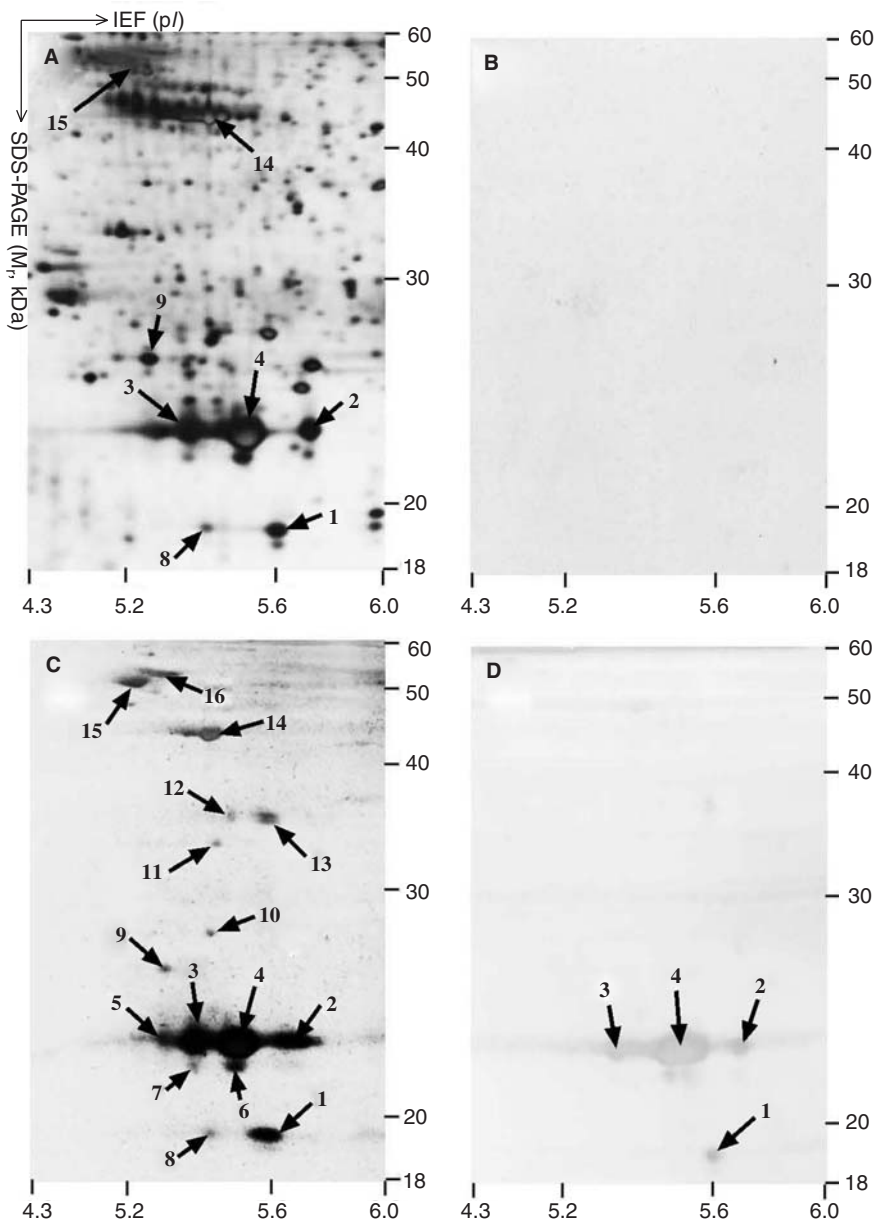


Fig. 7. Two-dimensional Western blot analysis of anti-3-nitrotyrosine proteins in a human pituitary (70 μg protein per 2D gel). (A) Silver-stained image on a 2D gel before transfer of proteins to a PVDF membrane. (B) Silver-stained image on a 2D gel after transfer of proteins to a PVDF membrane. (C) Western blot image of anti-3-nitrotyrosine proteins (anti-3-nitrotyrosine antibodies + secondary antibody). (D) Negative control of Western blot to show the cross-reaction of the secondary antibody (only the secondary antibody; no anti-3-nitrotyrosine antibody). Reproduced from Zhan and Desiderio [19], with permission from Elsevier Science (USA), copyright 2004.

Table 2
Phosphoproteins from human pituitary digests identified by LC–MS/MS

Protein(Swiss-Prot number)	Peptide identified	Phosphorylation site	SEQUEST scores (X_{corr})	MASCOT scores
Human growth hormone (P01241)	[121]SVFANSLVYGASDSNVYDLLK[141]	Ser 132	3.47	48
	[172]FDTNSHNDDALLK[184]	Ser 176	3.63	32
Chromogranin A (P10645)	[319]GGKSGELEQEEER[331]	Ser 322	3.61	44
Secretogranin I (P05060)	[134]ADEPQWSLYPSDSQVSEEVK[153]	Ser 149	3.26	58
	[397]MAHGYGEESEEER[409]	Ser 405	2.67	24
60S acidic ribosomal protein P1 (P05386) and/or 60S acidic ribosomal protein P2 (P05387)	[98]KEESEESDDDMGFGLFD[114]	Ser 101	4.84	35
	[99]KEESEESDDDMGFGLFD[115]	Ser 102	4.84	35
DnaJ homolog subfamily C member 5 (Q9H3Z4)	[8]SLSTSGESLYHVLGLDK[24]	Ser 10	4.24	62
Galanin (P22466)	[108]LLDLPAAASSEDIERS[123]	Ser 117	5.06	41

Reproduced from Giorgianni et al. [18], with permission from Wiley-VCH, copyright 2004.

independently and accurately determined each amino acid sequence before SEQUEST analysis. Four different nitrated peptides were characterized, and were matched to four different nitroproteins (Table 3).

3.2.4. Analysis of protein isoforms

Protein isoforms result from PTMs, splicing variants, etc. Each protein isoform has its own *p*I and M_r values; 2DGE, or 2DGE coupled with corresponding protein antibodies, is an effective method to array those different isoforms of each protein. MS, especially MS/MS, plays an important role in the characterization of each PTM and splicing variant. We found that prolactin had multiple isoforms in human pituitary control tissues—six 2D gel spots that contained prolactin were identified [32; Evans et al., in preparation]. Further experiments are needed to determine whether the ratio of each prolactin isoform changes in NF pituitary adenomas and prolactinomas.

Twenty-four 2D gel spots that contained human GH were found in human pituitary control tissues [34]. Those hGHs in the 24 2D gel spots were classified into the four types of hGH splicing isoforms, 1–4. The expression proportion of those

Table 3
Human pituitary nitroproteins determined by amino acid sequencing (SEQUEST; *de novo*) from LC–MS/MS

Spot number	Protein name	Swiss-Prot number	Nitrotyrosyl peptide	Tyr nitration site	SEQUEST (X_{corr})	Interpretation
1	Synaptosomal-associated protein	O60641	^{228}(K)GQC$^{\#}$KD ALEI*YK^{238}	237	3.44	SEQUEST
			^{231}KDALEI*YK^{238}	237		*de novo*
2	Immunoglobulin alpha Fc receptor	P24071	^{223}(D)*YTTQNLIR^{230}	223	1.87	SEQUEST
			^{223*}YTTQNLIR^{230}	223		*de novo*
4	Immunoglo bulin alpha Fc receptor	P24071	^{223}(D)*YTTQNLIR^{230}	223	1.95	SEQUEST
			^{223*}YTTQNLIR^{230}	223		*de novo*
14	Actin	P03996 or P12718 or P04270	^{294}(K)DL*YANNV LSGGTTMYPGI ADR^{314}	296	3.25	SEQUEST
			^{293}(K)DL*YANNV LSGGTTMYPGI ADR^{313}	295		
			^{294}(K)DL*YAN NVLSGGTTM YPGIADR^{314}	296		
15	cGMP-dependent protein kinase 2	Q13237	^{352}(K)GE*YFGEK ALI^{361}	354	1.96	SEQUEST
			^{352}GE*YFGEK^{358}	354		*de novo*

*Y: nitrotryrosine. C$^{\#}$: Cys-CAM. The bracket refers to the amino acid residue that preceded the N-terminus of the nitrated peptide. Reproduced from Zhan and Desiderio [19], with permission from Elsevier Science (USA), copyright 2004.

four isoforms was isoform 1 (87.5%) > isoform 2 (8.1%) > isoform 3 (3.3%) > isoform 4 (1.1%); a significant statistical difference was found among those isoforms. PTM analysis demonstrated that, among those 24 GH spots, some spots had a measurably different *p*I, but the same M_r; that result could be due to the deamidation of asparagine to aspartate that was identified with MALDI–TOF MS. That deamidation caused a change in charge. Other spots had a measurably different M_r, but the same apparent *p*I; that result could be due to N-glycosylation, polymer formation, or proteolysis. MS/MS data demonstrated that the hGH in 1 of the 24 2D-gel spots was a phosphoprotein with three phosphate groups (Ser-77,

Ser-132, and Ser-176). Further study is required to determine whether the ratio of GH isoform changes in human pituitary adenoma compared to controls.

3.3. Challenge of comparative proteomics in the study of human pituitary adenomas

The gel comparative proteomics method is highly reproducible (>98%), and is the most direct method to elucidate the complexity of systems that cause human pituitary adenomas [2,12,16,17,30,31]. However, that gel-based system is challenged in terms of its ability to locate and characterize "all" tumor-related proteins [8,12,14,16]:

(i) Pituitary adenomas and control tissues cannot be obtained from the same patient. An adenoma tissue is commonly obtained from neurosurgery, whereas a control pituitary is obtained from a postmortem autopsy; e.g., a death from another disease, a gun-shot wound, or an accident. Thus, in order to obtain the DEPs between pituitary adenoma and control tissues, an accurate evaluation of the heterogeneity of different control pituitary tissue proteomes is needed as a function of gender, age, and race. We analyzed the heterogeneity of a human pituitary proteome with eight control pituitaries [16]. An expanded number of samples is still needed for the heterogeneity analysis.

(ii) Most of the currently identified proteins in the 2DGE map of a pituitary adenoma or control tissue are cytoplasmic, which is only one "window" into the complete proteome of a pituitary adenoma or a control tissue. That window occurs because only one protein-extraction protocol cannot extract "all" of the proteins from a pituitary because there are different classes of proteins—especially the hydrophobic proteins that include most of the membrane and nuclear proteins.

(iii) In all reviewed documents, the total proteins extracted from pituitary control tissues or adenoma tissues were separated with 2DGE, based on a IPG strip pH 3–10 NL. Although the pH range of an IPG strip is 3–10, most of well-separated protein spots in the 2DGE maps from pituitary control and adenoma were distributed in the area of pH 4–8 and M_r 15–100 kDa. The extremely acidic (pI < 3.5 or 4)/basic (pI > 7.5 or 8) proteins and the extremely high-mass (>150 kDa) or low-mass (<10 kDa) proteins were either not well separated or cannot be separated. Thus, any DEPs that occur within those areas on a 2D gel were not detected.

(iv) The detection of low-abundance DEPs was limited in the current method due to two factors: (a) some low-abundance DEPs cannot be visualized with the silver-stain method; (b) some low-abundance DEPs have been detected, but cannot be MS-characterized—possibly due to the small amount of protein in the gel spot. Other high-abundance proteins (albumin, hemoglobin, somatotropin, prolactin [12]) possibly hindered the separation of other low-abundance proteins; pre-separation is needed.

(v) A pituitary adenoma is monoclonal in origin, and a pituitary adenoma tissue sample should be a pure cell type. However, due to the limitations imposed on any neurosurgical method to obtain tissue, mixed cell types are obtained. A control pituitary tissue is usually obtained from a postmortem source and includes multiple cell types. Therefore, any accurate comparison of proteomes between adenoma and control tissues is compromised.

Moreover, gel-based comparative proteomics were only used to analyze NF pituitary adenomas and prolactinomas. However, pituitary adenomas include different cell types: GH, ACTH, TSH, PRL, and LH/FSH. Theoretically, for those different cell types of adenomas, not only could a common mechanism occur in their formation, but also some differences could exist among the different cell types of an adenoma. It is necessary to expand gel-based comparative proteomics to other cell types of pituitary adenomas.

4. Discussion

4.1. Insights into the basic molecular mechanisms of pituitary tumor formation

Proteomics provides a unique insight on a global protein-system level into the basic molecular mechanisms that participate in the formation of a pituitary adenoma. Through our series of comparative proteomics, we have confirmed our initial hypothesis that the proteome differs between pituitary controls and adenomas—many DEPs were found that correlated to the comparative transcriptomics data and to the corresponding biological systems. Those protein systems summarized in Table 4 provide a hint of the multiple protein systems that are involved in the formation of a pituitary adenoma. Those DEPs provide a basis to determine the activities that are critical for the observed changes in expression—whether they occur either at the mRNA or protein level or by PTM, and whether they are involved in the formation of a human pituitary adenoma.

Several protein systems that are discussed below connect some of the critical aspects of pituitary adenoma formation: (i) The alteration of signal transduction systems is significantly involved in the pathophysiological processes of pituitary adenomas. We found that multiple signal pathways were changed, including G-proteins, cytokine-receptors (IFG, IL, EGF, TGF, and IFN), some signal system-regulated enzymes (MAPK4, phospholipase A2-IB, and cAMP-dependent protein kinase), etc. Other studies also found that G-proteins played important roles in pituitary tumors [35], and that members of the FGF family are implicated in pituitary tumorigenesis [36]. FGF receptor genes mediate FGF signaling, those genes encode receptor tyrosine kinases, and our DEP data also include a Tyr-protein kinase. Moreover, phosphorylation significantly participates in the regulation of signal pathways, and we found a Ser/Thr protein phosphatase 2A. (ii) The changes

Table 4
List of biological systems, DEPs (n = 56, Moreno et al. [32], Evans et al., in preparation), and DEGs (n = 128, Evans et al. [38])

Systems	DEPs	DEGs
(A) Signal transduction		
(1) G-proteins[a]	G_0 subunits 1, 2**	G-protein G-binding protein Regulator of G-protein signaling 16 Regulator of G-protein bind signaling 2, 5 Rho GTPase activating protein 5
(2) Cytokine-receptors		
(a) IGF	IGF-binding	IGF IGF-binding IGF-binding 5 IGF-binding 3 Protease Ser 11 IGF-binding
(b) IL	Splice isoform IL-15	
(c) EGF		EGF-containing fibulin ECM protein 1
(d) TGF		TGF_α receptor III
(e) IFN		Interferon-induced protein 56
(3) Signal-system enzymes		MAPK 4 Phospholipase A2-IB Protein kinase cAMP-dependent β
(4) Retinoic acid		Cellular retinoic acid-binding protein 2
(5) Phosphorylation	Ser/Thr protein phosphatase 2A	
(6) Others	Rab GDP dissociation inhibitor alpha	SH3-domain GRB2-like 2 *XIST* nuclear receptor 1, 3
(B) DNA/mRNA methylation[a]	^{6}N-adenosine methyltransferase**	RNA-binding protein 2 U4/U6-associated RNA splicing factor Eukaryotic translation initiation factor 3–5
(C) Tumor genes		
(1) Oncogenes		Neuroblastoma over-expressed gene
(2) Protooncogenes[a]	Protooncogene Tyr protein kinase FYN** L-myc-1 protooncogene protein	
(3) Tumor suppressor genes	Tumor rejection antigen (endoplasmin)	

(*continues*)

Table 4
Continued

Systems	DEPs	DEGs
(D) Hormones		
(1) PRL	Prolactin	PRL
(2) GH	Somatotropin	GHRH receptor
(3) FSH		FSH, β
(4) LH		LH, β
(5) TSH		TSH, β
(E) Excitatory (tachykinins)	Secretagogin	Reticulocalbin l, EF-hand calcium-binding domain Peptidylglycine α-amidating monooxygenase
(F) Reactive-oxygen species		
(1) HSP	HSP27	HSP 105 kDa
(2) Cytochrome P450		Cytochrome P450, III A
(3) GST	GSTμ-2	
(4) Peroxidase	Phospholipid hydroperoxide glutathione peroxidase	
(G) Energy	ATP synthase, mitochondrial ATP-binding protein	ATPase
(H) Immune system	Ig κ, λ	IgG CD58 Pre-B cell leukemia transcription factor 3 MHC class 1
(I) Cell cycle (G_1–S–G_2–M)		Death-associated protein Folate receptor C1
(J) Transcription factors		Pit 1 Basic transcription factor 3
(K) Intermediate filaments	Vimentin	Vimentin

[**]The identified DEP was confirmed by publication data.
[a] Publication data indicated that G-protein [35], methylation [37], c-myc [39], PTTG, gsp, ccnd1 [39], FGF [36], Men-1, Prop-1 [39], c-jun, Fos, and angiogenesis were related to pituitary adenomas.

of endocrine hormone levels are an important clinical feature of pituitary adenomas. We found that multiple hormones were altered, including GH, PRL, FSH, LH, and TSH. The synthesis and secretion of each hormone is tightly regulated by multiple signaling and related factors. For example, cGMP-dependent protein kinase and NO significantly regulate the signal process of water-soluble hormones. Recently, we found [19] that the cGMP-dependent protein kinase 2 was nitrated in human pituitary control tissues; those data hint that the nitration of a cGMP-dependent protein kinase might mediate the signal processing of water-soluble hormones. (iii) Oxidative/nitrative stresses are involved in a variety of

tumorigenesis mechanisms. Our studies demonstrated that many reactive-oxygen species (ROS)-related proteins were changed in human pituitary adenomas, including heat shock protein, cytochrome P450, GST, and phospholipid hydroperoxide gluta-thione peroxidase. (iv) DNA methylation has been studied in pituitary tumors [37]; we found a methylation-related DEP (Table 4), and a methylation-related DEG was published [32,38]. (v) Oncogene activation in pituitary tumors (such as c-Myc) was studied [39]; we found protooncogene-related DEPs (Table 4), and related DEGs were published [32,38].

Those data reveal an overview of multiple protein systems that could participate in human pituitary adenomas.

4.2. Discovery of potential biomarkers related to pituitary adenomas

Those DEPs (Table 1) that were identified by comparative proteomics and comparative transcriptomics studies constitute our novel source of candidate biomarkers [2,32]—for example, secretagogin [17]. Secretagogin is a neuroendocrine and pancreatic islet of Langerhans-specific Ca^{2+}-binding protein [40] that is also expressed in a high quantity in the secretory neurons of the anterior pituitary [41]. Secretagogin transfection lowered the growth rate of RIN-5F tumor cells [40] by down-regulating the transcription of the excitatory amino acid substance P (SP). The latter finding is consistent with our working hypothesis [42–46] that an imbalance between excitatory (tachykinin) and inhibitory (opioid) neuropeptidergic systems could contribute to the formation of a human pituitary macroadenoma. The comparative proteomics and the comparative transcriptomics analyses [17] both demonstrated that secretagogin was significantly down-regulated in a set of human NF pituitary adenomas with a statistically significant difference ($p < 0.05$) (Figs. 8 and 9). Moreover, the secretagogin protein expression correlated significantly with its mRNA expression. Further analyses will be needed to determine whether secretagogin plays an important role in the formation of a pituitary adenoma and in pituitary hormone secretion.

4.3. Pituitary hormone isoforms in human pituitary adenomas

Protein isoforms result from protein PTMs, splicing, etc. Our comparative proteomics study of human pituitary adenomas demonstrated that human pituitary hormones had multiple isoforms. The change of a neuroendocrine hormone level is an important clinical feature in a human pituitary adenoma. Not all of the isoforms are related to pituitary adenomas [32].

Somatotropin was significantly down-regulated at the protein and the mRNA levels in the NF pituitary adenoma [32] and in prolactinomas (Evans et al., in preparation); that finding is consistent with their monoclonal composition in origin [9,10]—a NF adenoma generated from gonadotrophs, and a prolactinoma

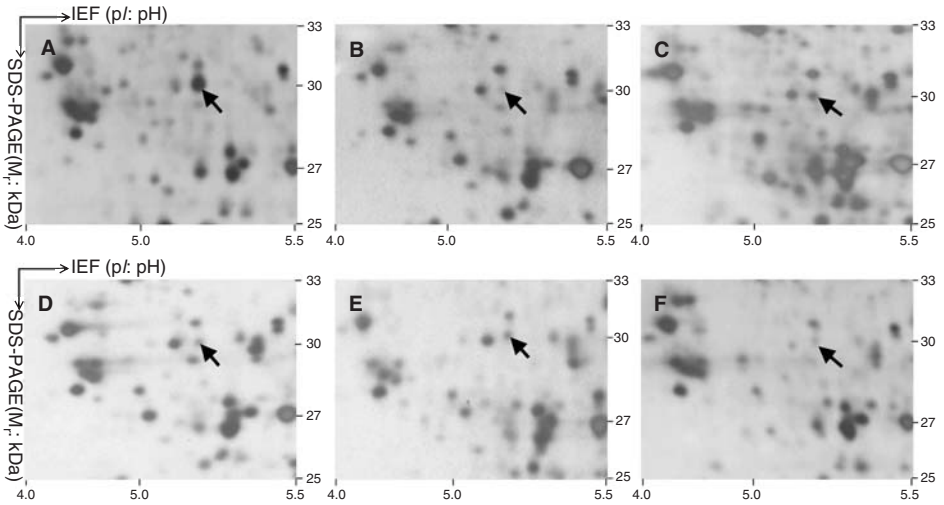


Fig. 8. Quantitative analysis of secretagogin that is contained in a differential protein spot (see arrow) in human non-functional pituitary adenoma 2D gels compared to control pituitary 2D gels (Zhan et al. [17]). (A) Control (C7-5); (B) NF^- adenoma (T217-2); (C) LH^+ adenoma (T204-2); (D) FSH^+ adenoma (T57-3); (E) FSH^+ + LH^+ adenoma (T65-1); (F) unknown cell type adenoma (T2-1). The extracted total proteins from each human pituitary adenoma or each control sample were 2DGE-separated. The first-dimension IEF was performed with 18 cm IPG strip pH 3–10 NL. The second-dimension SDS–PAGE was performed with 12% PAGE resolving gel. The silver-stained 2D gel was digitized, and analyzed with PDQuest 2D image analysis software. The total density in a gel image was used to normalize each spot volume in the gel image to minimize the effect of any experimental factors on spot volume. Reproduced from Zhan et al. [17], with permission from Kluwer Academic Publishers, copyright 2003.

from lactotrophs; the GH receptor gene was unchanged. Those data showed that GH hyposecretion in NF adenoma results from the hypoexpression of the GH gene. However, the more important finding is that the multiple GH isoforms (17 spots contain somatotropin) were detected ([32], Evans et al., in preparation); those data cannot be interpreted by the transcriptomics method. The down-regulated ratio of the different GH isoforms was different in each cell type of pituitary adenoma relative to the controls. Those data suggested that the proportion of the different GH isoforms changed in each cell type adenoma compared to controls. Other researchers showed that the proportion of the circulating GH isoform significantly changed in pituitary adenomas and other pituitary diseases [47,48]. The proportional change of the different GH isoforms might have an important value in the clinical evaluation of human pituitary adenomas. Recently, we found that GH isoforms were derived from a variety of splicing variants and PTMs, including phosphorylation. The phosphorylation of endogenous GH in the human pituitary [18,34] provided us with new insights into the mechanisms of growth hormone that participate in the neuroendocrine signal pathways.

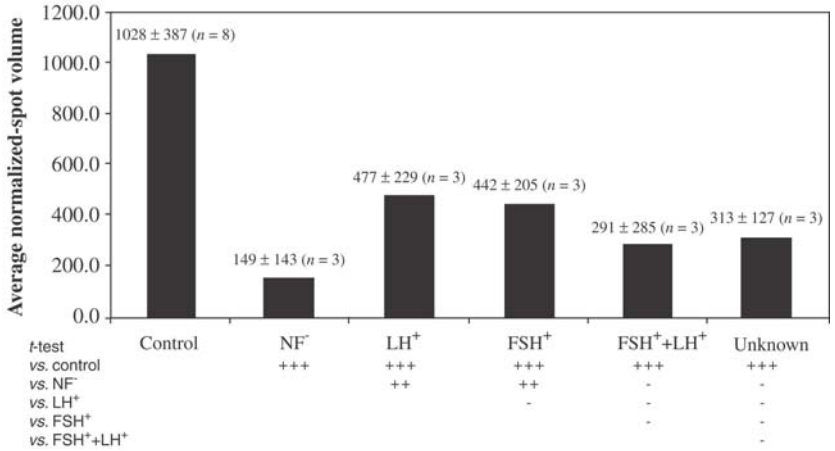


Fig. 9. Statistical analysis of secretagogin that was differentially expressed in human non-functional pituitary adenomas and controls (Zhan et al. [17]). The volume of each silver-stained 2D gel spot was quantified with PDQuest 2D-image software. Each 2D gel spot volume was normalized with the total density in a gel image. The normalized spot volume represented the content of each protein in a pituitary adenoma or control. The normalized spot volume was used to analyze the differential expression. The number of samples is labeled in the figure. $(+++)$ $p < 0.001$; $(++)$ $p < 0.01$; $(-)$ $p > 0.05$. Reproduced from Zhan et al. [17], with permission from Kluwer Academic Publishers, copyright 2003.

Prolactin is another important pituitary hormone. We detected six prolactin isoforms with our proteomics method ([32], Evans et al., in preparation). Similar to GH, each prolactin isoform was down-regulated in each cell type NF adenoma, with a different down-regulated ratio relative to controls. There was no significant expression change of the prolactin gene at the protein and mRNA levels in the prolactinoma relative to controls; that finding is consistent with a prolactinoma's monoclonal composition in origin. However, the proportion of six prolactin isoforms changed in prolactinomas compared to controls. Other researchers showed that glycosylation is an important prolactin modification [49] that produces different isoforms; glycosylation of human prolactin may down-regulate its hormone bioactivity and promote its metabolic clearance [50]. Other studies showed that the main variant of prolactin was the non-glycosylated form of PRL in human prolactinomas [51]; that finding is consistent with our result that some prolactin isoforms (Evans et al., in preparation) were down-regulated in prolactinomas compared to controls. Thus, we speculated that the prolactin isoform in those down-regulated spots is possibly a glycosylated prolactin.

Therefore, our detailed study of the different isoforms and PTMs of GH and prolactin could lead to new insights into the clinical importance of pituitary hormones in the formation of a pituitary adenoma (see Section 2). The ratio of the change of each isoform in human pituitary tissue has value for clinical research.

For example, could a change in the ratios of the prolactin isoforms be a potential index for diagnostics and prognostics of a prolactinoma patient? Could an alteration of circulating GH isoforms be a potential index for diagnostics and prognostics of a GH adenoma patient?

5. Future trends

5.1. Combination of gel-proteomics and non-gel quantitative proteomics

Proteomics generally includes gel-based methods and non-gel-based methods. Each method has its unique set of advantages and disadvantages, including detection sensitivity, the total separation of all proteins, dynamic range, and a limited ability to detect all proteins in a proteome. 2DGE is currently the only technique that can be routinely applied for the parallel quantitative expression profiling of large sets of complex protein mixtures from human tissues, and that can deliver a map of intact proteins that reflects any change in the protein expression level, isoforms, or PTMs. 2DGE is in contrast to LC–MS/MS methods, which analyze peptides, where M_r and *p*I information is lost, and where stable isotope-labeling is required for quantitative analysis [52]. For human tissue comparative proteomics analysis, non-gel methods are not yet sufficiently reproducible; therefore, gel-based methods are imperative to study a complex human tissue proteome, and to archive the proteome from precious human post-surgical pituitary adenoma tissue samples. However, LC–MS/MS quantitative proteomics systems might overcome some of the limitations of gel-based comparative proteomics such as the location, detection, and characterization of low-abundance, extremely acidic ($pI < 3.5$)/basic ($pI > 8$), and extremely high-mass (>150 kDa)/low-mass (<10 kDa) proteins [14]. Therefore, the combination of gel and non-gel-proteomics is needed to study human pituitary proteomes.

5.2. Comparative proteomics studies of PTM proteins in human pituitary adenomas

The PTM of a protein is an important factor in a cell and in proteomics [53], and cannot be detected with genomics or transcriptomics methods due to the many factors that are involved in the complex, multistep process DNA $\rightarrow$ mRNA $\rightarrow$ protein. A single gene could generate multiple gene products, PTM occurs in the majority of proteins, and the covalent alteration of a protein is not coded by the gene. A PTM is an important mechanism to maintain the structure of a protein, and to perform the wide range of functions of a protein. The number of documented protein PTMs is 351 (January 2005) in a database of protein PTM (http://www.abrf.org/index.cfm/dm.home?AvgMass=all).

Because of the important roles of phosphorylation, glycosylation, and nitration in the human body, phosphoproteomics [54], glycoproteomics [55,56], and nitroproteomics [19–21] have all emerged as rapidly increasing fields in proteomics. We have qualitatively analyzed phosphoproteins [18] and nitroproteins [19–21] in human pituitary control or adenoma proteomes. However, global comparative phosphoproteomics and comparative nitroproteomics are needed to discover any differentially expressed phosphoproteins and nitroproteins that might be related to human pituitary adenomas.

5.3. Integration of proteomics and transcriptomics to study human pituitary adenomas

Even though a proteome is much more complex than either a transcriptome or a genome due to a wide array of PTMs, protein translocations, protein–protein interactions, protein regulations, etc., the proteome, transcriptome, and genome are highly complementary systems. It is important to compare genomic, transcriptomic, and proteomic data to comprehensively clarify the basic molecular mechanisms that participate in the formation of a pituitary adenoma, and to discover any specific biomarkers for an early stage diagnosis, therapeutics, and prognosis of a human pituitary adenoma. Also, the combination of those three diverse methodologies, which has been referred as "operomics" [57,58], is necessary because each method has its own unique advantages and disadvantages, and each method provides unique molecular information. Recently, a combined comparative proteomics and comparative transcriptomics study was performed in human pituitary NF adenomas and prolactinomas. Nine genes demonstrated consistent changes at the mRNA and protein levels in human pituitary adenomas (Table 1). An expanded integrated "omics" study will be needed for the other types of pituitary adenomas.

5.4. Protein chips coupled with mass spectrometry to study human pituitary adenomas

A protein biochip is the counterpart of the array technology in the genomics field. Currently, the Ciphergen's ProteinChip® array surface-enhanced laser desorption-ionization (SELDI) MS system is available [59]. Chips with a broad range of binding properties, including immobilized metal affinity capture, and with biochemically characterized surfaces, such as antibodies and receptors, form the core of SELDI. Once the target proteins are captured on a SELDI protein biochip array, the proteins are detected with MALDI–TOF MS. A retentate map (proteins retained on the chip) is generated in which the individual protein is displayed as separate peaks on the basis of their mass-to-charge ratio (m/z). ProteinChip SELDI–MS could possibly be used to identify known biomarkers in a cancer, and to discover any potential markers that are either over- or under-expressed in a cancer. Protein

chip-based MS techniques could facilitate the screening, diagnostics, and prognostics of human pituitary adenomas. In order for SELDI to be an accurate method, it is crucial to rigorously and accurately regulate the physicochemistry that occurs on a SELDI probe surface. For example, one must calculate the ratio of the total number of moles of ionizable groups in all of the proteins in a sample to the total number of moles of interacting sites on the surface of a SELDI probe.

6. Conclusions

The elucidation of the molecular mechanisms that participate in the formation of a pituitary adenoma, and the discovery of any specific biomarker, are critical goals in any human pituitary adenoma study. Proteomics is valuable in that type of study because the proteome reflects the intrinsic genetic program of the cell and the impact of its immediate environment; therefore, a proteome could be used to identify potential biomarkers for adenoma early diagnosis, monitor disease progression, and identify therapeutic targets. Human pituitary proteomics has made important progress in several different areas [14]. The 2DGE-based comparative proteomics analysis system has been optimized and *established* for a human pituitary adenoma study. Some significant pituitary proteomics data have been obtained: a 2D gel protein reference map [12,15], the heterogeneity of a human pituitary proteome [16], the DEP profile associated with pituitary adenomas ([2,32], Evans et al., in preparation), a potential biomarker [17], the qualitative analysis of the phosphoproteome and nitroproteome in the human pituitary [18–21], and the integration of proteomic and transcriptomic data in human pituitary NF adenoma [2,32].

The overall progress of human pituitary proteomics is encouraging. In terms of the fact that the pituitary is the most well-protected tissue in the body, that the pituitary is the most critical neuroendocrine regulatory center, that pituitary proteomics is still in its infancy, and that proteomics is becoming an active field with a significant impact on neuroendocrinology, many issues of human pituitary proteomics remain to be improved and developed to their full potential for its clinical application. The current 2DGE method will remain one of the most suitable approaches to systematically characterize the differential proteome in the different cell types of human pituitary adenomas because of its wide availability, excellent reproducibility, ease of use, effectiveness, and low cost. Non-gel quantitative proteomics will complement gel-proteomics methods to expand the differential proteome coverage in a human pituitary adenoma. The comparative proteomics studies of PTM proteins will directly provide functional DEPs related to human pituitary adenomas. The integration of comparative proteomics and transcriptomics will provide global insights into the molecular processes of each DEP in human pituitary adenomas, and will provide important clues to the discovery of potential biomarkers for the clinical evaluation and accurate molecular treatment of human pituitary tumors.

Acknowledgments

The authors gratefully acknowledge the financial assistance (to DMD) from the National Institutes of Health (NS 42843, RR 10522, RR 14593, RR 016679, and DA) and the National Science Foundation (DBI 9604633). The scientific contributions of Francesco Giorgianni, Sarka Beranova-Giorgianni, Haibao Wan, Xianglin Yuan, Nelson M. Oyesiku, and Chheng-Orn Evans are acknowledged.

References

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J.; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921 (2001).

2. Desiderio, D.M. and Zhan, X., The human pituitary proteome: the characterization of differentially expressed proteins in an adenoma compared to a control. *Cell. Mol. Biol.*, **49**, 689–712 (2003).
3. Wilkins, M.R., Sanchez, J.C., Golley, A.A., Appel, R.D. HumpherySmith, I., Hochstrasser, D.F. and Williams, K.L., Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.*, **13**, 19–50 (1996).
4. Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.*, **19**, 1720–1730 (1999).
5. Anderson, L. and Seilhamer, J., A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, **18**, 533–537 (1997).
6. Godovac-Zimmermann, J. and Brown, L.R., Perspectives for mass spectrometry and functional proteomics. *Mass Spectrom. Rev.*, **20**, 1–57 (2001).
7. Adam, G.C., Sorensen, E.J. and Cravatt, B.F., Chemical strategies for functional proteomics. *Mol. Cell Proteomic.*, **1**, 781–791 (2002).
8. Melmed, S., Mechanisms for pituitary tumorigenesis: the plastic pituitary. *J. Clin. Invest.*, **112**, 1603–1618 (2003).
9. Alexander, J.M., Biller, B.M., Bikkal, H., Zervas, N.T., Arnold, A. and Klibanski, A., Clinically nonfunctioning pituitary tumors are monoclonal in origin. *J. Clin. Invest.*, **86**, 336–340 (1990).
10. Herman, V., Fagin, J., Gonsky, R., Kovacs, K. and Melmed, S., Clonal origin of pituitary adenomas. *J. Clin. Endocrinol. Metab.*, **71**, 1427–1433 (1990).
11. Asa, S.L., The pathology of pituitary tumors. *Endocrinol. Metab. Clin. North Am.*, **28**, 13–43 (1999).
12. Zhan, X. and Desiderio, D.M., A reference map of a human pituitary adenoma proteome. *Proteomics*, **3**, 699–713 (2003).
13. Medzihradszky, K.F., Campbell, J.M., Baldwin, M.A., Falick, A.M., Jahasz. P., Vestal, M.L. and Burlingame, A.L., The characteristics of peptide collision-induced dissociation using a high-performance MALDI–TOF/TOF tandem mass spectrometer. *Anal. Chem.*, **72**, 552–558 (2000).
14. Zhan, X. and Desiderio, D.M., Comparative proteomics analysis of human pituitary adenomas: current status and future perspectives. *Mass Spectrom. Rev.*, **24**, 783–813 (2005).
15. Beranova-Giorgianni, S., Giorgianni, F. and Desiderio, D.M., Analysis of the proteome in the human pituitary. *Proteomics*, **2**, 534–542 (2002).
16. Zhan, X. and Desiderio, D.M., Heterogeneity analysis of the human pituitary proteome. *Clin. Chem.*, **49**, 1740–1751 (2003).
17. Zhan, X., Evans, C.O., Oyesiku, N.M. and Desiderio, D.M., Proteomics and transcriptomics analyses of secretagogin down-regulation in human non-functional pituitary adenoma. *Pituitary*, **6**, 189–202 (2003).
18. Giorgianni, F., Beranova-Giorgianni, S. and Desiderio, D.M., Identification and characterization of phosphorylated proteins in human pituitary. *Proteomics*, **4**, 587–598 (2004).
19. Zhan, X. and Desiderio, D.M., The human pituitary nitroproteome: detection of nitrotyrosylproteins with two-dimensional Western blotting, and amino acid sequence determination with mass spectrometry. *Biochem. Biophys. Res. Commun.*, **325**, 1180–1186 (2004).
20. Zhan, X. and Desiderio, D.M., Nitroproteins from a human pituitary adenoma tissue discovered with a nitrotyrosine affinity column and tandem mass spectrometry. *Anal. Biochem.*, **354**, 279–289 (2006).
21. Zhan, X. and Desiderio, D.M., Linear ion-trap mass spectrometric characterization of human pituitary nitrotyrosine-containing proteins. *Int. J. Mass Spectrom.*, **259**, 96–104 (2007).
22. Nelson, R.W., Nedelkov, D. and Tubbs, K.A., Biosensor chip-mass spectrometry: a chip-based proteomics approach. *Electrophoresis*, **21**, 1155–1163 (2000).

23. Burgus, R., Dunn, T.F., Desiderio, D.M., Ward, D.N., Vale, W. and Guillemin, R., Characterization of ovine hypothalamic hypophysiotropic TSH-releasing factor. *Nature*, **226**, 321–325 (1970).
24. Yalow, R.S. and Berson, S., Immunoassay of endogenous plasma insulin in man. *J. Clin. Invest.*, **39**, 1157–1175 (1960).
25. Yalow, R.S., Radioimmunoassay: a probe for the fine structure of biologic systems. *Science*, **200**, 1236–1245 (1978).
26. Zhu, H., Bilgin, M. and Snyder, M., Proteomics. *Annu. Rev. Biochem.*, **72**, 783–812 (2003).
27. Ezzat, S., Pituitary tumor pathogenesis—the hunt for novel candidate genes continues. *J. Clin. Endocrinol. Metab.*, **88**, 5116–5118 (2003).
28. Liu, E.T. and Karuturi, K.R., Microarrays and clinical investigations. *N. Engl. J. Med.*, **350**, 1597–1599 (2004).
29. Herbert, B., Advances in protein solubilisation for two-dimensional electrophoresis. *Electrophoresis*, **20**, 660–663 (1999).
30. Zhan, X. and Desiderio, D.M., Differences in the spatial and quantitative reproducibility between two second-dimensional gel electrophoresis systems. *Electrophoresis*, **24**, 1834–1846 (2003).
31. Zhan, X. and Desiderio, D.M., Spot volume vs. amount of protein loaded onto a gel: a detailed, statistical comparison of two gel electrophoresis systems. *Electrophoresis*, **24**, 1818–1833 (2003).
32. Moreno, C.S., Evans, C.O., Zhan, X., Okor, M., Desiderio, D.M. and Oyesiku, N.M., Novel molecular signaling and classification of human clinically nonfunctional pituitary adenomas identified by gene expression profiling and proteomic analyses. *Cancer Res.*, **65**, 10214–10222 (2005).
33. Beranova-Giorgianni, S., Zhao, Y., Desiderio, D.M. and Giorgianni, F., Phosphoproteomic analysis of the human pituitary. *Pituitary*, **9**, 109–120 (2006).
34. Zhan, X., Giorgianni, F. and Desiderio, D.M., The proteomics analysis of growth hormone isoforms in human pituitary. *Proteomics*, **5**, 1228–1241 (2005).
35. Lania, A., Montovani, G. and Spada, A., Genetics of pituitary tumors: focus on G-protein mutations. *Exp. Biol. Med.*, **228**, 1004–1007 (2003).
36. Ezzat, S., Zheng, L., Zhu, X.F., Wu, G.E. and Asa, S.L., Targeted expression of a human pituitary tumor-derived isoform of FGF receptor-4 recapitulates pituitary tumorigenesis. *J. Clin. Invest.*, **109**, 69–78 (2002).
37. Alexander, J., Tumor suppressor loss in pituitary tumors. *Brain Pathol.*, **11**, 342–355 (2001).
38. Evans, C.O., Young, A.N., Brown, M.R., Brat, D.J., Parks, J.S., Neish, A.S. and Oyesiku, N.M., Novel patterns of gene expression in pituitary adenomas identified by complementary deoxyribonucleic acid microarrays and quantitative reverse transcription-polymerase chain reaction. *J. Clin. Endocrinol. Metab.*, **86**, 3097–3107 (2001).
39. Yu, R. and Melmed, S., Oncogene activation in pituitary tumors. *Brain Pathol.*, **11**, 328–341 (2001).
40. Wagner, L., Oliyarnyk, O., Gartner, W., Nowotny, P., Groeger, M., Kaserer, K., Waldhausl, W. and Pasternack, M.S., Cloning and expression of secretagogin, a novel neuroendocrine- and pancreatic islet of Langerhans-specific Ca^{2+}-binding protein. *J. Biol. Chem.*, **275**, 24740–24751 (2000).
41. Gartner, W., Lang, W., Leutmetzer, F., Domanovits, H., Waldhausl, W. and Wagner, L., Cerebral expression and serum detectability of secretagogin, a recently cloned EF-hand Ca^{2+}-binding protein. *Cereb. Cortex*, **11**, 1161–1169 (2001).
42. Desiderio, D.M., Kusmierz, J.J., Zhu, X., Dass, C., Hilton, D., Robertson, J.T. and Sacks, H.S., Mass spectrometric analysis of opioid and tachykinin neuropeptides in non-secreting and ACTH-secreting adenomas. *Biol. Mass Spectrom.*, **22**, 89–97 (1993).

43. Zhu, X. and Desiderio, D.M., Methionine enkephalin-like immunoreactivity, substance P-like immunoreactivity and beta-endorphin-like immunoreactivity post-mortem stability in rat pituitary. *J. Chromatogr.*, **616**, 175–187 (1993).
44. Zhu, X. and Desiderio, D.M., Effects of space flight stress on proopiomelanocortin, proenkephalin A, and tachykinin neuropeptidergic systems in the rat posterior pituitary. *Life Sci.*, **55**, 347–350 (1994).
45. Zhu, X., Robertson, J.T., Sacks, H.S., Dohan Jr., C., Tseng, J.L. and Desiderio, D.M., Opioid and tachykinin neuropeptides in prolactin-secreting human pituitary adenomas. *Peptides*, **16**, 1097–1107 (1995).
46. Desiderio, D.M., Mass spectrometric analysis of neuropeptidergic systems in the human pituitary and cerebrospinal fluid. *J. Chromatogr. B Biomed. Sci. Appl.*, **731**, 3–22 (1999).
47. Boguszewski, C.L., Johannsson, G., Bengtsson, B.A., Johansson, A., Carlsson, B. and Carlsson, L.M., Circulating non-22-kilodalton growth hormone isoforms in acromegalic men before and after transsphenoidal surgery. *J. Clin. Endocrinol. Metab.*, **82**, 1516–1521 (1997).
48. Boguszewski, C.L., Jansson, C., Boguszewski, M.C., Rosberg, S., Carlsson, B., Albertsson-Wikland, K. and Carlsson, L.M., Increased proportion of circulating non-22-kilodalton growth hormone isoforms in short children: a possible mechanism for growth failure. *J. Clin. Endocrinol. Metab.*, **82**, 2944–2949 (1997).
49. Gambino, G.M., Beck-Peccoz, P., Borgato, S., Faglia, G., Spada, A. and Persani, L., Bioactivity and glycosylation of circulating prolactin in various physiological and pathological conditions. *Pituitary*, **2**, 225–231 (1999).
50. Hoffmann, T., Penel, C. and Ronin, C., Glycosylation of human prolactin regulates hormone bioactivity and metabolic clearance. *J. Endocrinol. Invest.*, **16**, 807–816 (1993).
51. Hoffmann, T., Gunz, G., Brue, T., Jaquet, P. and Ronin, C., Prolactin isoforms secreted by human prolactinomas. *Horm. Res.*, **38**, 164–170 (1992).
52. Görg, A., Weiss, W. and Dunn, M.J., Current two-dimensional electrophoresis technology for proteomics. *Proteomics*, **4**, 3665–3685 (2004).
53. Mann, M. and Jensen, O.N., Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, **21**, 255–261 (2003).
54. Kalume, D.E., Molina, H. and Pandey, A., Tackling the phosphoproteome: tools and strategies. *Curr. Opin. Chem. Biol.*, **7**, 64–69 (2003).
55. Vosseller, K., Wells, L. and Hart G.W., Nucleocytoplasmic O-glycosylation: O-GlcNAc and functional proteomics. *Biochimie*, **83**, 575–581 (2001).
56. Wells, L., Vosseller, K., Cole, R.N., Cronshaw, J.M., Matunis, M.J. and Hart, G.W., Mapping sites O-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. *Mol. Cell Proteomic.*, **1**, 791–804 (2002).
57. Hanash, S.M., Operomics: molecular analysis of tissues from DNA to RNA to protein. *Clin. Chem. Lab. Med.*, **38**, 805–813 (2000).
58. Hanash, S.M., Bobek, M.P., Rickman, D.S., Williams, T., Rouillard, J.M., Kuick, R. and Puravs, E., Integrating cancer genomics and proteomics in the post-genome era. *Proteomics*, **2**, 69–75 (2002).
59. Merchant, M. and Weinberger, S.R., Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis*, **21**, 1164–1167 (2000).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 20

Mass spectrometry of proteinous allergens inducing human diseases

MARTINA MARCHETTI[a], JASMIN HIRSCHMANN[a], ELISABETH FÖRSTER-WALDL[b], and GÜNTER ALLMAIER[a,*]

[a]*Institute of Chemical Technologies and Analytics, Vienna University of Technology, Vienna, Austria*
[b]*Department of Paediatrics and Juvenile Medicine, Medical University of Vienna, Vienna, Austria*

1. Introduction

Allergies are unpleasant ailments and actually, in some cases, acute life-threatening diseases. The term "allergy" (from Greek *allos*—other and *ergon*—reaction) was first mentioned 100 years ago by an Austrian pediatrician named Clemens von Pirquet in 1906 [1]. He noticed that some of his vaccinated patients had more severe reactions to ongoing treatment and followed this to be a response to outside elicitors. Since this observation hypersensitivities were thought to result from unpredictable actions of immunoglobulins and in 1963 Gell and Coombs [2] proposed a classification of hypersensitivity reactions dividing these reactions into

*Corresponding author. Tel.: +43 158801 15160; Fax: +43 158801 15199.
E-mail: guenter.allmaier@tuwien.ac.at (G. Allmaier).

four subgroups (types I to IV) representing the main strategies a body uses to combat different classes of infectious agents. These classes, extended by a fifth one (type V) introduced in 2003 by Rajan [3], are widely accepted today. To understand the mechanisms of "allergic reactions" the corresponding elicitors have to be investigated more precisely. On this account great effort has already been put into the understanding of hypersensitivities against natural rubber latex products since the number of affected people has increased dramatically after the first case was reported in 1979 [4]. It was found that the most frequent triggers are accelerators, color pigments, preserving agents, or other organic components used during the production processes. Today, gas chromatography combined with mass spectrometry (GC–MS) is typically used for the determination and quantification of these allergological relevant compounds in disposable gloves [5]. Nevertheless, it was clear that not only these molecules but also proteins are responsible for immune responses. The principal proteinous allergy elicitor, a 14 kDa protein named rubber elongation factor (REF), was first described in 1993 on the amino acid sequence level after tryptic digestion and Edman degradation [6], and in 1989 fast-atom bombardment (FAB) mass spectrometry was first applied to determine molecular weights of enzymatically obtained peptides from REF [7]. Only recently, matrix-assisted laser desorption/ionization (MALDI) and nanoelectrospray ionization (nano-ESI) mass spectrometry in combination with enzymatic digestion after gel electrophoresis, a typical proteomics approach, were applied to show that REF besides a truncated form is present in commercially available latex gloves [8].

The predominant elicitor in a latex protein fraction with proteins smaller than 10 kDa is the so-called hevein, a 4.7 kDa polypeptide. Its IgE binding capacity was first demonstrated by Alenius et al. [9] who also determined the average mass of this allergen to be 4719 $\pm$ 1.9 Da by applying mass spectrometry. Chen et al. [10] detected two more minor components indicating the existence of hevein variations with additional Ser and Ser–Gly by MALDI-TOF-MS and ESI-MS. These findings were in good agreement with previously published ESI–quadrupole MS data reporting a ragged C-terminus of hevein and pseudohevein, a hevein analog with six amino acid replacements and several additional Gly residues at the C-terminus [11].

Individuals with latex protein allergy often exhibit reactions to plant-derived food and fresh fruits, such as avocado or banana, too. N-terminal hevein-like domains seem to be responsible for these mediated reactions in the so-called latex-fruit syndrome [12].

Besides these hevein-like domains, lipid transfer proteins (LTPs) play an important role in food allergy [13–15] and they have therefore been suggested as model plant food allergens. Only recently, the first citric LTPs were isolated from oranges and lemons with molecular masses of 9610 and 9618 Da, respectively, determined by MALDI-TOF-MS [16]. LTPs have also been identified as the major apricot allergens [17]. The molecular masses of the intact proteins were deduced

by ESI-quadrupole ion trap (QIT)-MS and were found to be 9170.4 and 7238.0 Da, respectively, which is in good agreement with the data published earlier describing the complete primary structure of a 9 kDa allergen in apricots [18] where liquid chromatography–mass spectrometry (LC–MS) analyses were used for mass determination of the purified intact protein and some selected tryptic peptides for verification of the most probable amino acid sequence.

The cross-reactivity between various fruits and different pollen allergens is a well-known fact and has already been studied in the 80s of the last century [19,20] using radioallergosorbent tests (RASTs). One of the first pollen allergens characterized by mass spectrometry is betv1, the major birch pollen allergen [21]. Plasma desorption mass spectrometry (PD-MS) was used to confirm the primary structures of the intact purified protein, of all potential isoforms and some selected proteolytic peptides and to investigate any possible posttranslational modifications.

Since mass spectrometry got applicable to larger biomolecules [22,23], a big effort has been put into the identification of further allergy elicitors and into the structure elucidation of major allergens relevant for mankind. Special attention has been drawn on plants responsible for summer hay fever such as birch, willow, elder, hazelnut, grass pollen, mugwort, or chrysanthemum. One of the most unstudied trees whose plant parts are widely used is the elderberry tree (*Sambucus nigra*).

Elderberry trees grow ubiquitously in regions of moderate climate, blossom during May to August with the main season in June and July [24]. *S. nigra* has a long-term history in folk medicine to treat influenza, common cold, or sinusitis [25–27] and its antiinfectious activity has already been proven in a double-blind placebo-controlled clinical study [28]. A number of proteins, maybe related to the mentioned medical applications, extracted from various parts of the tree, as bark, leaves, or fruits, have been intensively investigated and characterized. Certain attention has been drawn not only to their lectin activities [29] but also to their application in medical diagnosis [30–32].

Up to now little is known about the frequency of allergic reactions to elderberry because its flowering season overlaps with the seasonal allergy to grass and weed pollen [33]. Furthermore, patients suffering from summer hay fever are often not sensitized to a single allergenic plant, but seem to be polysensitized. Even though type I allergic reactions after inhalative contact with pollen of the elderberry plant have been suspected by clinicians during the last decade, the diagnosis of allergy against elderberry was disregarded.

In everyday life, allergic reactions may occur not only after inhalative contact but also after ingestion since the fruits, flowers, or leaves of the elderberry tree are widely used in nutrition (juices, sparkling wine, jam) or herbal medicine (teas). Although the cDNA of a 17.6 kDa protein from elderberry tissue with high homology to well-characterized food allergens of plants has been cloned and expressed, a 5 kDa protein with sequence homology to hevein, a major latex allergen, was cloned and expressed too [34], verification of the existence of type I allergy

against this plant has been achieved just once until now [35]. Detection and characterization of the allergens in the relevant plant materials have not been performed so far.

Along with other forms of allergic reactions, food allergy seems to be on a continuous rise. Allergies are usually caused by the proteinous components in food, within which most of the protein triggers seem to belong to one of three structurally related superfamilies: the prolamin superfamily, the cysteine proteases superfamily, or the cupin superfamily [36,37]. The prolamin superfamily contains alcohol-soluble storage proteins of cereals, the cysteine proteases—also known as the papain-like family—bear conserved glutamine, cysteine, histidine, and asparagine residues at the active site, and the cupin superfamily comprises proteins with one basic and one double-stranded α-helix domain. The most widely investigated cupins are the 7S and 11S globulin storage proteins. Besides the allergy-relevant structural characteristics of the polypeptide chains, the 7S globulin subunits are very often glycosylated. This is of special interest because carbohydrates, containing core-fucose and core-xylose glycoepitopes, have been made responsible for IgE cross-reactions [38–41].

Besides these three major protein classes, plant lectins, also known as plant agglutinins, were shown to react with the carbohydrate moieties of IgE, inducing histamine release and thus causing allergy-like symptoms [42,43]. One 31 kDa peanut agglutinin was already identified as a lectin binding specifically to the IgE epitope [44].

In this chapter we explore the role mass spectrometry plays in the identification of particular allergens and of the reaction the human body stages to the flowers and other parts of the elderberry tree. The presented approach can be viewed as a case study for the use of mass spectrometry in these endeavors.

1.1. Highlights for medical professionals

Summer hay fever sounds harmless, but is a severe illness. The number of people suffering from summer hay fever rises year after year. Up to 20% of the population in Europe and North America are estimated to suffer from hay fever though exact figures are hard to determine since many cases go unreported. Sneezing, wheezing, increased production of nasal mucosa and lachrymal fluid, plugged ears, and itching of the nose, throat, mouth, ear canal, eyes, and/or skin are rather harmless though very unpleasant symptoms, whereas inflammation of the gastrointestinal mucosa, sickness, asthma, and eczema can lead to severe problems and people can sometimes no longer conduct their everyday life.

According to the classification of Gell and Coombs [2], summer hay fever is a type I allergy for which symptoms occur immediately or up to 1 h after contact when antigens, e.g., pollen, have entered the body and activated mast cells. These types of immune defense cells carry antibodies (IgEs) attached to their surfaces

and therefore antigens can bind to them, resulting in the release of histamine, causing blood vessel dilation and/or narrowing of the respiratory tract.

Some patients suffering from allergy-mediated rhinoconjunctivitis and dyspnea reported to reveal these symptoms after close contact with elderberry flowers. Other patients complained about wheezing after drinking juices prepared by extracting elderberry blossoms. Type I allergy triggered by different products of the elderberry tree has been suspected by clinical allergologists but has not been further followed up. This is because, on one hand, many of the affected persons are polysensitized and, on the other hand, the flowering seasons of the trees and bushes completely overlap with more common elicitors for type I allergy such as mugwort, basswood, barley, wheat, pine, chestnut, acacia, birch, spruce, sorrel, rape, or willow [24,33]. Thus, type I allergy against elderberry can easily be overseen or underestimated. Furthermore, one should consider that elderberry plants belong to the botanic family of *Caprifoliaceae,* endemic plants, which have never been associated with allergies, such as the *Betulaceae, Rosaceae, Poaceae,* and *Compositae*. Therefore, the chance of a correct clinical diagnosis is reduced.

Recently, a study has shown that IgE from eight patient's sera binds specifically to proteins present in various elderberry tissues (pollen, berry extract, flowers) [35]. A protein with an apparent molecular weight of 33.2 kDa exhibited the predominant reaction in serological experiments, using SDS-PAGE and Western blots, in which the reaction of pollen extract exceeded the binding specificities of the flower and berry extract, respectively. Cross-reactivity studies showed only partial inhibition of specific IgE binding by birch and mugwort, representatives of the most important pollen allergen sources. Antibodies from mice immunized with elderberry flower extract exclusively gave positive results for elderberry but not for birch, grass, or mugwort pollen. MALDI and nano-ESI mass spectrometric studies showed that the amino acid sequence of the potential allergen has a very high homology to type 2 ribosome-inactivating proteins (RIPs). Only recently, the general antiinfectious activity of RIPs could be shown [45] and thus became of special interest for the therapy of infectious diseases of plants [46,47] and for diverse applications in human medicine such as abortification [48], immunotoxins [27,49], or antihuman immunodeficiency virus (HIV) agents [50–53]. Moreover, the antiproliferation activity of this very special class of proteins promoted research related to their application in therapy of malignancy [54–56].

1.2. Highlights for chemists

Only recently, a very interesting protein family, namely RIPs, has been isolated from *S. nigra* [57–60]. RIPs are RNA N-glycosidases inactivating ribosomes through a site-specific deadenylation of the large ribosomal RNA [61]. In addition to this, some RIPs have been reported to have superoxide dismutase [62,63] and phospholipase [64] type of activities. It is supposed that RIPs are defense-related

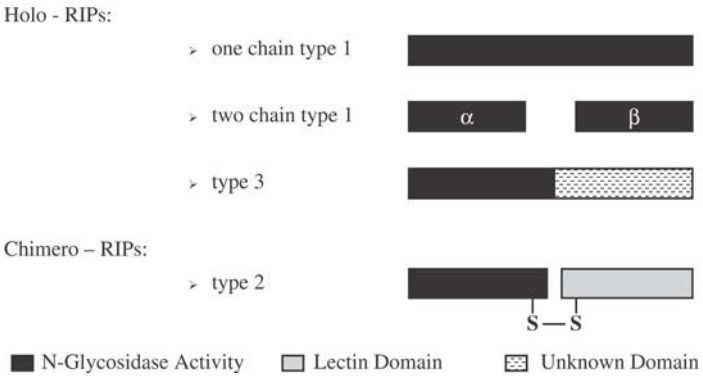


Fig. 1. Schematic representation of the structure of different types of ribosome-inactivating proteins (RIPs). Type 1 RIPs consist of one or two smaller polypeptide chains, held together by noncovalent interactions, featuring N-glycosidase activity. Type 3 RIPs consist of a polypeptide chain harboring N-glycosidase activity and a second domain of unknown function. Both types belong to the superfamily of holo-RIPs. Type 2 RIPs are representatives of chimero-RIPs consisting of two structurally unrelated polypeptide chains linked by a disulfide bridge.

plant proteins that are also involved in the aging process [65]. Two major classes can be distinguished: holo-RIPs and chimero-RIPs (Fig. 1). Holo-RIPs, generally referred to as type I RIPs, consist of a single polypeptide chain of ~30 kDa, although the promoter sequence sometimes is processed into two shorter sequences held together by noncovalent interactions (two-chain type 1 RIPs). Chimero-RIPs comprise one or more amino acid backbones containing a N-glycosidase domain and a structurally different and functionally unrelated lectin domain linked by a disulfide bridge (type 2 RIPs). Type 3 RIPs, members of the holo-RIP family, again consist of just one single polypeptide chain though consisting of two different domains, a N-glycosidase domain and a domain whose function is not yet clarified.

Four years ago, only recently, it could be shown that products from the elderberry tree can be elicitors for type I allergy [35]. It could be demonstrated by inhibition experiments by means of denaturing SDS-PAGE and subsequent Western blotting that the allergenic component is the major proteinous component of elderberry flower extract and has an apparent molecular weight of 33.2 kDa. Implementing positive linear MALDI-TOF-MS experiments into the studies ascertained the proper molecular weight to be 66.6 kDa. Edman sequencing of the reduced 33.2 kDa allergen, after blotting onto polyvinylidene difluoride (PVDF) membranes from one- and two-dimensional gels (1D and 2D PAGE) as well as from the purified protein, yielded the first 13 N-terminal amino acids. Database search for homologous sequences using BLAST (basic local alignment search tool) [66] revealed partial sequence similarity to type 2 RIPs from *Sambucus ebulus* (dwarf elderberry) but not in the same high degree to RIPs of *S. nigra*. Therefore, *in-gel* tryptic

digestion after 1D gel electrophoresis of the 33.2 kDa and of the 66.6 kDa protein, the latter observable under nonreducing SDS-PAGE conditions, was performed and the resulting tryptic peptides were further sequenced by MS/MS experiments. Partial and complete sequences of 12 major peptides could be assigned by means of a hybrid multistage MALDI-QIT/reflectron instrument in the low-energy CID mode. These results could be corroborated by a multistage nano-ESI-QIT mass spectrometry and provided high confidence in the obtained sequence data. Based on these data, intensive bioinformatic data mining delivered the information that a high homology to RIPs from *S. nigra* is given. 2D PAGE showed that the protein at 33.2 kDa consists of more than one isoform. Up to five distinguishable protein spots could be detected and consecutive IgE binding experiments pointed out that not just one of these spots is responsible for the positive allergy testing but all of them bind IgE. In addition, protein bands at 28 kDa (pI 4.5–6) and 17 kDa (pI 6–6.5) gave positive serological response. The identification and final characterization of the allergen is a very crucial point for pinning down the elicitor for type I allergy in elderberry plants.

2. Methodology

Flowers from the elderberry plant were collected in the Vienna Woods and in Upper Austria. For protein extraction the flowers were grinded and subsequently shaken overnight in a potassium phosphate buffer, pH 7.2, at 4ºC as described previously [67]. Centrifugation at 40,000 $\times$ *g* removed the insoluble particles and the supernatant was further dialyzed against double-distilled water using a dialysis membrane with a molecular weight cutoff of 8 kDa. Samples obtained by this means were lyophilized and stored at -20ºC. Applying such a straightforward and easy sample preparation should result in samples free of low-molecular-weight substances, such as mono- and oligosaccharides, lipids, or anthocyans whereupon the protein composition should not be altered to a very high extent; however, the amount of low-molecular-weight proteins might be reduced. Molecular weight pattern of the samples was determined by 1D gel electrophoresis according to the protocol from Laemmli [68]. Briefly, an aliquot of the elderberry extract was heated in denaturing and nondenaturing Laemmli buffer before loading it to lanes of SDS-PAGE gels (10 and 15%) run in vertical gel apparatus applying 200 V constantly. Protein molecular weights were estimated by comparing the protein bands, visualized by Coomassie Brilliant Blue staining, to prestained molecular weight markers. Gel electrophoresis is a very fast method for estimating the protein profile of a sample, but molecular weight determination is often hampered due to the low accuracy ($\pm$20%) of the method.

For subsequent IgE immunoblotting of proteins onto nitrocellulose or PVDF membranes was performed. Unspecific binding of serum proteins to nitrocellulose

strips was blocked by soaking the nitrocellulose strips in 50 mM sodium phosphate buffer (pH 7.5) containing Tween and bovine serum albumin. Sera from allergic patients were diluted 1:10 in blocking buffer and blot strips were preincubated. IgE binding was specifically detected by ^{125}I-rabbit antihuman IgE. To check for cross-reactivity between elderberry pollen and pollen from other sources, such as birch, grass, or mugwort, sera from elderberry immunized mice, containing antibodies directed against the antigens of interest, were pooled in equal parts and applied in 1:100 dilutions. Bound murine antibodies were detected with ^{125}I-labelled antimouse Ig antibodies using autoradiography.

For direct clinical tests, nine patients reporting rhinoconjunctivitis or dyspnea, disease pattern associated with type I allergy, after inhalative or dietary contact with elderberry flowers were tested by skin prick tests (SPTs) and RAST in order to diagnose specific allergy against elderberry. Histamine hydrochloride was used as positive reference. SPT is the most common method to gain results within about 20 min where several suspected allergens can be tested at the same time, whereas RAST is a laboratory test performed *in vitro* measuring specific IgE antibodies in blood. Although the blood test is less sensitive and more time-consuming than skin testing, it is still very helpful in cases of extensive dermatitis, marked dermatographism, or in children younger than 4 years of age.

The patient's sera were furthermore tested for their IgE-binding profiles of electrophoretically separated extracts from elderberry blossoms, pollen, and fruits to identify the immunological-relevant allergen. To get information about the composition of each single protein band in 1D SDS-PAGE, 2D gel electrophoresis was applied. 2D PAGE separates proteins due to their different isoelectric points related to possible posttranslational modifications like glycosylation, phosphorylation, or sulfatation. Cell metabolism causes a great variability of such protein modifications, very often resulting in a great number of protein spots scattered in a quite narrow p*I* range. On this account the combination of 2D PAGE with immunological studies opens the possibility to get hold of serological-relevant isoforms since in many cases not only one protein species but also a pool of closely related structures is responsible for a positive immune reaction. Admittedly, separating the multitude of isoforms has a drawback too. Very often immunologically interesting proteins are low-abundant proteins. Separating the isoforms of the same protein lowers the protein concentration of each single isoform and hence the absolute concentration for protein detection. Sometimes to an extent that some isoforms cannot be detected anymore by conventional staining protocols making consecutive protein identification impossible.

Since protein concentration of the serologically interesting protein in the elderberry flower extract was high enough for further analysis, 2D PAGE was performed according to protocols from Görg et al. [69]. The first dimension, a 7 cm immobilized linear pH gradient (IPG) strip from pH 3–7, was rehydrated in an appropriate volume of sample buffer containing urea and CHAPS besides IPG buffers, dithiothreitol,

and traces of bromophenol blue. Sample application was performed during the rehydration step by dissolving a lyophilized sample aliquot in the rehydration solution. Isoelectric focusing was run at room temperature for 11 h applying 63 kVh in total. 10% bis–tris gel run with MES buffer was used for the second dimension and protein visualization was again performed by Coomassie Brilliant Blue staining. Serological reactivity on 2D PAGE maps was tested after transferring the proteins onto a nitrocellulose membrane and incubation with human serum IgE as described for 1D gels.

For protein identification of the allergen, 1D- and 2D-separated elderberry flower extracts were blotted onto PVDF membranes and stained with 0.1% methanolic Coomassie Brilliant Blue solution. Bands known to bind IgE were cut from the nylon membrane and subjected to an automatic Edman degradation system, a widespread method for sequential protein and peptide degradation to get information about the N-terminus of a biomolecule. The chemistry of Edman reaction is well understood and works with high yields (>90%). Nevertheless, amino acid sequence analysis still has some problems that make sequencing difficult or impossible, especially in the lower picomole range. The sample should ideally consist of one single protein that has to have a homogenous N-terminus for unambiguous protein identification, whose salt, free amino acid, and detergent content should also be kept very low. A fundamental problem for amino acid sequencing by Edman degradation is the N-terminal blockage of the protein since the chemical reactions need a free amine group at the N-terminus, but about 50% of the naturally occurring proteins are N-terminally modified. Besides this some amino acids are undetectable by this method (nonderivatized cysteines) or the harsh chemical conditions of Edman degradation partly destroy some residues (e.g., dehydration of serine and/or threonine, oxidation of lysine). Edman degradation is therefore a very useful automated method but it cannot replace mass spectrometry in terms of an extensive amino acid sequence analysis. In the case of the elderberry extract it has been possible to detect 13 N-terminal amino acids by this method.

The use of gel electrophoresis, immunoblots, and Edman degradation made the localization of the protein within the very complex protein composition of the elderberry flower extract possible and gave first results for the N-terminus of the allergen. Therefore, emphasis could be shifted on identifying the allergen by a mass spectrometric approach. The very high salt content of the original elderberry flower extract was a drawback for analysis (e.g., mass spectrometry) and the protein of interest had to be further purified prior to manipulation. For this purpose the elderberry extract was loaded onto a Sephadex G50 chromatography column equilibrated in phosphate buffer (pH 7.4), containing glycerol, dithiotreitol, and sodium azide. Protein concentrations of fractions collected in minute intervals were controlled *online* using a UV detector at 280 nm. An extinction coefficient of 0.785 for 1 mg/mL solution was taken for calculations according to Gill and von Hippel [70]. The sample containing the main allergen, controlled by SDS-PAGE, Coomassie

staining, and IgE immunoblotting, was subsequently applied to a reversed-phase chromatographic system using a RP-C4 column. Aqueous trifluoroacetic acid (0.1% TFA) and isopropanol were applied as mobile phases. A UV detector controlled the protein elution at 280 nm whereupon the protein of interest eluted in a single peak. For exact molecular mass determination, an aliquot of the fraction gained from reversed-phase liquid chromatography was lyophilized and reconstituted in 0.1% TFA. MALDI-TOF-MS experiments were carried out on a linear TOF instrument equipped with a 337 nm N_2 laser, a 0.7 m flight tube (MALDI IV, Shimadzu Biotech Kratos Analytical, Manchester, UK) in the positive-ion mode with the extraction voltage set to 24 kV. The purified sample was prepared with sinapic acid using the dried-droplet method [71]. Mass calibration was performed externally using the doubly and singly charged molecular ions of bovine serum albumin (prepared in the same manner as the samples) whereby a mass accuracy of $\pm 0.1\%$ for the major allergenic component could be achieved.

To yield information about the amino acid sequence *in-gel* digestion applying trypsin and subsequent MALDI-TOF-MS, MALDI-MS/MS and nano-ESI-MS/MS experiments were performed. Enzymatic degradation was accomplished according to protocols suitable for mass spectrometry [72]. The Coomassie-stained protein band of the interesting allergen was excised from the gel after SDS-PAGE, cut into small cubes, and destained by consecutive washing steps including double-distilled water and acetonitrile. Furthermore, disulfide bonds of the protein were reduced using dithiothreitol at 56ºC for 45 min and further derivatized by iodoacetamide for 30 min in the dark. Afterwards the gel bands were soaked in an enzymatic solution containing 12.5 ng trypsin (sequencing grade from bovine pancreas) and digestion was carried out overnight at 37ºC. The very next day the supernatant containing the first fraction of the tryptic peptides was collected and the remaining peptides were extracted from the gel pieces by incubation with acetonitrile and 0.1% TFA. The extracts and the first removed supernatant were pooled, lyophilized, reconstituted in a 0.1% TFA solution, and directly desalted using Zip-Tip technology [73]. Since only volatile buffers such as ammonium hydrogen carbonate and solvents of high purity with only minimal contaminations of inorganic salts (e.g., sodium, potassium) were used, this microscaled desalting step is highly efficient for gaining samples best suited for subsequent mass spectrometry concerning their salt content.

One aliquot of the tryptic digest was submitted to peptide mass fingerprint analysis. α-Cyano-4-hydroxy-cinnamic acid dissolved in acetone was used as matrix using the thin-layer sample preparation method [71]. Mass spectra were acquired on a linear MALDI-TOF-MS instrument (AXIMA LNR, Shimadzu Biotech Kratos Analytical, Manchester, UK) in the positive-ion mode, also equipped with a nitrogen laser, but allowing 1.2 m flight path length. Extraction voltage of 20 kV was applied and the delayed-extraction (DE) mode for enhanced mass resolution was used [74]. Another aliquot of the desalted tryptic digest was used for

MALDI low-energy CID MS/MS experiments on an AXIMA-QIT instrument. This instrument is a first-generation hybrid-type mass spectrometer consisting of a 3D-QIT where the selected precursor ions are extracted with 4 kV and can be further fragmented by CID through an argon pulse in the QIT. The fragments are subsequently separated in a high-resolution reflectron analyzer [75,76]. In some cases up to 5000 consecutive unselected laser shots were necessary to generate high-quality MS/MS spectra for amino acid sequencing. Therefore, sample preparation had to be modified to produce a matrix layer which is not ablated too fast. 0.2 M aqueous diammonium hydrogen citrate, the peptide solution, and a methanolic solution of 2,4,6-trihydroxyacetophenone were mixed directly on the target in the given order and dried at room temperature in a gentle stream of air [77]. External mass calibration with fullerite was used on this instrument. To corroborate and supplement the sequence results obtained with the hybrid instrument, another still available aliquot of the desalted *in-gel* tryptic digest was submitted to a multistage nano-ESI-QIT mass spectrometer (Esquire 3000^{plus}, Bruker Daltoniks, Bremen, Germany) in the *off-line* mode to generate low-energy CID spectra for obtaining sequence tags.

After generating the peptide mass fingerprint on the MALDI-TOF instruments the *m/z* values of all relevant peptides, excluding autodigest products of recombinant trypsin, were analyzed using various *on-line* protein identification tools such as Mascot [78], ProteinProspector [79], and ProFound. All these tools correlate the submitted *m/z* values to *in silico* generated peptides from proteins available through publicly accessible databases such as SWISS-PROT/TrEMBL or the comprehensive, nonidentical protein database at NCBI [80] and list thereafter potential candidates for the protein of interest in a score-based order. Some restrictions such as reasonable taxonomy (green plants), fixed and variable modifications (carbamidomethylation, oxidation of methionines), or peptide molecular mass tolerance for the average ($\pm$1 Da) and/or monoisotopic data points ($\pm$0.5 Da) were set considering the specifications of every single experiment. Amino acid sequences resulting from MS/MS experiments were submitted to database search through BLAST, searching for short, nearly exact sequence matches of proteins originating from *Viridiplantae* listed in the NCBI database.

3. Mass spectrometric identification of proteinous allergens

The list of elicitors of type I allergy is continuously increasing [81]. Studies about the allergenic potential of products from *S. nigra* have just once been carried out [35] but not been scrutinized although allergologists have suspected the possibility of elderberry trees triggering symptoms correlated to this kind of allergy. Especially the fact that these trees are intensively flowering over a period of approximately 2 months in the early summer season, when other major allergenic plants such as

grass or birch are blooming, may have resulted in underestimation or misjudgment of the allergenic potential of elderberry products.

Nine patients with a long history in summer hay fever were tested for symptoms after inhalative and dietary contact with elderberry products. All of them reported rhinoconjunctivitis; four of them even exhibited asthmatic symptoms. As patients may be exposed to the allergens hailing from *S. nigra* via the oral route—flowers and fruits have been used in plant remedies and food for centuries—it has been of special interest that one patient developed upper-airway obstruction when drinking elderberry juices. Four patients showed strong reactions after SPT, medium response was observed in two cases, and negative results were received for three persons including the patient exhibiting airway obstruction. IgE serum levels measured by RAST also varied significantly. In some cases no serum IgE was detectable; in another case up to 4080 kU/L was measured.

It has been of interest to identify and characterize the molecules responsible for type I allergy to elderberry. 1D gel electrophoresis of elderberry flowers, pollen and berry extracts showed a very complex protein composition (Fig. 2). Under nonreducing conditions clearly two dominant gel bands at 33 and 66 kDa were visible after Coomassie staining, which coincide at 33 kDa under reducing circumstances

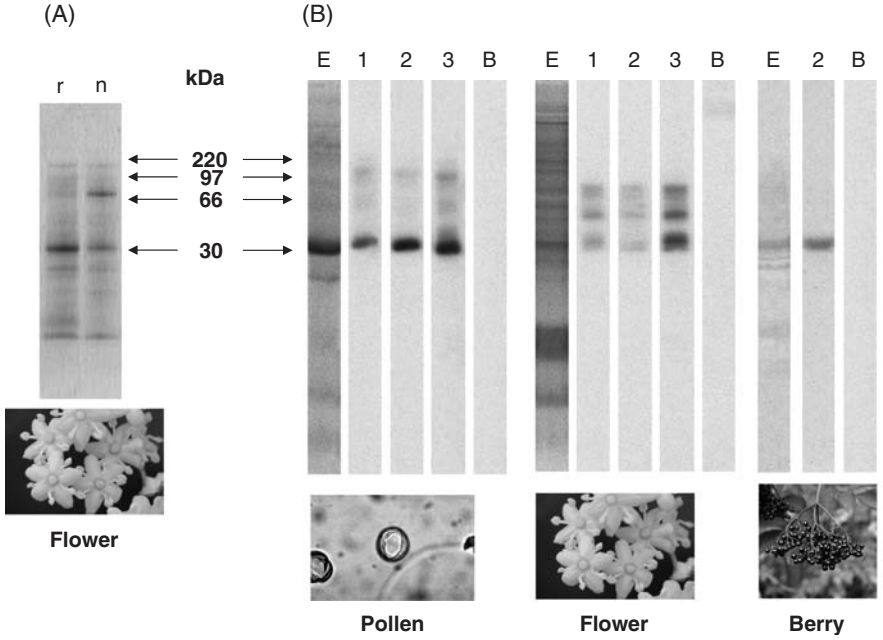


Fig. 2. (A) Coomassie-stained SDS-PAGE gels of elderberry flower extract under denaturing (reducing condition, lane r) and nondenaturing (lane n) conditions exhibiting dominant protein bands at 33 and 33 kDa plus 66 kDa, respectively. (B) IgE binding of representative patients' sera in immunoblotting experiments showing the particular binding of the 33 kDa protein band (lane E: Coomassie-stained gels of the corresponding extracts; lane B: serum of a nonallergic volunteer; lanes 1–3: allergic patients' sera).

(Fig. 2A). Disrupting the disulfide bond(s) of the 66 kDa protein leads to the observation of just one protein band at 33 kDa giving hint to the fact that the protein is consisting of two subunits of identical molecular weight linked via one or more disulfide bonds. However, out of these numerous proteins comprised in the flower extract just one predominant protein at about 33 kDa under reducing conditions could be defined as allergenic by testing patients' sera for their IgE-binding profile using immunoblots (Fig. 2B). This very same allergen could be determined within all pollen, berry, and flower extracts whereby the concentration in pollen seems to be highest compared to fruits and blossoms. Identity of the detected proteins at 33 kDa in the plant materials was validated by testing a serum pool from mice immunized with elderberry pollen extract. It has to be mentioned that only a few number of plants (e.g., birch) show a similar profile of containing just a single allergen relevant for IgE binding [82]. Clinical records of the selected patients suggested cross-reactivity to other well-known summer hay fever elicitors such as grass, mugwort, or birch, but only partial cross-reactivity of specific IgE binding to an elderberry blot at 33 kDa could be observed with birch (*Betula verrucosa*) and mugwort (*Artemisia vulgaris*). Although cross-reactivity between a large number of plants and fruits has been verified in many cases [36,83–85], mostly explained by phylogenetically conserved proteins in plant species, these previous findings could not be asserted for the *S. nigra* allergen. Inhibition, and therefore cross-reactivity, was not observable using major allergens, such as Bet v 1, grass (*P. pretense*) pollen extract, or other recombinant predominant allergens such as Phl p1, Phl p2, or Phl p5, with the result that the 33 kDa protein has to represent a novel type of allergen.

2D gel electrophoresis, carried out to further characterize the physicochemical properties of the allergen complex, clearly exhibited five spots (labeled a–e in Fig. 3A)

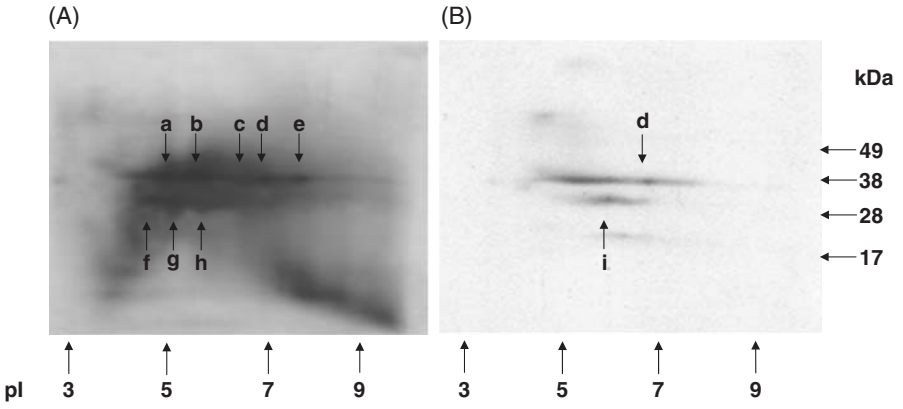


Fig. 3. (A) Coomassie staining of the gel-separated elderberry flower extract exhibited five spots at 33 kDa (marked a–e) and three spots at 28 kDa (marked f–g). (B) Subsequent immunoblotting pointed out that most distinct spot at 33 kDa has a p*I* of 7.1 and correlates very well with spot d on the Coomassie-stained gel.

at the same molecular weight of about 33 kDa. The p*I*s of these protein isoforms span from p*I* 5.0 over p*I* 5.3, 5.9, 6.7 to p*I* 7.1. In addition, three spots at 28 kDa at p*I* 4.7, 5.1, and 5.5 (labeled f–h in Fig. 3A) could be detected after Coomassie staining. Serological investigations pointed out that the most prominent binding of human IgE occurred at 33 kDa/p*I* 7.1, which correlates very well with spot d in the Coomassie-stained gels, with diffuse smearing into the more acidic region (Fig. 3B). The most distinct IgE binding of the 28 kDa protein band was observed at p*I* 6.0. Interestingly, in the Coomassie-stained gel no protein spot could be clearly detected at this position; just a faint smear was visible. Moreover, the more acidic spots at 28 kDa did not show any immune reaction but gave distinct spots after colorimetric staining. The scattering of the protein band at 33 kDa after 2D gel electrophoresis can be explained by the fact that either this allergen is posttranslationally modified (e.g., glycosylated) or single amino acids are exchanged within the polypeptide chain (may be caused by point mutations on the DNA level). Alternatively, more than one protein can be responsible for the allergic reactions against elderberry plants keeping in mind that the polypeptide chains have to be very similar since the molecular weights and p*I*s do not differ very much.

Subsequent Edman degradation after cutting out the relevant region of the PVDF membrane of the 33 kDa allergen at p*I* 7.1 yielded the N-terminal amino acid sequence RDYPFTSRISGGD. Database search for short, in the majority of cases, exact matches within nonredundant protein databases using BLAST but restricting the search to green plants did not give any relevant plant-specific results. A germane hit could not be generated until constraining the taxonomy to the extreme narrow field of *asterids* whereby intrinsic homology to the N-terminus of the β-chain of Ebulin 1, a nontoxic type 2 RIP previously isolated from bark and fruits of *S. ebulus*, could be determined [59]. Girbes et al. identified 25 amino acids of the N-terminus of this protein by Edman degradation, and the molecular weight of the β-chain was estimated to be 26 kDa (by means of SDS-PAGE).

For exact molecular weight determination of the allergen from elderberry flowers, the protein was purified by size exclusion and reversed-phase liquid chromatography, whereupon the purity was controlled by SDS-PAGE, and subsequently submitted to MALDI-TOF-MS analysis. External calibration using bovine serum albumin as calibrant revealed a molecular weight of 66.6 kDa for the intact allergen (Fig. 4) with a mass accuracy of ±0.1%. The observed signal at 33.2 kDa can either display the doubly charged molecular ion of the ion at m/z 66,600 or represent the molecular weight of the subunit already observed in SDS-PAGE under nonreducing conditions. Anyway a small peak at the high mass side of the asymmetric peak could be observed for the 33.2 kDa compound providing indications for present glycoheterogeneities.

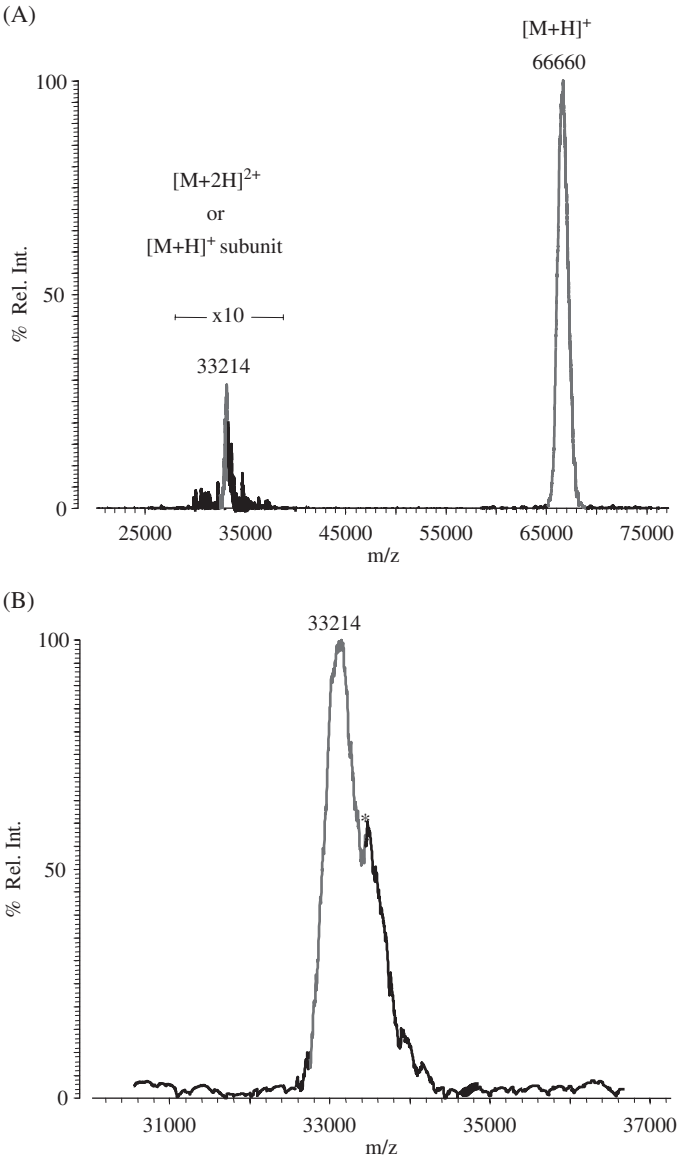


Fig. 4. (A) Positive-ion MALDI mass spectrum in the linear TOF mode of the intact allergen purified from elderberry flower extract. The peak at 33.2 kDa is either the doubly charged molecular ion of the allergen or representing the $[M+H]^+$ ion of the subunit of the allergen. (B) Blowup of the asymmetric peak at m/z 33214 (gray line) exhibiting a small peak marked with an asterisk at the trailing edge ($\Delta m = +0.5$ kDa).

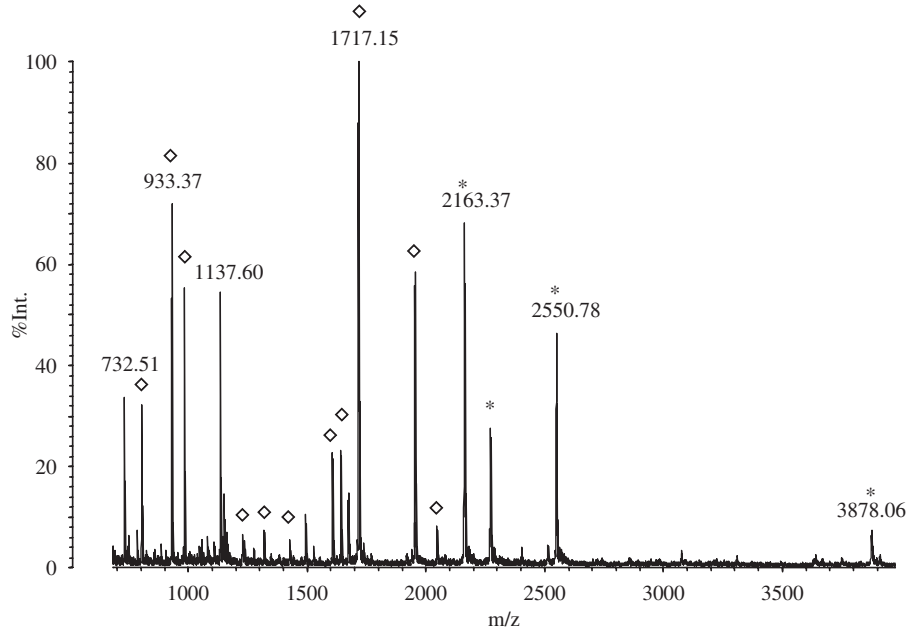


Fig. 5. Representative peptide mass fingerprint for the 33.2 and 66.6 kDa proteins resulting from an *in-gel* digestion of the 66.6 kDa gel band obtained by MALDI-TOF mass spectrometry. Peptides marked with an asterisk result from autodigestion products of recombinant trypsin. Peptides labeled with a diamond were selected for MS/MS experiments carried out on a MALDI-QIT/RTOF and a nano-ESI-QIT mass spectrometer.

Breaking down the protein structure to the peptide level using trypsin for *in-gel* digestion of the 33.2 and 66.6 kDa protein gel band revealed peptide mass finger prints exhibiting the very same peptide masses for the high- and the low-molecular-weight compounds (Fig. 5), indicating that this protein consists of two very similar subunits. Submitting the peptide molecular masses revealed in multiple experiments to publicly available databases using web portals such as Mascot, ProteinProspector, or ProFound did not give search results relevant for *Sambucus* species nor for proteins with homologies to RIPs to unambiguously identify the allergen. For this reason peptides not resulting from trypsin autolysis were selected and subsequently fragmented on a MALDI-QIT/RTOF and a nano-ESI-QIT mass spectrometer to gain information on the amino acid sequences. Table 1 summarizes the CID results for each selected tryptic peptide. Interestingly, both instruments delivered in some cases very good results with long stretches of amino acid sequences (e.g., from the precursor ions *m/z* 1717.15 and 1957.06), which resulted in useful data after BLAST search. It pointed out that the sequences IANNVQPIITSIV and EIIIYQPTGNPNQQWR have a very high homology to the lectin chain of type 2 RIPs originating from *S. nigra* and *S. ebulus* and therefore, due to the highly conserved domain of lectins, also to other lectins from the elderberry plant.

Table 1
Summary of the 12 peptides and their amino acid sequences revealed after MS/MS experiments obtained either on a MALDI-QIT/RTOF or on an off-line nano-ESI-QIT mass spectrometer

Peptide $[M+H]^+$	Amino acid sequence determined by MALDI-QIT/RTOF-MS	Amino acid sequence determined by ESI-QIT-MS	Combined sequence information used for BLAST	Relevant BLAST result
804.43	—	QSDVS[I|L]R	QSDVSLR	✓
933.37	—	[N|D]GLCVDVR	DGLCVDVR	✓
986.34	[{QS}|{NT}]YPFT [{SR}|{TK}]	TDYPFTSR	TDYPFTSR	—
1323.59	(429.40)EQW(323.31) [I|L|Q]	—	EQW or WQE	Sequence too short
1425.34	(358.32)WTW{HV} QVE	—	WTWHVQVE or EVQVHWTW	—
1608.71	(446.31)DKDF(657.26)	KQWTFDKDGDV (256.26)	QWTFDKDGDVR	✓
1642.03	(657.97)GSGDASV [{G[I|L]}|{VA}|{PA}] [Q|I|L]	—	GSGDASV or VSADGSG	—
1704.93	—	WALYGD	WALYGD	✓
1717.15	[{QP}|{EN}|{[I|L]S}] T[I|L]{P[I|L]}QVNN[{GN}|{GQ}|W]	IANNVQ{P[I|L]} [I|L]TS[I|L]V (350.75)	IANNVQPILTSIV	✓
1957.06	EIIIYQPTGN{NP}Q (489.32)	VMYQPTGNPN QQW	EIIIYQPTGNPN QQWR	✓
2048.88	(558.00)QW[I|L](364.01) [I|L][I|L][I|L](359.69)	—	III and QWI or IWQ	Sequences too short
2579.5	—	V[Q|K]P[I|L][I|L] TS[I|L]V	VQPILTSIV	✓

Complementing results of the instruments made the elucidation of long stretches feasible resulting in seven positive BLAST results relevant for *Sambucus* species.

The peptide at *m/z* 1717.15 is moreover a very good example to illustrate the fact that both techniques (MALDI and ESI) and instruments (QIT/RTOF and QIT) complement each other to corroborate and supplement the results (Fig. 6). After interpretation of the MALDI CID spectra sequence, uncertainties due to too many possible dipeptide masses in the low mass range existed. However, missing information on parts of the sequence elucidated from the nano-ESI CID spectrum could be unequivocally reassigned after taking the information from the MALDI MS/MS spectra into account.

Combining the information present in each type of CID spectrum made the determination of the complete sequences of some peptides possible, e.g., the peptide at *m/z* 1957.06 (theoretical monoisotopic molecular weight for the sequence EIIIYQPTGNPNQQWR $[M+H]^+$: 1956.99 Da; observed mass: 1957.06 Da; mass accuracy: +0.07 Da).

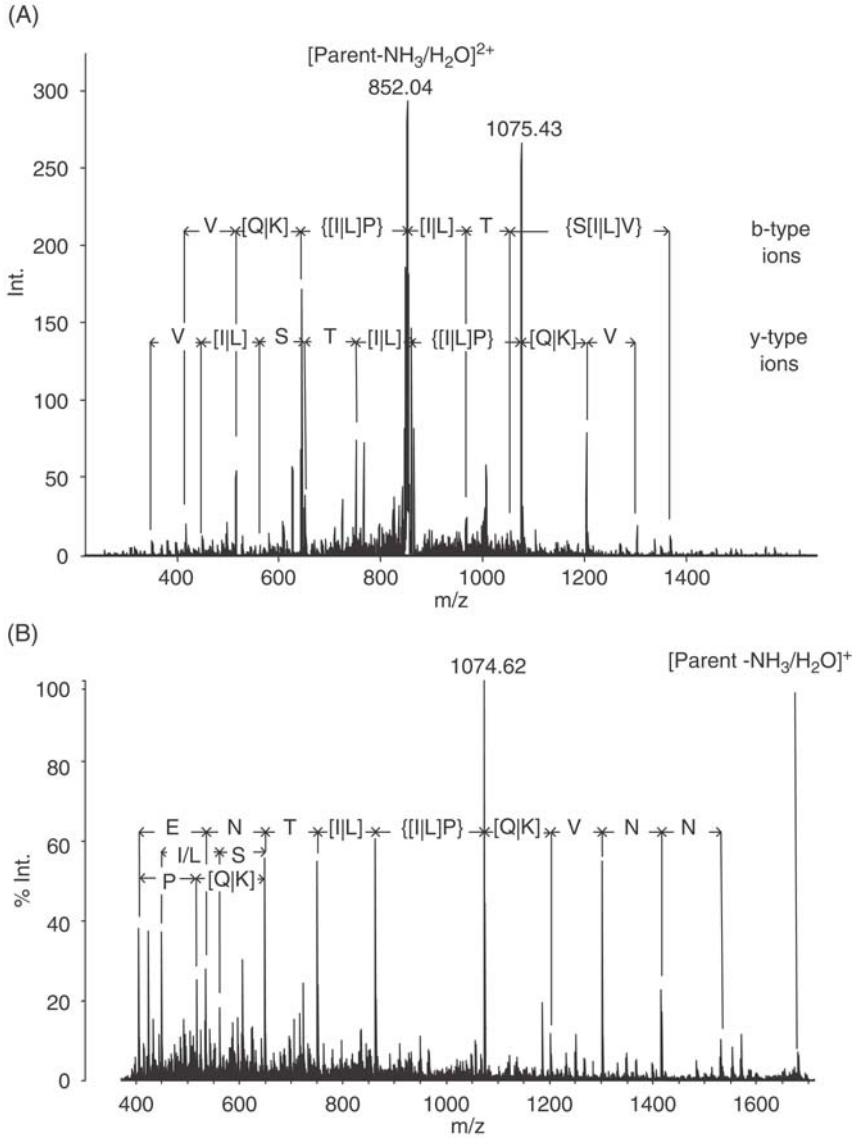


Fig. 6. (A) Low-energy CID spectrum (obtained by means of a nano-ESI-QIT instrument) of the doubly charged precursor ion (*m/z* 859.07) of the peptide with the amino acid sequence IANNVQPILTSIV. (B) Low-energy CID spectrum (obtained by means of a MALDI-QIT/RTOF instrument) of the singly charged precursor ion (*m/z* 1717.15) of the above-mentioned peptide.

In other cases, CID spectra could only be generated on one type of multistage instrument. The peptides with the $[M+H]^+$ precursor ions of *m/z* 804.43, 933.37, 1704.93, and 2579.5 could only be observed on the nano-ESI-QIT mass spectrometer giving peptide sequences of QSDVSLR, DGLCVDVR, WALYGD, and VQPILTSLV where only L/I and Q/K were questionable due to the low-energy CID

characteristics of the QIT itself. All these peptides again were homologs to partial sequences of RIPs. After assignment of the complete sequence it additionally turned out that the peptide at *m/z* 1717.15 (observed on both instruments) is already a fragment of the peptide at *m/z* 2579.5 (only observed in nano-ESI-QIT experiments). This observation can be explained by the amino acid sequence close to the N-terminus of the peptide. Amide bonds near consecutive multiple asparagines (N) are very labile and therefore can easily break during the energetic cooling process in the ion trap, generating a kind of an in-quadrupole ion trap artifact.

Tryptic peptides detected at *m/z* 1323.59 and 2048.88 were just observed on the MALDI-QIT/RTOF instrument and gave only short amino acid sequences useless for BLAST search. Two peptides (*m/z* 1642.03 and 1425.34) revealed the amino acid sequences of GSGDASV/VSADGSG and WTWHVQVE/EVQVHWTW, respectively. Unfortunately it was not possible to gain information on the N- and C-termini of these peptides (amino acid sequences) and therefore the reading direction could not be unambiguously given. BLAST searches taking both possibilities into account gave homologies to proteins not descending from elderberry trees. The peptide at *m/z* 986.34 revealed the amino acid sequence TDYPFTSR which is not homologous to proteins coming from elderberry plants at first sight, but was partially already found after Edman sequencing [35]. In contrary, the sequence QWTFDKDGDVR (*m/z* 1608.71) again showed homologies to RIPs.

By carrying out a gapped BLAST search taking all insights into account (mass spectrometric and Edman sequencing data) and thereafter attaching the identified amino acid sequence tags in the most possible order the best alignment result was obtained for a type 2 RIP from *S. nigra* first identified in 1997 [86]. Interestingly, type 2 RIPs have been reported to carry carbohydrate moieties, which might explain the various protein spots ("spot train") detected after 2D gel electrophoresis. Already in 1997 Van Damme et al. showed that a type 2 RIP from elderberry bark contained about 11 hexose units per native protein [87] using the phenol/H_2SO_4 method [88]. Furthermore, recently the observation that type 2 RIPs are glycoproteins has been verified by NMR experiments using isolated glycopeptides obtained from the β-chain of cinnamomin, a type 2 RIP from *Cinnamomum camphora* [89]. This is corroborated by the extreme peak width in the MALDI-TOF mass spectrum of the intact molecule.

4. Future trends

During the last years more emphasis has been laid on clearing up problems concerning allergies. Great achievements have been made elucidating the metabolic pathway of histamine release, one of the most important representatives of the actual inflammatory mediators for allergic reactions. In the course of these studies, great attention has been paid to food allergy in particular because just 7–10 foodstuffs are responsible for the majority of allergies in the western world,

including quite a number of plant origin. In this case study we could show that products originating from elderberry trees can induce symptoms characteristic for type I allergies such as sneezing, respiratory tract obstruction, or wheezing. Gel electrophoretic experiments and intensive MALDI/ESI mass spectrometric investigations showed that the predominant elicitor for these symptoms is a 66.6 kDa protein consisting of two very similar/identical subunits, which exhibits isoelectric points scattered between p*I* 5 and 7. The fuzzy characteristic of the immunodetected spots after 2D gel electrophoresis pointed out that the allergy elicitor may be just one protein, posttranslationally modified, e.g., by carbohydrate moieties, or that the immunological response results from highly homologous proteins with just minor variation in their polypeptide sequence. Mass spectrometry was a very powerful tool to identify this allergen as a homolog to type 2 RIPs. Although several long stretches of amino acid sequences could be identified by multistage low-energy CID experiments on two different types of mass spectrometers, a unique identification and full characterization of the protein were not possible until now.

Combining mass spectrometry with extremely selective sample preparation and protein analytical separation techniques can be the future trend to clearly pin down the primary structure of the type I elicitor in elderberry products. This can, for instance, be done by affinity purification taking the possible carbohydrate moieties into account and therefore enriching the protein by specific sugar interactions. Alternatively, affinity chromatography based on immobilized monoclonal antibodies directed against the allergen itself can be used. Affinity chromatography is a well-known and successfully established analytical method that can be carried out in preparative or analytical scales and can be efficiently coupled to mass spectrometry in the *off-line* [90] or *on-line* mode [91,92]. Particularly, the *off-line* mode has the enormous advantage that samples can be spotted onto membranes for ongoing serological investigations on the intact protein carried out in parallel to mass spectrometry investigating structural components by spotting the analyte onto MALDI sample plates for analyses on high-resolution TOF instruments to detect the exact molecular weight of the intact protein.

Besides affinity chromatography, capillary zone electrophoresis can easily be coupled to an ESI high-performance RTOF mass spectrometer. The high accuracy achievable in molecular mass determination by newly developed reflectron mass spectrometers makes the assignment of isomeric glycoproteins feasible [93] and can furthermore be a powerful tool to discriminate between different serologically relevant molecules.

Breaking down the protein into smaller fragments by enzymatic treatment makes the elucidation of the immunological-relevant epitope of the allergen approachable. Enrichment of the possibly present glycopeptides by affinity-, ion-exchange, or straight-phase chromatography and subsequent clarification of the carbohydrate moiety substructures by high- (tandem MS) and low-energy CID

(multistage MS) experiments may reveal new sugar structures responsible for immune responses or confirm the latest published results of carbohydrate core structures (core α1,3-fucose, core xylose) to be responsible for antibody binding [38,39,42].

5. Conclusions

Fruits and flowers of elderberry trees are widely used in herbal medicine as remedies for cold, influenza, and catarrhal inflammation. Type I allergy to this plant has been suspected by clinicians over a long period of time but has never been further studied.

Recently published data gave first evidence that *S. nigra* is a plant truly harboring allergenic potential and that a 33.2 kDa protein, a subunit of the 66.6 kDa intact component, is the predominant elicitor for allergic reactions. Remarkably, initially suspected cross-reactivity of concerned individuals to major triggers of summer hay fever such as grass, mugwort, or birch could only partly be confirmed. Partial cross-reactivity to birch and mugwort was observed but not to the other main actuators.

Gel electrophoretic techniques pointed out that the predominant human IgE binding protein consists of two subunits having identical molecular weight and more than one isoform with isoelectric points scattered between p*I* 5 and 7 whereupon the isoform at p*I* 7 showed the strongest immune response. N-terminal sequencing of the dominant allergen resulted in 13 amino acids giving first indication for a type 2 RIP.

Mass spectrometry was a very powerful tool to substantiate these findings. *In-gel* digestions after 1D SDS-PAGE of the 33.2 and 66.6 kDa proteins were performed and the identical resulting tryptic peptides were further sequenced by multistage low-energy CID experiments. Shorter and longer stretches of amino acid sequences of eight peptides could be assigned by means of a hybrid multistage MALDI-QIT/RTOF instrument in the low-energy CID mode. These results were corroborated and supplemented by low-energy CID experiments with a nano-ESI-QIT mass spectrometer providing thereafter three complete and nine partial amino acid sequences of tryptic peptides with very high confidence. Based on these data, intensive bioinformatic data mining delivered the information that a high homology to lectins, in particular to type 2 RIPs, from *S. nigra* is given. Considering that dietary lectins can induce histamine release [42] this has been a notably interesting result. Unfortunately, the various gel spots detected in 2D PAGE could not be clearly assigned to potential protein homologues or possible carbohydrate isoforms until now, but these investigations are in progress. Furthermore, the knowledge that protein sequence variation, protein conformations, and posttranslational modifications, such as specific carbohydrate structures, are involved in the generation of

IgE-reactive epitopes was not of assistance for clearing up the uncertainties of protein identification but complicated the situation for clarification.

Sophisticated purification steps to get hold of the highly purified allergen have to be developed including, e.g., affinity purification by either immobilizing antibodies directed against the allergen itself or working with affinity ligands taking the possible carbohydrate moieties of the allergen into account. Nevertheless, all the observations until now indicate that the predominant elicitor for type I allergy induced by elderberry flowers is a type 2 RIP. This is of particular interest for immunology as the family of RIPs has recently gained importance in anticancer and antiviral therapy due to their antiproliferative and antimitogenic activities.

References

1. Turk, J.L., Von Pirquet, allergy and infectious diseases: a review. *J. R. Soc. Med.*, **80**, 31–33 (1987).
2. Gell, P.G.H. and Coombs, R.R.A. (Eds.), *The Classification of Allergic Reactions Underlying Disease.* Blackwell, Oxford, 1963.
3. Rajan, T.V., The Gell-Coombs classification of hypersensitivity reactions: a re-interpretation. *Trends Immunol.*, **24**, 376–379 (2003).
4. Nutter, A.F., Contact urticaria to rubber. *Br. J. Dermatol.*, **101**, 597–598 (1979).
5. European Commission, Determination of allergological relevant compounds in disposable gloves. Correlation of chemical, allergological and immunological data. Final report MAT1—CT 940060. In: European Commission Directorate General XII. Science, Research and Development. Directorate C: Industrial and material technologies programme measurement and testing. Brussels, Belgium, 1998.
6. Czuppon, A.B., Chen, Z., Rennert, S., Engelke, T., Meyer, H.E., Heber, M. and Baur, X., The rubber elongation factor of rubber trees (*Hevea brasiliensis*) is the major allergen in latex. *J. Allergy Clin. Immunol.*, **92**, 690–697 (1993).
7. Dennis, M., Henzel, W., Bell, J., Kohr, W. and Light, D., Amino acid sequence of rubber elongation factor protein associated with rubber particles in Hevea latex. *J. Biol. Chem.*, **264**, 18618–18617 (1989).
8. Dürauer, A., Csaszar, E., Mechtler, K., Jungbauer, A. and Schmid, E., Characterisation of the rubber elongation factor from ammoniated latex by electrophoresis and mass spectrometry. *J. Chromatogr. A*, **890**, 145–158 (2000).
9. Alenius, H., Kalkkinen, N., Reunala, T., Turjanmaa, K. and Palosuo, T., The main IgE-binding epitope of a major latex allergen, prohevein, is present in its N-terminal 43-amino acid fragment, hevein. *J. Immunol.*, **156**, 1618–1625 (1996).
10. Chen, Z., Posch, A., Lohaus, C., Raulf-Heimsoth, M., Meyer, H.E. and Baur, X., Isolation and identification of hevein as a major IgE-binding polypeptide in Hevea latex. *J. Allergy Clin. Immunol.*, **99**, 402–409 (1997).
11. Soedjanaatmadja, U.M., Hofsteenge, J., Jeronimus-Stratingh, C.M., Bruins, A.P. and Beintema, J.J., Demonstration by mass spectrometry that pseudo-hevein and hevein have ragged C-terminal sequences. *Biochim. Biophys. Acta*, **1209**, 144–148 (1994).
12. Karisola, P., Kotovuori, A., Poikonen, S., Niskanen, E., Kalkkinen, N., Turjanmaa, K., Palosuo, T., Reunala, T., Alenius, H. and Kulomaa, M.S., Isolated hevein-like domains, but not 31-kd endochitinases, are responsible for IgE-mediated in vitro and in vivo reactions in latex-fruit syndrome. *J. Allergy Clin. Immunol.*, **115**, 598–605 (2005).

13. Schad, S.G., Trcka, J., Vieths, S., Scheurer, S., Conti, A., Brocker, E.B. and Trautmann, A., Wine anaphylaxis in a German patient: IgE-mediated allergy against a lipid transfer protein of grapes. *Int. Arch. Allergy Immunol.*, **136**, 159–164 (2005).
14. Fernandez-Rivas, M., Gonzalez-Mancebo, E., Rodriguez-Perez, R., Benito, C., Sanchez-Monge, R., Salcedo, G., Alonso, M.D., Rosado, A., Tejedor, M.A., Vila, C. and Casas, M.L., Clinically relevant peach allergy is related to peach lipid transfer protein, Pru p 3, in the Spanish population. *J. Allergy Clin. Immunol.*, **112**, 789–795 (2003).
15. Pastorello, E.A., Pompei, C., Pravettoni, V., Farioli, L., Calamari, A.M., Scibilia, J., Robino, A.M., Conti, A., Iametti, S., Fortunato, D., Bonomi, S. and Ortolani, C., Lipid-transfer protein is the major maize allergen maintaining IgE-binding activity after cooking at 100 degrees C, as demonstrated in anaphylactic patients and patients with positive double-blind, placebo-controlled food challenge results. *J. Allergy Clin. Immunol.*, **112**, 775–783 (2003).
16. Ahrazem, O., Ibanez, M.D., Lopez-Torrejon, G., Sanchez-Monge, R., Sastre, J., Lombardero, M., Barber, D. and Salcedo, G., Lipid transfer proteins and allergy to oranges. *Int. Arch. Allergy Immunol.*, **137**, 201–210 (2005).
17. Conti, A., Fortunato, D., Ortolani, C., Giuffrida, M.G., Pravettoni, V., Napolitano, L., Farioli, L., Perono Garoffo, L., Trambaioli, C. and Pastorello, E.A., Determination of the primary structure of two lipid transfer proteins from apricot (Prunus armeniaca). *J. Chromatogr. B Biomed. Sci. Appl.*, **756**, 123–129 (2001).
18. Pastorello, E.A., Ortolani, C., Baroglio, C., Pravettoni, V., Ispano, M., Giuffrida, M.G., Fortunato, D., Farioli, L., Monza, M., Napolitano, L., Sacco, M., Scibola, E. and Conti, A., Complete amino acid sequence determination of the major allergen of peach (Prunus persica) Pru p 1. *Biol. Chem.*, **380**, 1315–1320 (1999).
19. Lahti, A., Bjorksten, F. and Hannuksela, M., Allergy to birch pollen and apple, and cross-reactivity of the allergens studied with the RAST. *Allergy*, **35**, 297–300 (1980).
20. Halmepuro, L., Vuontela, K., Kalimo, K. and Bjorksten, F., Cross-reactivity of IgE antibodies with allergens in birch pollen, fruits and vegetables. *Int. Arch. Allergy Appl. Immunol.*, **74**, 235–240 (1984).
21. Swoboda, I., Jilek, A., Ferreira, F., Engel, E., Hoffmann-Sommergruber, K., Scheiner, O., Kraft, D., Breiteneder, H., Pittenauer, E., Schmid, E., Vicente, O., Heberle-Bors, E., Ahorn, H., and Breitenbach, M., Isoforms of Bet v 1, the major birch pollen allergen, analyzed by liquid chromatography, mass spectrometry, and cDNA cloning. *J. Biol. Chem.*, **270**, 2607–2613 (1995).
22. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse, C.M., Electrospray ionization for mass spectrometry of large biomolecules. *Science*, **246**, 64–71 (1989).
23. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y. and Yohida, T., Protein and polymer analyses up to *m/z* 1,00,000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, **2**, 151–153 (1988).
24. Durham, S., ABC of allergies. Summer hay fever. *Br. Med. J.*, **316**, 843–845 (1998).
25. Barak, V., Halperin, T. and Kalickman, I., The effect of Sambucol, a black elderberry-based, natural product, on the production of human cytokines: I. Inflammatory cytokines. *Eur. Cytokine Netw.*, **12**, 290–296 (2001).
26. Moerman, D.E. and Estabrook, G.F., Native Americans' choice of species for medicinal use is dependent on plant family: confirmation with meta-significance analysis. *J. Ethnopharmacol.*, **87**, 51–59 (2003).
27. Borchers, A.T., Keen, C.L., Stern, J.S. and Gershwin, M.E., Inflammation and Native American medicine: the role of botanicals. *Am. J. Clin. Nutr.*, **72**, 339–347 (2000).
28. Zakay-Rones, Z., Varsano, N., Zlotnik, M., Manor, O., Regev, L., Schlesinger, M. and Mumcuoglu, M., Inhibition of several strains of influenza virus in vitro and reduction of symptoms by an elderberry extract (*Sambucus nigra L.*) during an outbreak of influenza B Panama. *J. Altern. Complement. Med.*, **1**, 361–369 (1995).

29. Peumans, W.J., Van Damme, E.J., Barre, A. and Rouge, P., Classification of plant lectins in families of structurally and evolutionary related proteins. *Adv. Exp. Med. Biol.*, **491**, 27–54 (2001).
30. Baum, L.G., Derbin, K., Perillo, N.L., Wu, T., Pang, M. and Uittenbogaart, C. Characterization of terminal sialic acid linkages on human thymocytes. Correlation between lectin-binding phenotype and sialyltransferase expression. *J. Biol. Chem.*, **271**, 10793–10799 (1996).
31. Chintalacharuvu, S.R. and Emancipator, S.N., The glycosylation of IgA produced by murine B cells is altered by Th2 cytokines. *J. Immunol.*, **159**, 2327–2333 (1997).
32. Tong, H.H., James, M., Grants, I., Liu, X., Shi, G. and DeMaria, T.F., Comparison of structural changes of cell surface carbohydrates in the eustachian tube epithelium of chinchillas infected with a Streptococcus pneumoniae neuraminidase-deficient mutant or its isogenic parent strain. *Microb. Pathog.*, **31**, 309–317 (2001).
33. Pollenflugkalender URL http://www.die-seite.at/special/pollen/pollenkalender.php, February 15, 2005.
34. Van Damme, E.J.M., Charels, D., Roy, S., Tierens, K., Barre, A., Martins, J.C., Rouge, P., Van Leuven, F., Does, M. and Peumans, W.J., A gene encoding a hevein-like protein from elderberry fruits is homologous to PR-4 and class V chitinase genes. *Plant Physiol.*, **119**, 1547–1556 (1999).
35. Förster-Waldl, E., Marchetti, M., Schöll, I., Focke, M., Radauer, C., Kinaciyan, T., Nentwich, I., Jäger, S., Schmid, E.R., Boltz-Nitulescu, G., Scheiner, O. and Jensen-Jarolim, E., Type I allergy to elderberry is elicited by a 33.2 kDa allergen with significant homology to ribosomal inactivating proteins (RIPs). *Clin. Exp. Allergy*, **33**, 1703–1710 (2003).
36. Breiteneder, H. and Ebner, C., Molecular and biochemical classification of plant-derived food allergens. *J. Allergy Clin. Immunol.*, **106**, 27–36 (2000).
37. Mills, E.N.C., Madsen, C., Shewry, P.R. and Wichers, H.J., Food allergens of plant origin—their molecular and evolutionary relationships. *Trends Food Sci. Technol.*, **14**, 145–156 (2003).
38. Wilson, I.B., Harthill, J.E., Mullin, N.P., Ashford, D.A. and Altmann, F., Core alpha1,3-fucose is a key part of the epitope recognized by antibodies reacting against plant N-linked oligosaccharides and is present in a wide variety of plant extracts. *Glycobiology*, **8**, 651–661 (1998).
39. Bardor, M., Faveeuw, C., Fitchette, A.C., Gilbert, D., Galas, L., Trottein, F., Faye, L. and Lerouge, P., Immunoreactivity in mammals of two typical plant glyco-epitopes, core alpha(1,3)-fucose and core xylose. *Glycobiology*, **13**, 427–434 (2003).
40. Corti, V., Cattaneo, A., Bachi, A., Rossi, R.E., Monasterolo, G., Paolucci, C., Burastero, S.E. and Alessio, M., Identification of grass pollen allergens by two-dimensional gel electrophoresis and serological screening. *Proteomics*, **5**, 729–736 (2005).
41. Leonard, R., Petersen, B.O., Himly, M., Kaar, W., Wopfner, N., Kolarich, D., van Ree, R., Ebner, C., Duus, J.O., Ferreira, F. and Altmann, F., Two novel types of O-glycans on the mugwort pollen allergen art v 1 and their role in antibody binding. *J. Biol. Chem.*, **280**, 7932–7940 (2005).
42. Haas, H., Falcone, F.H., Schramm, G., Haisch, K., Gibbs, B.F., Klaucke, J., Poppelmann, M., Becker, W.M., Gabius, H.J. and Schlaak, M., Dietary lectins can induce in vitro release of IL-4 and IL-13 from human basophils. *Eur. J. Immunol.*, **29**, 918–927 (1999).
43. Shibasaki, M., Sumazaki, R., Isoyama, S. and Takita, H., Interaction of lectins with human IgE: IgE-binding property and histamine-releasing activity of twelve plant lectins. *Int. Arch. Allergy Immunol.*, **98**, 18–25 (1992).
44. Burks, A.W., Cockrell, G., Connaughton, C., Guin, J., Allen, W. and Helm, R.M. Identification of peanut agglutinin and soybean trypsin inhibitor as minor legume allergens. *Int. Arch. Allergy Immunol.*, **105**, 143–149 (1994).
45. Barbieri, L., Ciani, M., Girbes, T., Liu, W.-Y., Van Damme, E.J.M., Peumans, W.J. and Stirpe, F., Enzymatic activity of toxic and non-toxic type 2 ribosome-inactivating proteins. *FEBS Lett.*, **563**, 219–222 (2004).

46. Vandenbussche, F., Desmyter, S., Ciani, M., Proost, P., Peumans, W.J. and Van Damme, E.J., Analysis of the in planta antiviral activity of elderberry ribosome-inactivating proteins. *Eur. J. Biochem.*, **271**, 1508–1515 (2004).
47. Chen, Y., Peumans, W.J. and Van Damme, E.J.M., The *Sambucus nigra* type-2 ribosome-inactivating protein SNA-I′ exhibits in planta antiviral activity in transgenic tobacco. *FEBS Lett.*, **516**, 27–30 (2002).
48. Chow, L.P., Chou, M.H., Ho, C.Y., Chuang, C.C., Pan, F.M., Wu, S.H. and Lin, J.Y., Purification, characterization and molecular cloning of trichoanguin, a novel type I ribosome-inactivating protein from the seeds of *Trichosanthes anguina*. *Biochem. J.*, **338**(**Pt 1**), 211–219 (1999).
49. Hartley, M.R. and Lord, J.M., Cytotoxic ribosome-inactivating lectins from plants. *Biochim. Biophys. Acta*, **1701**, 1–14 (2004).
50. McGrath, M.S., Hwang, K.M., Caldwell, S.E., Gaston, I., Luk, K.C., Wu, P., Ng, V.L., Crowe, S., Daniels, J., Marsh, J., Deinhart, T., Lekas, P.V., Vennari, J.C., Yeang, H.-W. and Lifson, J.D., GLQ223: an inhibitor of human immunodeficiency virus replication in acutely and chronically infected cells of lymphocyte and mononuclear phagocyte lineage. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 2844–2848 (1989).
51. Zarling, J.M., Moran, P.A., Haffar, O., Sias, J., Richman, D.D., Spina, C.A., Myers, D.E., Kuebelbeck, V., Ledbetter, J.A. and Uckun, F.M., Inhibition of HIV replication by pokeweed antiviral protein targeted to CD4+ cells by monoclonal antibodies. *Nature*, **347**, 92–95 (1990).
52. Au, T.K., Collins, R.A., Lam, T.L., Ng, T.B., Fong, W.P. and Wan, D.C.C., The plant ribosome inactivating proteins luffin and saporin are potent inhibitors of HIV-1 integrase. *FEBS Lett.*, **471**, 169–172 (2000).
53. Wang, Y.X., Jacob, J., Wingfield, P.T., Palmer, I., Stahl, S.J., Kaufman, J.D., Huang, P.L., Lee-Huang, S. and Torchia, D.A., Anti-HIV and anti-tumor protein MAP30, a 30 kDa single-strand type-I RIP, shares similar secondary structure and beta-sheet topology with the A chain of ricin, a type-II RIP. *Protein Sci.*, **9**, 138–144 (2000).
54. Sandvig, K. and van Deurs, B., Entry of ricin and Shiga toxin into cells: molecular mechanisms and medical perspectives. *Embo J.*, **19**, 5943–5950 (2000).
55. Kodama, T., Doukas, A.G. and Hamblin, M.R., Delivery of ribosome-inactivating protein toxin into cancer cells with shock waves. *Cancer Lett.*, **189**, 69–75 (2003).
56. Citores, L., Miguel Ferreras, J., Munoz, R., Benitez, J., Jimenez, P. and Girbes, T., Targeting cancer cells with transferrin conjugates containing the non-toxic type 2 ribosome-inactivating proteins nigrin b or ebulin l. *Cancer Lett.*, **184**, 29–35 (2002).
57. Peumans, W.J., Hao, Q. and Van Damme, E.J., Ribosome-inactivating proteins from plants: more than RNA N-glycosidases? *FASEB J.*, **15**, 1493–1506 (2001).
58. Van Damme, E.J., Barre, A., Rouge, P., Van Leuven, F. and Peumans, W.J., The NeuAc(alpha-2,6)-Gal/GalNAc-binding lectin from elderberry (*Sambucus nigra*) bark, a type-2 ribosome-inactivating protein with an unusual specificity and structure. *Eur. J. Biochem.*, **235**, 128–137 (1996).
59. Girbes, T., Citores, L., de Benito, F.M., Inglesias, R. and Ferreras, J.M., A non-toxic two-chain ribosome-inactivating protein co-exists with a structure-related monomeric lectin (SNA III) in elder (*Sambucus nigra*) fruits. *Biochem. J.*, **315**, 343–344 (1996).
60. Battelli, M.G., Citores, L., Buonamici, L., Ferreras, J.M., de Benito, F.M., Stirpe, F., Girbes, T., Inglesias, R., Van Damme, E.J., Barre, A., Rouge, P., Van Leuven, F., and Peumans, W.J., Toxicity and cytotoxicity of nigrin b, a two-chain ribosome-inactivating protein from *Sambucus nigra*: comparison with ricin, a non-toxic two-chain ribosome-inactivating protein co-exists with a structure-related monomeric lectin (SNA III) in elder (*Sambucus nigra*) fruits. The NeuAc(alpha-2,6)-Gal/GalNAc-binding lectin from elderberry (*Sambucus nigra*) bark, a type-2 ribosome-inactivating protein with an unusual specificity and structure. *Arch. Toxicol.*, **71**, 360–364 (1997).

61. Endo, Y., Mitsui, K., Motizuki, M. and Tsurugi, K., The mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. The site and the characteristics of the modification in 28 S ribosomal RNA caused by the toxins. *J. Biol. Chem.*, **262**, 5908–5912 (1987).
62. Li, X.D., Chen, W.F., Liu, W.Y. and Wang, G.H., Large-scale preparation of two new ribosome-inactivating proteins—cinnamomin and camphorin from the seeds of *Cinnamomum camphora*. *Protein Expr. Purif.*, **10**, 27–31 (1997).
63. Sharma, N., Park, S.W., Vepachedu, R., Barbieri, L., Ciani, M., Stirpe, F., Savary, B. J. and Vivanco, J.M., Isolation and characterization of an RIP (ribosome-inactivating protein)-like protein from tobacco with dual enzymatic activity. *Plant Physiol.*, **134**, 171–181 (2004).
64. Helmy, M., Lombard, S. and Pieroni, G., Ricin RCA60: evidence of its phospholipase activity. *Biochem. Biophys. Res. Commun.*, **258**, 252–255 (1999).
65. Park, S.W., Vepachedu, R., Sharma, N. and Vivanco, J.M., Ribosome-inactivating proteins in plant biology. *Planta*, **219**, 1093–1096 (2004).
66. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25**, 3389–3402 (1997).
67. Jensen-Jarolim, E., Gerstmayer, G., Kraft, D., Scheiner, O., Ebner, H. and Ebner, C., Serological characterization of allergens in poppy seeds. *Clin. Exp. Allergy*, **29**, 1075–1079 (1999).
68. Laemmli, U.K., Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680–685 (1970).
69. Görg, A., Postel, W. and Gunther, S., The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **9**, 531–546 (1988).
70. Gill, S.C. and von Hippel, P.H., Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.*, **182**, 319–326 (1989).
71. Kussmann, M., Nordhoff, E., Rahbek-Nielsen, H., Haebel, S., Rossel-Larsen, M., Jakobsen, L., Gobom, J., Mirgorodskaya, E., Kroll-Kristensen, A., Palm, L. and Roepstorff, P., Matrix-assisted laser desorption/ionization mass spectrometry sample preparation techniques designed for various peptide and protein analytes. *J. Mass Spectrom.*, **32**, 593–601 (1997).
72. Shevchenko, A., Wilm, M., Vorm, O. and Mann, M., Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.*, **68**, 850–858 (1996).
73. Pluskal, M.G., In the laboratory: microscale sample preparation. *Nat. Biotechnol.*, **18**, 104–105 (2000).
74. Tsarbopoulos, A., Bahr, U., Pramanik, B.N. and Karas, M., Glycoprotein analysis by delayed extraction and post-source decay MALDI-TOF-MS. *Int. J. Mass Spectrom. Ion Processes*, **169/170**, 251–261 (1997).
75. Martin, R.L. and Brancia, F.L., Analysis of high mass peptides using a novel matrix-assisted laser desorption/ionisation quadrupole ion trap time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.*, **17**, 1358–1365 (2003).
76. Demelbauer, U.M., Zehl, M., Plematl, A., Allmaier, G. and Rizzi, A., Determination of glycopeptide structures by multistage mass spectrometry with low-energy collision-induced dissociation: comparison of electrospray ionization quadrupole ion trap and matrix-assisted laser desorption/ionization quadrupole ion trap reflectron time-of-flight approaches. *Rapid Commun. Mass Spectrom.*, **18**, 1575–1582 (2004).
77. Zehl, M. and Allmaier, G., Investigation of sample preparation and instrumental parameters in the matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of noncovalent peptide/peptide complexes. *Rapid Commun. Mass Spectrom.*, **17**, 1931–1940 (2003).
78. Perkins, D.N., Pappin, D.J.C., Creasy, D.M. and Cottrell, J.S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567 (1999).

79. Clauser, K.R., Baker, P. and Burlingame, A.L., Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.*, **71**, 2871–2882 (1999).
80. Wilkins, M.R., Williams, K.L., Appel, R.D. and Hochstrasser, D.F. (Eds.), *Proteome Research: New Frontiers in Functional Genomics.* Springer Verlag, Berlin Heidelberg, New York, 1997.
81. Gendel, S.M., Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv. Food Nutr. Res.*, **42**, 63–92 (1998).
82. Jarolim, E., Rumpold, H., Endler, A.T., Ebner, H., Breitenbach, M., Scheiner, O. and Kraft, D., IgE and IgG antibodies of patients with allergy to birch pollen as tools to define the allergen profile of Betula verrucosa. *Allergy*, **44**, 385–395 (1989).
83. Valenta, R., Duchene, M., Ebner, C., Valent, P., Sillaber, C., Deviller, P., Ferreira, F., Tejkl, M., Edelmann, H., Kraft, D. and Scheiner, O., Profilins constitute a novel family of functional plant pan-allergens. *J. Exp. Med.*, **175**, 377–385 (1992).
84. Mikkola, J.H., Alenius, H., Kalkkinen, N., Turjanmaa, K., Palosuo, T. and Reunala, T., Hevein-like protein domains as a possible cause for allergen cross-reactivity between latex and banana. *J. Allergy Clin. Immunol.*, **102**, 1005–1012 (1998).
85. Barral, P., Batanero, E., Palomares, O., Quiralte, J., Villalba, M. and Rodriguez, R., A major allergen from pollen defines a novel family of plant proteins and shows intra- and interspecies [correction of interspecie] cross-reactivity. *J. Immunol.*, **172**, 3644–3651 (2004).
86. Van Damme, E.J., Roy, S., Barre, A., Rouge, P., Van Leuven, F. and Peumans, W.J., The major elderberry (*Sambucus nigra*) fruit protein is a lectin derived from a truncated type 2 ribosome-inactivating protein. *Plant J.*, **12**, 1251–1260 (1997).
87. Van Damme, E.J.M., Barre, A., Rouge, P., Van Leuven, F. and Penumans, W.J., Isolation and molecular cloning of a novel type 2 ribosome-inactivating protein with an inactive B chain from elderberry (*Sambucus nigra*) bark. *J. Biol. Chem.*, **272**, 8353–8360 (1997).
88. Dubois, M., Gilles, K., Hamilton, J.K., Rebers, P.A. and Smith, F., A colorimetric method for the determination of sugars. *Nature*, **168**, 167 (1951).
89. Pu, Z., Li, Y., Hou, F.J., He, F., Zhang, N., Wu, H. and Liu, W.Y., Structural studies of the glycopeptides of B-chain of cinnamomin—a type II ribosome-inactivating protein by nuclear magnetic resonance. *Glycoconj. J.*, **17**, 749–759 (2000).
90. Demelbauer, U.M., Plematl, A., Kremser, L., Allmaier, G., Josic, D. and Rizzi, A., Characterization of glyco isoforms in plasma-derived human antithrombin by on-line capillary zone electrophoresis-electrospray ionization-quadrupole ion trap-mass spectrometry of the intact glycoproteins. *Electrophoresis*, **25**, 2026–2032 (2004).
91. Ng, E.S., Yang, F., Kameyama, A., Palcic, M.M., Hindsgaul, O. and Schriemer, D.C., High-throughput screening for enzyme inhibitors using frontal affinity chromatography with liquid chromatography and mass spectrometry. *Anal. Chem.*, **77**, 6125–6133 (2005).
92. Zhang, B., Palcic, M.M., Schriemer, D.C., Alvarez-Manilla, G., Pierce, M. and Hindsgaul, O., Frontal affinity chromatography coupled to mass spectrometry for screening mixtures of enzyme inhibitors. *Anal. Biochem.*, **299**, 173–182 (2001).
93. Neusüß, C., Demelbauer, U. and Pelzing, M., Glycoform characterization of intact erythropoietin by capillary electrophoresis-electrospray-time-of-flight mass spectrometry. *Electrophoresis*, **26**, 1442–1450 (2005).

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 21

Mass spectrometry in clinical treatment

ANDRÁS TELEKES[a,1], MÁRTA HEGEDŰS[a,2], and ISTVÁN KISS[b,3]

[a]*National Institute of Oncology, Budapest, Hungary*
[b]*St. Imre Government Teaching Hospital, Budapest, Hungary*

1. Introduction

The first clinical applications of mass spectrometry (MS) date back to 1966, when the use of gas chromatography coupled to mass spectrometry (GC–MS) for identification of organic acidurias in children was reported [1]. Twenty years later tandem mass spectrometry (MS/MS) was introduced into clinical laboratories, and first applied to the evaluation of children at risk of inborn errors of metabolism [2–4]. Using stable-isotope-labeled (i.e., not radioactive) internal standards, the method was further enhanced. These standards are identical to the native analytes, except that their molecular masses are slightly different. Addition of these isotopes at known concentration to the sample before analysis serves as a positive control, helping identification and quantification of the analytes. The use of MS is, however, not yet routine in many fields where it could influence clinical decisions. While medical research using MS is flourishing, few applications have become part of the standard "bedside" practice. This is partly because the transition of MS from a research tool to a reliable clinical diagnostic platform requires rigorous standardization, spectral quality control and assurance, standard operating procedures for robotic

[1]Tel.: +36-204-119080; Fax: +36-139-52835. E-mail: andras.telekes@gmail.com (A. Telekes).
[2]E-mail: hegedus.marta@gmail.com (M. Hegedűs).
[3]E-mail: ikiss@enternet.hu (I. Kiss).

and automatic sample application, and standardized controls to ensure generation of highly reproducible spectra [5]. Reliable identification of protein expression patterns and associated protein biomarkers that differentiate disease from health or that distinguish different stages of a disease has now started to become feasible. There are many MS-based techniques for identification of biomarkers and protein expression patterns; see, e.g., Chapters 6 and 8. Among these, surface-enhanced laser desorption ionization (SELDI) has gained popularity in the clinical field, mainly due to its ease of use. Note that the same technique is often indicated as SELDI, SELDI–TOF, or SELDI TOF–MS. Recently, however, some results obtained using SELDI have been questioned, but this does not diminish the importance of proteomics-based biomarkers.

In a recent study the systematic variability of SELDI experiments was evaluated using biological and technical replicates. Systematic biases on plates, chips, and spots were not found. Reproducibility of SELDI experiments was demonstrated by examining the resulting low coefficient of variances of five peaks presented in all 144 spectra from quality control samples that were loaded randomly on different spots in the chips of six bioprocessor plates. The authors developed a method to detect and discard low quality spectra prior to proteomic profiling data analysis. This quality control tool involved a correlation matrix to measure the similarities among SELDI mass spectra obtained from similar biological samples. The reproducibility of experiments was acceptable and the profiling data for subsequent data analysis were reported to be reliable [6].

Valid biological information from SELDI-MS requires attention to experimental design, sample handling, and data processing. In the literature information on the biological aspects can be found and computer-learning algorithms have been applied to locate sets of biomarkers. Focus is needed on locating and measuring proteins across mass spectra, optimizing the trade-off between sensitivity and false discovery. Furthermore, the identified features must be biologically meaningful, representing identifiable chemical species for further investigation. Carlson et al. have developed an approach to address the deficiencies in reproducibility and comparability that exist across published studies. This approach, simultaneous spectrum analysis (SSA), was designed to locate proteins across spectra, measure their abundance, subtract baselines, exclude irreproducible measurements, and compute normalization factors. Two key parameters are used for feature detection and one parameter each for quality thresholds on spectra and peaks. Compared with other approaches, SSA improved the number and quality of between-group differences among lower signal peaks, and was less likely to introduce systematic bias with normalized spectra [7].

By overcoming technical difficulties, it is predicted that the role of MS in clinical decision-making will substantially increase in the coming years. This is due to not only the extreme sensitivity and high throughput of MS but also because it helps answer clinically relevant questions. Completion of the human genome

project extends medical practice beyond the identification of genes involved in the appearance, progression, and treatment of a disease (genomics). It has now become possible to investigate what these specific genes do and how they interact in communication networks (functional genomics), and the role played by the protein products of those genes in molecular pathways (proteomics) [8]. The human genome contains ~30,000 genes, each creating several transcripts per gene. Generally, these transcripts are not yet functional, but are translated into functional proteins by posttranslational modifications such as proteolysis, glycosylation, phosphorylation, etc., sometimes with great functional impact. The plasma proteome has an important position at the intersection between genes and diseases, and clinical laboratories must adapt to a new era of tests based on proteomics and genomics. Plasma has the potential to come into contact with all cells in the body, and thus can offer pointers to the diagnosis and treatment of disease. With our increasing ability to detect and characterize trace proteins, the discovery of novel therapeutics and biomarkers can be expected [9]. As biomarkers are typically low in abundance, a crucial step of biomarker discovery is to separate clinically relevant sets of proteins that might define disease stages and/or predict disease development. It is anticipated that a multidimensional fractionation system (MDFS) will provide an efficient means of separating low abundance proteins from plasma, resulting in lowering detection limits. However, when using an MDFS to analyze the plasma proteome it is important to consider how sample processing, yield, resolution, and throughput potential may influence the detection limit. In fact, the recent advances in MDFS research could be characterized according to "4RS criterion" (4R: resolution, reproducibility, recovery, and robustness; 4S: simplicity, speed, selectivity, and sensitivity) [10]. Obviously, measurement of a particular set of rigorously validated biomarkers results in a higher level of discriminatory power than a single biomarker. This may be particularly relevant in the context of heterogeneous patient populations and heterogeneous disease states.

MS can also play a major role in new therapeutic approaches. It can be used to predict therapeutic sensitivity to a given therapy, for monitoring drug treatments, or identifying pharmacological interactions. Research is now focusing on the discovery of highly sensitive and specific biomarkers to enable disease detection at the earliest possible stage. We can also expect "tailor-made" individual therapies for the treatment of complex diseases such as cancer. The low molecular weight (LMW) range of the circulatory proteome is a promising source of information in this regard. MS platforms can rapidly map the LMW proteome with high resolution, and we can expect that developments in nanotechnology will enable the amplification and harvesting of these LMW biomarkers, thus laying the foundation for the discovery and characterization of molecules which will improve disease detection and diagnostics [11].

The authors of this chapter are fully aware that methods published in the last few months cannot be considered as routine clinical recommendations. These

papers are pointing to the direction in which medicine (including diagnoses, therapeutic decision-making, or drug development) will change in near future. It may be surprising that most examples of the clinical use of MS are found in pediatrics and oncology. Considering the vast amount of ongoing research, this will probably change in the near future. It is our hope that the compilation of these data will illustrate that MS is likely to have an invaluable role in further development of evidence-based medicine.

2. Pediatrics

Pediatrics has its own special viewpoint for diagnoses and therapy. One example is that pharmacokinetic parameters are age dependent. Thus, age is one of the main factors in selecting drug doses as well as determining sensitivity to drug effects. Consequently, the treatment of the very young, as well as the elderly or pregnant, needs special considerations. Another aspect is that malignant tumors of childhood origin are greatly different from adult-onset malignancies, in terms of clinical course and biological nature. Finally, there are diseases that may occur exclusively or almost 100% in childhood. The most common acidurias and fatty acid oxidation disorders belong to this group. They are already discussed in detail in Chapter 16 and partly also in Chapter 12.

MS is widely used for newborn screening. This information is, however, not sufficiently well known. In a study prenatal care providers were evaluated regarding their attitudes for providing information about newborn screening (which is mostly based on MS/MS). A survey of 6197 prenatal care providers in California regarding their experience with newborn and prenatal screening services showed that 80% of respondents believe newborn screening is very important, only 33% of them discuss it with all their patients. More than 50% believe that either pediatricians (38%) or hospital staff (36%) will do this. Despite state legislation that requires that all pregnant patients should receive the educational booklet, only 61% of responders provided this. Communication about newborn screening to care providers and the public needs to improve [12].

The newborn screening tests vary from country to country, and (within the US) from state to state. In some cases screening is mandatory for only 3 conditions while in others places for as many as 43 diseases. In most cases MS/MS is used for screening. There is still no universally accepted consensus in this issue. Two attempts had been made previously in the US, one in 1975 by the National Academy of Sciences report [13] and the other one in 1988 by the United States Congress Office of Technology Assessment report [14]. Despite rapid developments in many areas including genetics, proteomics, and screening tools, etc., the next recommendation was only published in 2006 [15]. All available data were exhaustively analyzed and evaluated to develop these recommendations. Experts in various areas of medicine

were asked to assess and rank conditions that require screening. Each candidate condition was evaluated according to three main categories: the availability and features of a screening test; availability and complexity of diagnostic tools and services; and possible treatment options and their efficacy. In a two-step approach professionals in the legal, health policy, and public health sectors as well as specialists in primary care, neonatal screening, and pediatrics worked with a steering committee. In the first step a set of principles to guide analysis was developed by fixing the criteria to evaluate conditions. Then, supporting evidence and references from the scientific literature were investigated and compared to the selected criteria by a large group of experts. Data collection and a survey allowed quantification of the expert opinions. This was particularly important since some of the criteria were subjective. Three scoring categories were developed (high scoring, moderately scoring, or low scoring/absence of a newborn screening test) and based on the statistical analysis of data each disease was ranked. In the second step further analyses were carried out regarding the evidence associated with each disease. Detailed information was gathered from different sources (e.g., via systematic reviews of reference lists including MedLine, PubMed, and others; books; Internet searches; professional guidelines; clinical evidence; and cost/economic evidence and modeling). For each condition a fact sheet was prepared, reflecting the outcome of the overall analyses and this was once more reviewed by at least two highly respected experts. These experts reassessed the data, checked the associated references related to each criterion, evaluated the quality of the studies that established the evidence, and assigned a value to the level of evidence. They also made corrections where appropriate (e.g., due to significant variances of the survey data). The information obtained from these two tiers of assessment was then refined by means of technology, condition-specific or cost-effectiveness driven recommendations. Finally each condition was assigned to one of the following categories regarding screening recommendation: (A) core panel (newborn screening is unanimously recommended), (B) secondary targets (those diseases which should be separated from the core panel [differential diagnoses]), and (C) not appropriate for newborn screening (no screening test is available). As the final conclusion, 29 diseases were selected in the core panel, and 25 into the secondary target category. Another 27 conditions were determined to be inappropriate for newborn screening at present. The 29 diseases for which screening is recommended as well as the screening method suggested are seen in Table 1.

In diagnoses of some rare conditions, such as bile acid synthetic defects, MS can also be utilized. Nowadays it is possible to screen and rapidly diagnose potential or real inborn errors in bile acid synthesis from urinary bile acid analysis by means of MS. Specific mutations in the genes that encode the enzymes responsible for bile acid synthesis can be identified by molecular techniques. Of the seven known genetic defects that cause progressive cholestatic liver disease, syndromes of fat-soluble vitamin malabsorption, and neurological disease, six have been properly characterized [16].

Table 1
Childhood diseases primarily recommended for screening in ref. [15]

Disease	Screening test
(1) Isovaleric acidemia	MS/MS
(2) Glutaric acidemia type I	MS/MS
(3) 3-Hydroxy-3-methyl glutaric aciduria	MS/MS
(4) Multiple carboxylase deficiency	MS/MS
(5) Methylmalonic acidemia	MS/MS
(6) 3-Methylcrotonyl-CoA-carboxylase deficiency	MS/MS
(7) Methylmalonic acidemia (CblA, B)	MS/MS
(8) Propionic acidemia	MS/MS
(9) β-Ketothiolase deficiency	MS/MS
(10) Medium-chain acyl-CoA dehydrogenase deficiency	MS/MS
(11) Very long-chain acyl-CoA dehydrogenase deficiency	MS/MS
(12) Long-chain 3-OH acyl-CoA dehydrogenase deficiency	MS/MS
(13) Trifunctional protein deficiency	MS/MS
(14) Carnitine uptake defect	MS/MS
(15) Phenylketonuria	MS/MS, fluorometric, enzyme
(16) Maple syrup (urine) disease	MS/MS
(17) Homocystinuria	MS/MS
(18) Citrullinemia	MS/MS
(19) Argininosuccinic acidemia	MS/MS
(20) Tyrosinemia type I	MS/MS
(21) Hb SS disease (sickle cell anemia)	HPLC, IEF
(22) Hb S/β-thalassemia	HPLC, IEF
(23) Hb S/C disease	HPLC, IEF
(24) Congenital hypothyreoidism	RIA, ELISA
(25) Congenital adrenal hyperplasia	RIA, ELISA, MS/MS
(26) Biotinidase deficiency	Colorimetric assay, MS/MS (inconsistent)
(27) Classic galactosemia	Microbiological for G-1-P, and galactose and fluorometric assays for GALT activity
(28) Hearing loss	Audiometry
(29) Cystic fibrosis	Immunoreactive trypsinogen + second tier DNA

Note: (1–9) Organic acid disorders, (10–14) fatty acid oxidation disorders, (15–20) amino acid disorders, (21–23) hemoglobinopathies, (24 and 25) endocrinopathy, (26) other inborn error of metabolism, (27) carbohydrate disorders, (28) miscellaneous genetic conditions, and (29) infectious diseases. MS/MS: tandem mass spectrometry, HPLC: high pressure liquid chromatography, IEF: isoelectrofocusing, RIA: radioimmuno assay, and ELISA: enzyme-linked immunosorbent assay.

Inborn metabolic disorders of the pyrimidine degradation pathway were evaluated in children with unspecific neurological symptoms. Stable-isotope-labeled reference compounds were used as internal standards to determine uracil and thymine as well as their degradation products in urine by means of reversed-phase HPLC coupled with electrospray ionization MS/MS. Data obtained from the control group were used to develop age-related reference ranges of all pyrimidine degradation products. The study was able to identify patients with ornithine transcarbamylase deficiency based on the elevated level of uracil, dihydrouracil and β-ureidopropionate, and dihydropyrimidine dehydrogenase (DPYD) deficiency. Treatment-related increase of β-alanine was detected in the urine of a number of patients. The authors conclude that in children with unexplained neurological symptoms, especially epileptic seizures with or without psychomotor retardation, pyrimidine metabolites should be quantitatively investigated. The MS-based method and the age-related reference ranges provide a useful tool for diagnosis in clinical practice to detect of partial enzyme deficiencies [17].

3. Oncology

Oncology is one of the most innovative fields of medicine. For effective cancer drug therapy it is a prerequisite to thoroughly understand tumor biology, cell kinetics, pharmacology, and drug resistance. In many cases insufficient therapeutic results are the main driving force for exploring new therapeutic possibilities or research tools. Thus, oncology participated at the birth of, among others, controlled clinical trials, immunotherapy, antiapoptotic therapy, antiangiogenic therapy, genomics (including pharmacogenomics), proteomics, bioinformatics, etc. The key issues for further progress of oncology are the assignment of new therapeutic targets, developing and evaluating novel treatment entities, and combining these with existing therapies. Rational drug treatment indicates that the selection of therapy is based on considerations of mechanism of action, pharmacokinetics, interactions, and the side effect profile of the drug applied. In other words, without deep understanding of the drug used no rational drug treatment can be performed. A profound knowledge of drugs, however, is not sufficient for optimal outcome. In addition, patients' expectations, their relation to the disease, their psychic status, cultural and educational background, family support as well as concomitant diseases, physical performance status, etc., should be known and considered. In fact, modern oncology should combine the most advanced therapeutic innovations with a holistic approach.

Thanks to proteomic applications in diagnosis of cancer, several research groups have identified proteomic patterns associated with ovarian, prostatic, colorectal, lung, and other cancers. While the sensitivity and specificity of these patterns are

highly satisfactory, there are still some open questions concerning standardization, reproducibility, and inter-laboratory agreement of these data.

Colorectal cancer (CRC) can be screened by detecting blood in the stool, but this is not specific for gastro-intestinal cancer since many other diseases may yield a similar outcome. The low specificity and sensitivity of the presently used carcinoembryonic antigen test make it a not very good biomarker for detection of CRC. Various proteomic approaches have been developed and evaluated for distinguishing individuals with CRC from healthy individuals, based on simultaneous detection and analysis of multiple proteins. In one investigation serum samples were studied by SELDI–MS. In order to separate the healthy group from the CRC patients a multilayer artificial neural network with a back propagation algorithm was developed. The healthy samples were separated from the CRC samples with a specificity of 93% and sensitivity of 91%. The four top-scoring peaks in the SELDI spectra were selected as the potential detection "fingerprints." This combination of SELDI–MS with artificial neural networks was shown to be an efficient technique for detection and diagnosis of CRC [18]. In another study to characterize the serum proteomic patterns of CRC and tumor staging, SELDI–MS technology was coupled to a CM10 ProteinChip. Patients with different stages of the disease were investigated. Stage models were developed and validated. Model I comprised six protein peaks and could distinguish local CRC patients (Stages I and II) from regional CRC patients (Stage III) with 86.67% accuracy. Model II comprised three protein peaks and could distinguish locoregional CRC patients (Stages I–III) from metastatic CRC patients (Stage IV) with 75% accuracy. Further models were developed to distinguish Stages I and II; I and III; II and III; II and IV; III and IV. Different stage groups could also be distinguished by two-dimensional scatter-plots. This method is applied in the preoperative phase [19].

MS techniques in clinical treatment are used not only to *diagnose or stage CRC* patients but also to *predict the efficacy of chemotherapy*. Oxaliplatin, for example, is a Pt-containing anticancer drug for treating advanced CRC. An association between the levels of oxaliplatin–protein complexes in patients and treatment efficacy was reported in a study using size-exclusion HPLC with ICP–MS and nano-ESI–MS. Blood samples from 19 CRC patients were collected at 1 and 48 h following infusion of oxaliplatin. HPLC/ICP–MS quantification of the oxaliplatin–protein complexes showed reduction in the levels of Pt–protein complexes in plasma samples at 48 h of ca. 50% compared to those at 1 h, and no significant change in hemolysates. The concentrations of hemoglobin (Hb)–oxaliplatin complexes ranged from 3.1 to 8.7 μmol. Three distinct mass spectral profiles of the Hb–oxaliplatin complexes were identified by nano-ESI–MS analysis of the hemolysates. Multivariate analysis of the potential predictors showed that Hb–oxaliplatin complex concentration, performance status, baseline neutrophil count, and whether the site of the primary cancer was the colon or rectum were the statistically significant variables. The hazard ratio of 2.4 for the concentration of the

Hb–oxaliplatin complexes indicates that an increased amount of Hb–oxaliplatin complexes in patients closely correlates with an enhanced risk of cancer progression. The level of the Hb–oxaliplatin complexes in erythrocytes is therefore a potential biomarker for indicating inter-patient variations in the treatment efficacy of oxaliplatin [20].

The *mechanism of action* of anticancer agents can also be investigated by means of MS. Interaction of intact human holo-transferrin (holo-Tf) with oxaliplatin was reported; the complex comprised an intact holo-Tf and an oxaliplatin molecule and was detected using nanospray ionization MS. The molecular weight of this complex was 80,077 Da, an increase of 397 mass units compared to the 79,680 Da protein alone. This indicated that a parent drug molecule was bound to the intact protein. Interaction between the intact protein and oxaliplatin was further examined using size-exclusion HPLC coupled to ICP–MS. HPLC was used to separate the protein complex and free oxaliplatin, followed by quantitative determination by simultaneous ICP–MS monitoring of 195Pt and 56Fe. Pt and Fe signals were detected at the same retention time, identifying the protein–drug complex. The Fe signal was not affected by an increase in the incubation time of the reaction mixture containing holo-Tf and oxaliplatin, while the Pt signal increased over time, and the authors concluded that formation of this complex does not affect the protein-bound Fe. The nanospray and ICP–MS results are evidence that holo-Tf and oxaliplatin molecules form complexes through noncovalent binding; therefore, holo-Tf may be a useful carrier for oxaliplatin delivery [21].

MS can also be used for *optimizing cancer therapy*. A method for the quantification of plasma 2′-deoxyuridine (UdR) has been developed and validated. Only 1 ml plasma is required, which is subjected to a clean-up step with anion-exchange solid-phase extraction followed by HPLC separation and atmospheric pressure CI-MS detection in a selected-ion monitoring mode. The method has the sensitivity, precision, accuracy, and selectivity required for routine analysis, the limit of quantitation being 5 nmol/l, which is certainly sufficient for clinical studies. Cancer patients treated with the fluoropyrimidine analog capecitabine (N4-pentoxycarbonyl-5′-5-fluorocytidine) have significantly elevated plasma UdR after 1 week of treatment, which is consistent with inhibition of thymidylate synthase (TS). The authors suggest that the mechanism of antiproliferative toxicity of capecitabine is at least partly due to TS inhibitory activity of its active metabolite 5-fluoro-2′-deoxyuridine monophosphate. Monitoring of plasma UdR concentrations can help clinicians to optimize scheduling of capecitabine or other TS inhibitors in clinical trials. They also found marked differences of plasma UdR between humans and rodents. This simple, selective, and sensitive method facilitates pharmacodynamic studies of TS inhibitors [22].

Pancreatic adenocarcinoma is one of the most devastating and rapidly progressive forms of cancer. Currently, fewer than 5% of patients survive more than 5 years after diagnosis, mainly because most patients present with advanced disease.

Consequently, early diagnosis may improve their prognosis. In a recent study investigators aimed to identify unique, tissue-specific protein biomarkers capable of differentiating pancreatic adenocarcinoma (PC) from adjacent uninvolved pancreatic tissue (AP), benign pancreatic disease (B), and nonmalignant tumor tissue (NM). Tissue samples were analyzed on hydrophobic protein chip arrays by SELDI TOF–MS. They found that 13 protein peaks differentially expressed between PC and AP, 8 between PC and B, and 12 between PC and NM tissue. Using logistic regression and cross-validation they were able to identify overlapping panels of peaks to develop a training model that distinguished PC from AP (77.4% sensitivity, 84.1% specificity), B (83.9% sensitivity, 78.9% specificity), and NM tissue (58.1% sensitivity, 90.5% specificity). The final panels selected correctly classified 80.6% of PC and 88.6% of AP samples, 93.5% of PC and 89.5% of B samples, and 71.0% of PC and 92.1% of NM samples. Identification of these proteins is important to understanding the biology of pancreatic cancer. The authors conclude that these protein panels could have important diagnostic implications [23]. In another attempt to improve serological diagnosis of pancreatic cancer, SELDI protein chip MS was used to analyze serum samples from patients with and without pancreatic cancer. Serum samples from patients with resectable pancreatic adenocarcinoma were compared with samples from age- and sex-matched patients with nonmalignant pancreatic diseases, as well as healthy controls. The number of proteins that could potentially be identified was increased by a fractionation process using anion exchange and profiling on two ProteinChip surfaces (metal affinity capture and weak cation exchange). A set of protein peaks could discriminate between patient groups. The unified maximum separability algorithm compared the performance of the individual marker panels alone or in conjunction with CA19-9. The two most discriminating protein peaks for distinguishing between patient groups, as identified by SELDI profiling, could differentiate patients with pancreatic cancer from healthy controls with a sensitivity of 78% and specificity of 97%. They performed significantly better than the current standard serum marker, CA19-9. The investigators could further improve the diagnostic accuracy of the two markers by using them in combination with CA 19-9. A combination of three SELDI markers and CA19-9 was superior to CA19-9 alone. SELDI markers were superior to CA19-9 in distinguishing pancreatic cancer from pancreatitis. The investigators concluded that SELDI profiling of serum can be used to accurately differentiate individuals with pancreatic cancer from those with other pancreatic diseases and from healthy controls [24].

Breast cancer is one of the main causes of cancer-related death for woman affecting more than 1 million females annually throughout the world. To screen for and identify treatment-responsive proteins, the protein expression profile of serum from breast cancer patients was determined after 4, 8, 24, and 48 h after docetaxel infusion using SELDI–MS. The relative expression levels of target proteins were compared across time. Two representative proteins, kininogen and apolipoprotein

A-II, were identified. Protein expression profiles determined by MS are thus useful for identifying treatment-responsive proteins [25]. Toxicity can be severe in case of chemotherapy. MS-based techniques may also be applied to reduce side effects. A study evaluated the association between exposure to unbound docetaxel and neutropenia in cancer patients and identified factors influencing unbound docetaxel clearance. Pharmacokinetic studies and toxicity assessments were performed during the first cycle of therapy. Total docetaxel concentrations were determined by HPLC–MS–MS. The authors conclude that as exposure to unbound docetaxel is closely related to drug-induced hematologic toxicity, this needs to be considered in future pharmacological investigations [26].

Ricolleau et al. have looked for novel *prognostic biomarkers* to help direct treatment decisions by typing subgroups of node-negative breast cancer patients. In a proteomic approach SELDI–MS was used to identify differentially expressed proteins with a prognostic impact in node-negative breast cancer patients with no relapse vs. patients with metastatic relapse. Ubiquitin and ferritin light chain (FLC) proved to be interesting in this regard. Their differential expression was further confirmed by Western blotting analyses and immunohistochemistry. The mass spectrometric protein profiling in this study shows that a high level of cytosolic ubiquitin and/or a low level of FLC indicate a good prognosis [27].

In another study SELDI–MS was employed in a comparative analysis of lobular invasive vs. ductal invasive breast tumor tissue samples. The aim was to identify differentially expressed proteins and peptides, and to validate the technique for biomarker identification. Mass signals corresponding to an estimated 140 native peptides and proteins in each tumor were identified. Only 14% of the mass signals were present in more than six samples of either HMEC or MCF-7, showing a large degree of great heterogeneity. The authors conclude that the low amount of identified peptides and proteins and the large heterogeneity suggest that SELDI is not well suited for biomarker identification in complex samples [28].

Prognostic markers of the aggressive phenotype of HER-2/neu-positive breast cancer were also studied by MS. It is known that the tyrosine kinase receptor ErbB2 (HER-2/neu) is overexpressed in up to 30% of breast cancers and is associated with poor prognosis and an increased likelihood of metastasis, especially in node-positive tumors. Differentially expressed proteins in two subsets of tumor cells from HER-2/neu-positive and HER-2/neu-negative tumors were identified by 2D electrophoresis and MALDI–TOF/TOF MS/MS. Differential expression of several key cell cycle modulators was found, which were linked with increased proliferation of the HER-2/neu-overexpressing cells. The findings suggest that HER-2/neu signaling may result in enhanced activation of various metabolic, stress-responsive, antioxidative, and detoxification processes within the breast tumor microenvironment. Thus, it was hypothesized that these identified changes in the cellular proteome are likely to drive cell proliferation and tissue invasion and that the key cell cycle modulators might serve as useful targets for the development

of therapeutic strategies to negate the metastatic potential of HER-2/neu-positive breast tumors [29].

One-third to half of early breast cancer (EBC) patients are considered to be at high risk of developing metastatic recurrence. Work is being done to improve predicting of clinical outcomes in order to optimize and tailor therapeutic strategies. Goncalves et al. identified a protein signature that correlates with metastatic relapse, using SELDI profiling of early postoperative serum from high-risk EBC patients. Several protein peaks were differentially expressed. Using chemometric (bioinformatics) tools, a multiprotein model was built that correctly predicted outcome in 83% of patients. The multiprotein index was used to classify the "good prognosis" and "poor prognosis" patients whose 5-year metastasis-free survival rates were 83 and 22%, respectively. The authors conclude that the postoperative serum protein pattern could have an important prognostic value in high-risk EBC [30].

Not all biomarkers are peptides or proteins. Various estrogens may also serve as diagnostic or treatment indicative tools. Also, individual patterns of estrogen metabolism can influence an individual's risk of developing breast cancer. An HPLC–MS method to measure the concentrations of 15 endogenous estrogens and their metabolites in human urine has been developed and validated. The limit of quantitation for each estrogen is 0.02 ng/0.5 ml urine sample, which is well within clinical relevance. This method gives accurate, precise, and specific measurements of endogenous estrogen metabolites. It will be useful in future research on breast cancer prevention, screening, and treatment [31].

Gene expression analysis is considered a promising tool for predicting the clinical course of malignant disease and the response to antineoplastic therapy. Very little information is available regarding the protein expression pattern of human tumors. Proteins of interest can now be identified by their expression and/or modification pattern in 2-DE. Hudelist et al. identified a proteomic pattern that is characteristic for malignant breast epithelium by differential 2-DE analysis of sets of microdissected malignant breast epithelia and corresponding adjacent normal breast epithelia from patients with invasive breast carcinoma. They found that 32 protein spots were selectively regulated in malignant epithelium. MALDI–TOF and/or immunoblotting for protein identification was then applied, and identified 13 proteins that were not previously associated with breast cancer. This brings us a step further in understanding oncogenesis. In addition, this strategy can be used in the characterization of the malignant phenotype of individual tumors, and thereby identify novel targets for antineoplastic therapy [32].

Leptomeningeal metastasis (LM) occurs in 5% of patients with breast cancer. This complication can lead to neurological deterioration without early diagnosis and treatment. This is complicated by the fact that 25% of cerebrospinal fluid (CSF) samples produce false-negative results when examined cytologically. Dekker et al. have developed an MS-based method to investigate protein expression

patterns in the CSF from patients with breast cancer with and without LM. CSF samples from these patients and controls were digested with trypsin, and the resulting peptides were quantified by MALDI–TOF MS. Mass spectral analysis and a comparison between patient groups with bioinformatics tools showed 895 possible peak positions, of which 164 discriminated between the patient groups. On the basis of these discriminatory masses, a classifier was built to distinguish breast cancer patients with and without LM, having a maximum accuracy of 77% with a sensitivity of 79% and a specificity of 76%. This is a step forward in diagnosing LM in patients with breast cancer. This method is transferable to diagnostic assays for other neurological disorders [33].

Currently, no satisfactory biomarkers are available to screen for *lung cancer*. Serum SELDI proteomic patterns have been applied to distinguish lung cancer patients from healthy individuals. Serum samples from lung cancer patients and controls were randomly divided into a training set and a blinded test set, both of which included sera from patients with Stages I/II lung cancer, Stages III/IV lung cancer, and healthy controls. Five protein peaks were automatically chosen as a biomarker pattern in the training set. When the SELDI marker pattern was tested with the blinded test set, sensitivity was 86.9%, and specificity 80.0%, with a positive predictive value of 92.4%. The SELDI marker pattern showed a sensitivity of 91.4% in the detection of nonsmall cell lung cancers. For lung cancers in Stages I/II detection sensitivity was 79.1%. SELDI–TOF–MS can thus be considered a potential tool for the screening of lung cancer [34].

References

1. Tanaka, K., Budd, M.A., Efron, M.L. and Isselbacher, K.J., Isovaleric acidemia: a new genetic defect of leucine metabolism. *Proc. Natl. Acad. Sci. U.S.A.*, **56**, 236–242 (1966).
2. Gaskell, S.J., Guenat, C., Millington, D.S., Maltby, D.A. and Roe, C.R., Differentiation of isomeric acylcarnitines using tandem mass spectrometry. *Anal. Chem.*, **58**, 2801–2805 (1986).
3. Millington, D.S., Maltby, D.A. and Roe, C.R., Rapid detection of argininosuccinic aciduria and citrullinuria by fast atom bombardment and tandem mass spectrometry. *Clin. Chim. Acta*, **155**, 173–178 (1986).
4. Roe, C.R., Millington, D.S., Maltby, D.A. and Kinnebrew, P., Recognition of medium-chain acyl-*o*-dehydrogenase deficiency in asymptomatic siblings of children dying of sudden infant death or Reye-like syndromes. *J. Pediatr.*, **108**, 13–18 (1986).
5. Petricoin, E.F. and Liotta, L.A., Proteomic approaches in cancer risk and response assessment. *Trends Mol. Med.*, **10**(**2**), 59–64 (2004).
6. Hong, H., Dragan, Y., Epstein, J., Teitel, C., Chen, B., Xie, Q., Fang, H., Shi, L., Perkins, R. and Tong, W., Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of flight (TOF) mass spectrometry (MS). *BMC Bioinformatics*, **6**(**Suppl. 2**), S5 (2005).
7. Carlson, S.M., Najmi, A., Whitin, J.C. and Cohen, H.J., Improving feature detection and analysis of surface-enhanced laser desorption/ionization–time of flight mass spectra. *Proteomics*, **5**(**11**), 2778–2788 (2005).

8. Righetti, P.G., Castagna, A., Antonucci, F., Piubelli, C., Cecconi, D., Campostrini, M., Rustichelli, C., Antonioli, P., Zanusso, G., Monaco, S., Lomas, L. and Boschetti, E., Proteome analysis in the clinical chemistry laboratory: myth or reality? *Clin. Chim. Acta*, **357**(**2**), 123–139 (2005).
9. Lathrop, J.T., Hayes, T.K., Carrick, K. and Hammond, D.J., Rarity gives a charm: evaluation of trace proteins in plasma and serum. *Expert Rev. Proteomics*, **2**(**3**), 393–406 (2005).
10. Lee, H.J., Lee, E.Y., Kwon, M.S. and Paik, Y.K., Biomarker discovery from the plasma proteome using multidimensional fractionation proteomics. *Curr. Opin. Chem. Biol.*, **10**(**1**), 42–49 (2006).
11. Calve, K.R., Liotta, L.A. and Petricoin, E.F., Clinical proteomics: from biomarker discovery and cell signaling profiles to individualized personal therapy. *Biosci. Rep.*, **25**(**1–2**), 107–125 (2005).
12. Faulkner, L.A., Feuchtbaum, L.B., Graham, S., Bolstad, J.P. and Cunningham, G.C., The newborn screening educational gap: what prenatal care providers do compared with what is expected. *Am. J. Obstet. Gynecol.*, **194**(**1**), 131–137 (2006).
13. National Research Council, Committee for the Study of Inborn Errors of Metabolism, *Genetic Screening: Programs, Principles, and Research*. National Academy of Sciences, Washington, DC, 1975.
14. U.S. Congress, Office of Technology Assessment, *Healthy Children: Investing in the Future, Appendix 1: Data and Methods used in OTA'S Cost-Effectiveness Analysis of Strategies for Newborn Screening*. Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402-9325, 1988, available at http://www.wws.princeton.edu/cgi-bin/byteserv.prl/ota/disk2/1988/8819/881919.PDF, accessed April 7, 2006.
15. Watson, M.S., Lloyd-Puryear, M.A., Mann, M.Y., Rinaldo, P. and Howell, R.R., Main report. *Genet. Med.*, **8**(**Suppl. 5**) 12S–252S (2006).
16. Setchell, K.D. and Heubi, J.E., Defects in bile acid biosynthesis—diagnosis and treatment. *J. Pediatr. Gastroenterol. Nutr.*, **43**(**Suppl. 1**), S17–S22 (2006).
17. Schmidt, C., Hofmann, U., Kohlmuller, D., Murdter, T., Zanger, U.M., Schwab, M. and Hoffmann, G.F., Comprehensive analysis of pyrimidine metabolism in 450 children with unspecific neurological symptoms using high-pressure liquid chromatography–electrospray ionization tandem mass spectrometry. *J. Inherit. Metab. Dis.*, **28**(**6**), 1109–1122 (2005).
18. Chen, Y.D., Zheng, S., Yu, J.K. and Hu, X., Artificial neural networks analysis of surface-enhanced laser desorption/ionization mass spectra of serum protein pattern distinguishes colorectal cancer from healthy population. *Clin. Cancer Res.*, **10**(**24**), 8380–8385 (2004).
19. Xu, W.H., Chen, Y.D., Hu, Y., Yu, J.K., Wu, X.G., Jiang, T.J., Zheng, S. and Zhang, S.Z., Preoperatively molecular staging with CM10 ProteinChip and SELDI–TOF–MS for colorectal cancer patients. *J. Zhejiang Univ. Sci. B*, **7**(**3**), 235–240 (2006).
20. Mandal, R., Sawyer, M.B. and Li, X.F., Mass spectrometry study of hemoglobin–oxaliplatin complexes in colorectal cancer patients and potential association with chemotherapeutic responses. *Rapid Commun. Mass Spectrom.*, **20**(**17**), 2533–2538 (2006).
21. Zhao, Y.Y., Mandal, R. and Li, F.X., Intact human holo-transferrin interaction with oxaliplatin. *Rapid Commun. Mass Spectrom.*, **19**(**14**), 1956–1962 (2005).
22. Li, K.M., Rivory, L.P. and Clarke, S.J., Rapid quantitation of plasma 2′-deoxyuridine by high-performance liquid chromatography/atmospheric pressure chemical ionization mass spectrometry and its application to pharmacodynamic studies in cancer patients. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **820**(**1**), 121–130 (2005).
23. Scarlett, C.J., Smith, R.C., Saxby, A., Nielsen, A., Samra, J.S., Wilson, S.R. and Baxter, R.C., Proteomic classification of pancreatic adenocarcinoma tissue using protein chip technology. *Gastroenterology*, **130**(**6**), 1670–1678 (2006).
24. Koopmann, J., Zhang, Z., White, N., Rosencweig, J., Fedarko, N., Jagannath, S., Canto, M.I., Yeo, C.J., Chan, D.W. and Goggins, M., Serum diagnosis of pancreatic adenocarcinoma using

surface-enhanced laser desorption and ionization mass spectrometry. *Clin. Cancer Res.*, **10**(**3**), 860–868 (2004).

25. Heike, Y., Hosokawa, M., Osumi, S., Fujii, D., Aogi, K., Takigawa, N., Ida, M., Tajiri, H., Eguchi, K., Shiwa, M., Wakatabe, R., Arikuni, H., Takaue, Y. and Takashima, S., Identification of serum proteins related to adverse effects induced by docetaxel infusion from protein expression profiles of serum using SELDI ProteinChip system. *Anticancer Res.*, **25**(**2B**), 1197–1203 (2005).
26. Baker, S.D., Li, J., ten Tije, A.J., Figg, W.D., Graveland, W., Verweij, J. and Sparreboom, A., Relationship of systemic exposure to unbound docetaxel and neutropenia. *Clin. Pharmacol. Ther.*, **77**(**1**), 43–53 (2005).
27. Ricolleau, G., Charbonnel, C., Lode, L., Loussouarn, D., Joalland, M.P., Bogumil, R., Jourdain, S., Minvielle, S., Campone, M., Deporte-Fety, R., Campion, R. and Jezequel, P., Surface-enhanced laser desorption/ionization time of flight mass spectrometry protein profiling identifies ubiquitin and ferritin light chain as prognostic biomarkers in node-negative breast cancer tumors. *Proteomics*, **6**(**6**), 1963–1975 (2006).
28. Traub, F., Feist, H., Kreipe, H.H. and Pich, A., SELDI–MS-based expression profiling of ductal invasive and lobular invasive human breast carcinomas. *Pathol. Res. Pract.*, **201**(**12**), 763–770 (2005).
29. Zhang, D., Tai, L.K., Wong, L.L., Chiu, L.L., Sethi, S.K. and Koay, E.S., Proteomic study reveals that proteins involved in metabolic and detoxification pathways are highly expressed in HER-2/neu-positive breast cancer. *Mol. Cell Proteomics*, **4**(**11**), 1686–1696 (2005).
30. Gonclaves, A., Esterni, B., Bertucci, F., Sauvan, R., Chabannon, C., Cubizolles, M., Bardou, V.J., Houvenaegel, G., Jacquemier, J., Granjeaud, S., Meng, X.Y., Fung, E.T., Birnbaum, D., Maranichi, D., Viens, P. and Borg, J.P., Postoperative serum proteomic profiles may predict metastatic relapse in high-risk primary breast cancer patients receiving adjuvant chemotherapy. *Oncogene*, **25**(**7**), 981–989 (2006).
31. Xu, X., Veenstra, T.D., Fox, S.D., Roman, J.M., Issaq, H.J., Falk, R., Saavedra, J.E., Keefer, L.K. and Ziegler, R.G., Measuring fifteen endogenous estrogens simultaneously in human urine by high-performance liquid chromatography–mass spectrometry. *Anal. Chem.*, **77**(**20**), 6646–6654 (2005).
32. Hudelist, G., Singer, C.F., Pischinger, K.I., Kaserer, K., Manavi, M., Kubista, E. and Czerwenka, K.F., Proteomic analysis in human breast cancer: identification of a characteristic protein expression profile of malignant breast epithelium. *Proteomics*, **6**(**6**), 1989–2002 (2006).
33. Dekker, L.J., Boogerd, W., Stockhammer, G., Dalebout, J.C., Ciccama, I., Zheng, P., Bonfrer, J.M., Verschuuren, J.J., Jenster, G., Verbeek, M.M., Luider, T.M. and Smitt, P.A., MALDI–TOF mass spectrometry analysis of cerebrospinal fluid tryptic peptide profiles to diagnose leptomeningeal metastases in patients with breast cancer. *Mol. Cell Proteomics*, **4**(**9**), 1341–1349 (2005).
34. Yang, S.Y., Xiao, X.Y., Zhang, W.G., Zhang, L.J., Zhang, W., Zhou, B., Chen, G. and He, D.C., Application of serum SELDI proteomic patterns in diagnosis of lung cancer. *BMC Cancer*, **5**, 83 (2005).

This page intentionally left blank

Part V
Emerging Areas

This page intentionally left blank

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 22

Biomarker discovery

BRADLEY J. THATCHER[a,*] and EMILIA CAPUTO[a,1]

[a]*Institute of Genetics and Biophysics-IGB, CNR, Naples, Italy*

*Corresponding author. Address: C.da Forma, Villa Iris, 82034 Guardia, Sanframondi (BN), Italy; Tel.: 00 39335 7441346. E-mail: brad_thatcher@hotmail.com (B.J. Thatcher).
[1]caputo@igb.cnr.it (E. Caputo).

1. Introduction

The interest in biomarker discovery has increased in the last several years. This is especially true in the area of clinical research where biomarkers can be used to affect patient care. Biomarkers are tools for the diagnosis, monitoring of therapy, and screening for a number of diseases. Biomarkers can also be used to detect disease risk factors allowing the physician to recommend or prescribe more intensive monitoring or testing of a patient. Over the past 20 years we have seen the emergence of tests based on disease biomarkers in the clinic. Several molecules have been identified as markers, but only some of them are actually good enough [1] to be used as screening tests that have become standards of care in many countries and thus are used to examine entire populations.

1.1. Chemical background for medical doctors

The primary issues of concern at the start of a biomarker discovery experiment are the clinical questions. It is worth the time to stop and think for a moment about the biochemistry of disease and the chemistry of potential biomarkers associated with disease. There are very significant chemical differences between classes of biomolecules and this affects the design and experimental protocols that are used to find, separate, identify, and assay each compound that may be of interest.

The majority of separation tools available to the biochemist for the simplification of biological material are dependent on the physical and chemical properties of the molecules being studied. It is important to understand that it is not possible in most cases to look at diverse classes of molecules at the same time and from the same sample preparation methods. Separation is required during the process of biomarker discovery since most biological samples are much complex to allow direct analysis. Separation tools continue to be improved but are a long way from the level of resolution required to separate even a few percent of the diversity of a single class of compound at one time. The logical extension of this is simply to say that we will use multiple runs to separate all the components. This idea, although logical, does not work well in the real world since each separation causes significant sample loss; the use of three or four methods one after another will consume the majority of the sample and leave little to be analyzed. Thus, one should discuss the separation options for a discovery experiment before sample collection has started so that the quantity and type of sample collected will be of maximal utility. This does not preclude the use of archived samples for validation and biomarker purification and characterization, but it is best to collect the discovery samples for optimal analysis.

The tools used to detect biomarkers are also varied, and it might not be possible to adapt instrumentation from one purpose to another. This is also something to discuss with a biochemist, chemist before the research plan is decided. Even with these limitations, well-designed and executed biomarker projects have been very successful in the past and will continue to be more successful in the future.

The key to this success is a multidisciplinary approach to biomarker discovery where the different members of a team communicate well and try to understand what will and will not work and what is and is not possible before planning and starting the project.

1.2. Basic medical aspects

Medicine and disease pathology are immensely complicated issues. In either of these cases there are often a lot of questions for which there are no clear answers. This is why the medical field is devoting so much time, effort, and money in the search for biomarkers and the development of diagnostics. Disease pathology and the progression of diseases can vary from case to case, but there are often a number of similarities; this is in fact how diseases are classified. It is a good idea to have a discussion with a medical doctor about the disease, its symptoms, and progression to develop an understanding of some of the variables that might be observed in the samples. It is also noteworthy that in the process of biomarker discovery it is the consistency of potential biomarkers between disease samples that one is trying to identify and not the variability between disease samples. Bioinformatics will help in this process in two ways: by identifying molecules that are consistent in the disease and normal samples, and by focusing on the most significant changes between the control and disease samples.

From a chemical and biochemical point of view, biomarker discovery use a range of familiar techniques and tools that have been around for a long time such as SDS-PAGE, HPLC, mass spectrometry, etc. What is new is the number of samples and the data analysis that are often new and complicated. Biological systems especially involving human subjects are highly variable and lack a much of the consistency that is seen in other experimental systems such as animal models. Thus, it is important to choose methods that can be applied to large numbers of samples in order to deal with the statistical requirements of human samples. In a clinical study it is not uncommon to think of hundreds or thousands of samples to validate a biomarker. This will require the use of bioinformatics to analyze and interpret data and to keep track of the vast amount of data and information that will be used in the process of biomarker discovery and validation.

1.3. Basic concepts

The search for biomarkers is not a simple task. There are several challenges in the pathway to a validated biomarker. These challenges include sample availability, the use of large numbers of samples for validation, technical issues in experimental design, assay development, the necessity for bioinformatics, and asking the right question so that the results of the experiment will have meaning and be of practical use in the clinic.

The recent advancements in separation technology, mass spectrometry, and informatics for the biological sciences have been very useful in accomplishing this

task [2,3]. Mass spectrometry has improved the ability to detect, quantify, and identify biomarkers with an increased speed and sensitivity. These technologies have become much more user-friendly and their availability to researchers has increased. These advancements have facilitated biomarker discovery on clinically relevant numbers of samples [4–6].

However, biomarker discovery is one of the most difficult types of projects in biology. This is partly due to the level of complexity and inherent inconsistencies that are present in biological systems [7]. In addition, the pathology of a disease is rarely simple and there can be closely related conditions that complicate the diagnosis and thus the search for biomarkers.

To this end it is important to take a multidisciplinary approach to biomarker discovery. There are a number of difficult tasks to be accomplished in the process of biomarker discovery, each requiring expertise in different fields, separation technology, medicine, pathology of the disease, the chemistry of the type of molecule that is the target of the study, and statistical analysis. The conclusion from these facts is that a multidisciplinary approach to biomarker discovery is necessary. Once the group is assembled, it is then vital that members of the group *listen* to each other about the capabilities and weaknesses in each area of a project before starting the actual work.

This chapter will focus on assisting in providing some suggestions to minimize these challenges and hopefully assist in the design of successful biomarker discovery projects. We will not cover detailed methodologies that are admirably covered by other contributors to this work. What we will concentrate on are the philosophy and practical aspects of designing a successful biomarker discovery project; we will provide some suggestions concerning ways to validate the prospective biomarkers to prepare for assay development and the use of biomarkers in the clinic. We will also try and identify and explain some of the biggest challenges faced by a researcher during the development of a biomarker discovery experiment.

2. Biomarkers in medicine

The utility of biomarkers is the reason for the funding of projects based on new biomarkers, which has increased in the last few years. Once a successful biomarker diagnostic is developed the cost for screening large populations is reduced and this results in lower health-care costs in the short term. In the long term, this will facilitate the increased use of screening tests for a greater number of diseases. The increased screening of populations facilitates early diagnosis, better control of chronic conditions, and improved health in large populations, thereby improving patient care and reducing the cost of health care. Current screening tools can be resource intensive, expensive, and requiring expert examination of the data (mammograms, ultrasound, biopsy, etc.), and this causes their use to be limited. A biomarker test is generally less

expensive and invasive than other forms of testing facilitating a more efficient use of medical resources while improving health care. This becomes clear when we look at the use of diagnostic tests in the clinic. The measurement of cholesterol, for example, is so inexpensive and noninvasive that many people are tested twice a year in their entire life. This results in early detection of a problem and early treatment reducing the risk of disease resulting from high cholesterol.

The ultimate goal of biomarker discovery is the development of screening tests to detect diseases before they are symptomatic or at a stage where they are more effectively treated. This is an ambitious goal and will require many years or decades of basic and medical research. Each discovered and validated biomarker is a contribution toward this goal.

The short-term goal of many researchers in the field of medical biomarker discovery is to achieve two things. The first is the discovery and validation of biomarkers for diagnostic/prognostic purposes and to improve patient care. The second is to provide information that will assist in the understanding of the pathology of the disease. Both of these goals are worked on together using the same data and the same experiments.

The data generated from the discovery of diagnostic biomarkers are valuable information for basic science and research into disease pathology. Information about changes in the concentration of a biomolecule, modified forms of a compound, changes in protein expression, and posttranslational modifications, just to name a few, provides clues to the changes in the cellular machinery and pathways. This information can be used to determine where one should look for the changes that contribute to the disease pathology. Biomarkers will not provide all of the information, but they are a good tool for sorting out where to start looking.

The real goal here is to improve the information that a test can provide the physician in:

- Differentiating diseases that are currently difficult to separate or diagnose
- Detecting a disease condition earlier when treatments are more effective
- Understanding the pathology of a condition that is affecting a patient
- Assisting in choosing the best course of treatment
- Reducing the negative side effects of a treatment
- Monitoring the course of treatment to determine effectiveness

In summary, the proposed role of the biomarkers in medicine is to facilitate early diagnosis, the customization of treatment, and improved quality and quantity of life for the patient.

3. Important definitions

The terms used in this chapter must be understood in the context of their use. To simplify this we think that it is important that some of the terms that will be used a number of times are clearly defined.

Disease sample—a sample that is obtained from a patient who *has* the condition that one is looking for biomarkers. This should be confirmed by as complete a diagnosis as possible.

Control sample—a sample from a person who *does not have* the condition/disease that one is looking for biomarkers, but may have a related condition, another disease that has a similar diagnosis or pathology, or a healthy individual.

Matched samples—samples that are the same in as many parameters other than the disease as possible, factors such as age, sex, racial background, geographical location, etc. A perfect example of this would be samples from the same person before and after a successful treatment.

Biomarker discovery—an experiment designed to observe the maximum diversity of a particular class of molecule between disease and control samples to find differences resulting from a specific condition. This could include differences in quantity, structure, new molecules, modifications to common molecules, changes in structure, function, posttranslational modifications, etc. The number of samples in this type of experiment is only large enough to permit reliable statistical analysis to be done. (6–15 samples).

Primary validation—the second part of discovery where a larger number of samples are used to verify or eliminate potential biomarkers. This type of experiment is normally focused on the best (statistical) biomarkers observed in the discovery or identified in the discovery experiment and may use different methodology to facilitate a larger number of samples to be used to increase the statistical confidence and eliminate weak or poor biomarkers from the discovery experiment. This is also the stage of the experiment where one starts using samples from related conditions, other geographical areas, etc. The goal of the validation experiment is to focus on the potential biomarkers that are most likely to answer the research question, and to reduce the number of possible biomarkers from the discovery phase before further time, effort, and money are invested in purification, identification, and characterization of the molecule (30–50 samples).

Validation—the running of a large number of samples to look at the performance of the biomarkers in a population. This requires a sample set large and diverse enough to allow a statistical sampling of a population (hundreds to thousands of samples), in short a clinical trial.

4. Biomarkers discovery and complexity of biological systems

Biomarkers discovery is one of the most difficult tasks in biology partly due to the level of complexity of biological systems. To start with there are the problems of "normal biological variation" of 10–40%. This type of variation is impossible to reduce or eliminate from an experiment and thus must be accounted for in the statistical analysis. This variation arises from the differences in individuals resulting

from genetic background, environment, diet, age, sex, and an almost limitless set of variables. In well-controlled systems like animal models, a number of these variables can be controlled and, as a result, we see that the normal biological variation can be reduced and, to some degree, controlled. In a medical environment using humans for the study subjects this is not possible or practical. In fact for a number of commonly used medical tests there is a normal range for the amount of an analyte and values outside this range are considered a problem. Well-established tests such as cholesterol, blood glucose, triglycerides, etc., all have a normal range, not a single normal value. The same turns out to be true for other biomarkers as well. In the process of discovery, validation, and assay development, it will be necessary to use statistical analysis to determine the normal and disease concentration ranges for each biomarker.

The problem of biological variation can be minimized by good study design and careful statistical analysis of data. By carefully selecting the patients for the discovery phase of the experiment one can focus the search on biomarkers that are directly related to the question of interest. To this end it is advisable during the biomarker discovery experiments to select samples where the diagnosis is clear and the patient data are as uncomplicated as possible. Uncomplicated patient data indicate samples from patients with as few medical complications as possible other than those resulting from the disease state under investigation. For example, selecting a sample from a person with cancer as their only medical condition is preferable to a cancer patient with heart disease, high blood pressure, or diabetes for use in the discovery experiment. This type of sample selection will reduce the number of variables that will have to be analyzed, potential sources for biomarkers from other diseases, and the complexity of the data analysis. The initial discovery experiment can generally use a relatively small number of samples; a set of 6–15 diseased persons and an equal number of matched (age, sex, race, etc.) controls is a good number to work with provided that these samples are well chosen. The goal of the discovery experiment is to find as many potential markers as possible. The primary validation experiment should be larger: from 30 to 50 samples and matched controls. The goal of the primary validation experiment is to test the discovered markers and select the best of the discovery marker set to focus further efforts on.

The exact number of samples that you will need to work with in both the discovery and primary validation phases of the project depends on the techniques that are used and the complexity of the disease and the patient samples that you are working with.

The second issue the investigator will face is the immense number and diversity of biomolecules present in a biological system. These molecules can range from simple organic or inorganic compounds such as glucose and Na^+ to large complex biopolymers such as lipids, proteins, and carbohydrates. To further complicate the picture, biopolymers can be mixed with each other to form lipopolysaccharides, glycoproteins, lipoproteins, etc. From this complicated picture the investigator

looking for a biomarker must find the differences that highlight a particular condition from this complicated array of molecules. This might seem like an impossible task, but it is not. Successful biomarker discovery has aided the advancement of medicine over the last century.

5. Biomarker discovery and "omics"

In the discussion of biomarker discovery a number of "omics" terms have been used to divide the types of biomolecules into different fields: metabolomics, lipidomics, glycomics, proteomics, etc. This has some usefulness in defining what one is looking for, but unfortunately biology is not that simple. In many cases natural lines do not really exist between different types of biomolecules.

To highlight this point when we look in a textbook of biochemistry, we can easily find examples of lipopolysaccharides, glycoproteins, and lipoproteins. This complication is not simplified in the literature; as a practical example when searching the literature for glycomics one will quickly find cases of glycoproteins. With the ambiguities that we see in trying to classify lipidomics, glycomics, and proteomics, it is more important to think about the chemistry of the molecules that one is interested. The reason for this is that separation and identification tools are based on the physical and chemical properties of a molecule and thus understanding their chemistry is the key to working effectively with a class of compounds. One must keep in mind that there may (in rare cases) be some unexpected compounds in a sample preparation that should only contain one class of compound.

In this discussion of biomarker discovery, the arguments will be appropriate to all of the "omics" that have been coined. The fundamental differences between these different classes of biomolecules are the specific tools/techniques used for separation, detection, and identification. The principles governing design and development of biomarker discovery experiments are not affected by the type of compound or the tools used for the experiments.

6. The biomarker discovery project

In order to design a successful biomarker discovery project it is necessary to put together the right people and a solid experimental plan. A brief description of the biomarker discovery group and experiment are outlined in this section.

6.1. The biomarker discovery group

When putting together a group for biomarker discovery it is important to understand what type of expertise will be required to accomplish the goals of the project. This is

best done before the grant (budget) for the project has been applied for since the writing of the proposal will take care of most of the preliminary discussions and planning of the project and will allow time to refine the research question, ideas, and project before starting the work.

The anatomy of a successful biomarker discovery group generally contains the following skill sets:

- A biologist/biochemist who has research experience in the specific disease of interest
- A good understanding of separations, detection, and chemistry of the molecule type of interest
- A physician who has an active clinical practice and the ability to obtain the appropriate samples and an interest in research
- A mass spectroscopy expert with an understanding of the type of molecules that are to be studied and their identification and characterization
- A person with a good understanding of biological statistical analysis
- A reliable technician/student to do the sample preparation and experimental work

The research leader will need to put together as many people as necessary to obtain at least the skills listed above to run a successful project. It is also important that each member of the group is interested and committed to the project. When the group is assembled it is then time to develop a research question and an experimental plan to answer the question.

6.2. The biomarker discovery experiment

There are two main points we have to keep in mind before starting a biomarker discovery experiment.

6.2.1. Where do biomarkers come from?

There are several theories that try to explain the origin of biomarkers and why they are found. The simple truth of the matter is that biomarkers are produced by changes in biological processes of anabolism and catabolism, i.e., changes in metabolism. All biological molecules are used in metabolic processes; the exception to this is some xenobiotics but we will not discuss those here since they are not biomarkers. Changes in metabolism can result in the simple buildup or reduction in concentration of normal metabolites which can act as biomarkers. Furthermore, there is a possibility of modifications to biomolecules; using proteins as an example, markers are observed which are truncated, glutathionylated, and cystinilated, and with changes in the carbohydrate structure just to name a few. These types of modifications are caused by changes in enzymes and the metabolic pathways responsible for these posttranslational modifications. The results of these types of cellular changes are

generally the most common type of biomarkers observed since a small change in enzyme concentration, function, or pathway efficiency can produce large changes in the concentration of products or substrates. This produces an amplification of the metabolic changes in the organism. The results of these changes are usually easier to detect and quantify than the changes that produce them. Thus, the change in product concentration is observed and this becomes the surrogate biomarker for this metabolic change. In the majority of cases the actual change in the enzyme can only be detected/measured when it is looked for in a very specific way by using antibodies or a highly purified preparation or a detailed examination of enzyme activity/kinetics. Changes in metabolism can also cause the expression of proteins/enzymes not normally expressed in a tissue, the activation of enzymes that are normally inhibited, and the modification of enzymes to alter their activity; all these types of changes will produce products that can more easily be observed than the enzymatic change that produced them.

Example: Let us look at one example of this process currently used in medicine; the example that we will use is the control of diabetes in patients. For a number of reasons glucose and not insulin concentration is used to monitor this disease. The effect of a small amount of the hormone insulin can cause a big change in the concentration of glucose in the blood, which is easily monitored and provides an effective biomarker for monitoring control of this condition for both types of the diabetes (types I and II) (Chart 1).

Notes: This is probably not the best example, but it does illustrate another point about biomarkers and disease. Diabetes is a disease where the causative agents are proteins/peptides, insulin, and insulin receptor, but a good marker for the disease is the simple sugar glucose. This is not an isolated case. We can take examples from a number of other metabolic diseases where the cause of the disease is a protein/peptide and the resulting effect is most evident in the change in concentration of another class of molecule (lipid, carbohydrate, etc.). The pathology of the disease is a good tool to use as a starting point concerning the type of biomarker (protein, lipid, carbohydrate) that one could/should search for.

6.2.2. How many biomarkers are necessary?

This is a complicated question but one that must be thought about. The number of single biomarker assays that are failing in the market is high, and the FDA in the United States has not approved a single biomarker assay in the last year. The reason for this is that in most cases the single biomarker assay cannot show the level of sensitivity and specificity necessary to be an effective diagnostic.

The examples of this in the market place are several. CA-125 for ovarian cancer and CA15.3 for breast cancer cannot be used as screening tools only as a way of monitoring treatment [8–11]. The PSA test that is used as a screening tool shows sensitivity and specificity of 66%, and thus about one third of the cancers are missed and one third of patients go for unnecessary prostate biopsies. When

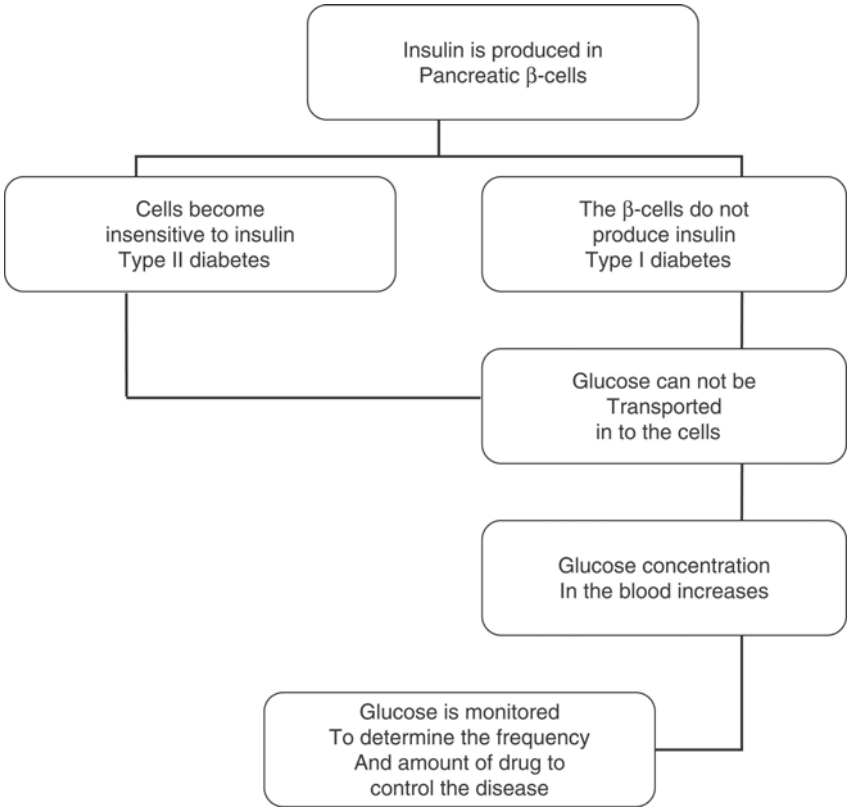


Chart 1. This is a schematic of the way the diagnostic for diabetes functions moving from the cause of the disease to the use of the biomarker glucose.

two or more of the current PSA tests (total PSA, bound PSA, and free PSA) are given together, the diagnostic performance improves significantly compared to any one of the single PSA tests alone [12–14]. The simple logic here is that the more information that is obtained the better the diagnosis. This also applies to biomarkers for the diagnosis of other conditions; the use of more than one marker can increase the diagnostic performance and reduce errors. By developing panels of markers (a combination of biomarkers) it is more likely that one will develop a high-quality clinical test. The logic behind this is that if one marker in a diagnostic panel fails in a patient, there are still other possibilities (other markers in the panel) that can be used to diagnose the disease. The use of patterns of biomarkers for diagnosis can reduce the problems for biomarkers assays resulting from disease heterogeneity, differences in disease pathology, and the effect of other medical conditions that occur in a population.

Disease pathology is probably a heterogeneous process; with some variation aspects of disease progression differently in different individuals, symptomology of the disease can also vary from case to case. This suggests that the pathology of

a disease is not necessarily the same from individual to individual and thus it is probable that there are slightly different biochemical processes going on as the disease develops. Therefore, ideally the markers should come from different pathways that have been altered by the disease process. This type of diagnostics is less prone to errors in diagnosis that are the result of different rates or types of disease progression or subclasses of the same disease. A lot of promise is being seen from the use of multimarker panels of biomarkers for the diagnosis of cancer [15]. Researchers in this field have found it necessary to use three or more proteins to obtain the desired diagnostic efficiency to make a useful test. The actual number of markers required to develop an effective diagnostic may be only three or four, but the process of primary validation will eliminate a large percentage of prospective markers that are found in the discovery experiment. Thus, it is important to find as many prospective biomarkers as possible so that there is the opportunity to evaluate and choose the very best group of markers to develop further.

7. Challenges in the biomarker discovery pathway

The process for a biomarker discovery process is very similar regardless of the type of molecule that one is looking for. Here, detailed descriptions of the challenges in the biomarker discovery pathway, which have to be kept in mind, are outlined. This section will also provide some information about how these challenges can be met by the careful researcher.

7.1. Asking the right question

It is important to think about the question that you are asking in several ways because the formation of the right question is the key to a successful biomarker discovery experiment. A good question has several qualities that help to direct the biomarker discovery work and focus the research efforts in a clear and organized manner. Is the question reasonable with respect to what is known about the disease and the disease state? For example, a poor question is: I want to find a diagnostic for cancer. This is a poor question for several reasons:

- Cancer is a very heterogeneous disease.
- There are a number of different cancer pathologies and tumor types.
- Each cancer has several stages that are different.
- Different forms of cancer affect different tissues.
- The number of samples required to represent all forms of cancer would be unmanageable.

A broad-ranging question may seem like a logical starting point, but one will quickly find that once the experimental work is started the resulting data will be complicated and virtually impossible to extract useful information from and thus

the project will be impossible to complete. However, by reducing the scope of the question to some aspect of the disease, it is possible to focus the research better and obtain useful results. Reducing the scope of the discovery experiment by design will allow all of the efforts to be focused on an achievable goal that can be accomplished in a reasonable time frame.

Some much better questions in the same field could be:

- Can I find markers that allow the differentiation between an aggressive and a nonaggressive form of the same cancer?
- Can I find markers that will predict which drugs may be most effective in treating this form of the disease?
- Can I find markers that facilitate determining the effectiveness of a course of treatment?

Answering these smaller questions will start on the road to answering much bigger questions. But more importantly there is a lot of evidence that these types of experiments are more often successful in both discovering useful biomarkers and publishing the results. By posing a question with a narrow scope one will focus the research efforts from sample collection to data analysis facilitating a more complete study with smaller sample numbers, less complicated data analysis, and a more significant outcome; thus developing clinical utility quicker. The use of narrowly defined questions will also provide time for the biomarker research group to develop the experience, techniques, and research tools that will be needed for projects of a larger scope.

Furthermore, the use of narrowly defined questions will also facilitate the building of a knowledge base that can then be used as a starting point for broader studies that can be accomplished by combining datasets and reanalyzing the data to examine possibilities and determine feasibility of answering or posing a broader question. Let us think for a moment as to how this could be accomplished. If a research group decides to look for biomarkers from different cancers of epithelia origin (for example) and does a number of very specific discovery experiments on colon, skin, and breast cancer and develops biomarkers for these diseases, one has a very successful biomarker discovery and validation program. The data obtained during this work can be combined and reanalyzed to look for markers that are common between these diseases and markers that differentiate these conditions and this could be the starting point for and answer to a bigger question involving several forms of cancer. This type of exploration will certainly elucidate the scope of the work and the possibility of success, and give indications concerning the number of samples required to answer the new bigger question.

In summary, the difference between good and bad questions is the scope of the study and the chance of success in the project. A good question has the following characteristics:

- The goal for the study is clearly defined.
- The question takes into account what is known about the disease.
- The study is possible with the samples available.

- The study could provide the answer to an important clinical question.
- The scope of the study is narrow enough to find clear answers.
- The collection of further samples for primary validation is possible.

There is a good review article covering this topic and it would be advisable to read this publication during the process of formulating your biomarker discovery experiment [16].

At this point it is worth discussing the entire biomarker discovery and validation process from one aspect. Each step in this process is built on the results of the previous steps. This means that errors and poor decisions will compound and amplify as one progresses along the experimental plan. For example, a poor experimental question will lead to a very large discovery experiment. This scope of the discovery experiment will require the use of a less than ideal set of discovery samples. This discovery experiment will produce a large volume of data that must be analyzed; this takes time and usually results in a large and confusing list of possible biomarkers since the discovery sample set was less than optimal. These prospective biomarkers will require a very large primary validation set of samples and a broad experimental scope for the primary validation process generating another very large dataset, which may only serve to complicate the original discovery experiment. These types of compounding problems will continue as one progresses to the purification, identification, and assay development with too many possibilities and too much work to accomplish. This will result in a failed biomarker discovery process and in not achieving the goal of the project due to the volume of work that will be required as well as the time involved and the experimental cost.

In contrast, a well-designed biomarker discovery process will result in increasing experimental focus because of a reduction of the number of prospective biomarkers as the number of samples increase. It is true that during the discovery process it is important to find as many biomarkers as possible; the process of primary validation should be used to reduce the number of markers that are being considered to answer the original question by selecting the best markers from the discovery pool. Thus, by generating data from more samples there are fewer markers that are being analyzed and thus the experimental work remains focused and the amount of data to be analyzed remains reasonable. This type of experiment will allow diagnostic models to be developed using a number of biomarkers and the data generated will facilitate the testing and refining of these models until a diagnostic group of biomarkers emerges. At this point, it is worth the work to identify, characterize, and develop assays and validate this small group of diagnostic biomarkers since there is a good chance that they will answer the initial question.

7.2. Sample availability

The definition of a good question also relies on the samples available, which is an important component to the formulation of the best question. Sample availability

is of primary concern; although discovery can be accomplished with relatively few samples, the process of primary validation and validation will require a large number of samples. The exact number of samples required for a discovery experiment will depend on the method used to find the marker. The statistics and errors that are inherent to the method will allow the researcher to calculate the number of samples required to produce a significant result that will facilitate the selection of prospective biomarkers. This process is also true for the validation phase of the experiment. The statistics and the question will determine the number of samples required to validate prospective biomarkers.

The availability of samples from different locations, such as cities, countries, and continents, is important. Samples from diverse locations are necessary to sort out markers that might be found in a particular group of people due to environmental, dietary, genetic, or race differences and are therefore not related to the condition under investigation. It is not vital to have samples from multiple centers to do the discovery work, but it is a good idea to think about obtaining these types of samples for primary validation and validation work.

7.3. Which sample should be used?

An important characteristic of the samples whether they are a biological fluid (sera, plasma, urine, etc.) or a tissue (muscle, liver, skin, etc.) is that they contain the type of molecules that one is looking for in concentrations that are above the detection limits of the techniques available for examining the molecules of interest. This must be determined by a series of test experiments to determine the concentration of the analyte in the sample. At this stage it is not important to determine the concentration of each analyte in the sample because there is enough of the class of compound (protein, lipid, carbohydrate, etc.) to make a study productive. This work is a simple optimization experiment to study the possible methods of sample preparation looking at extraction, fraction, separation, and detection of the class of molecules (lipids, metabolites, proteins, etc.) that one intends to examine, and then to put together the most promising methods to form the experimental protocol. During this stage of experiment while developing the separation methodologies that will be used for discovery, one should take note of the level of diversity that is seen in the samples. The reason for this is that the number of species that can be observed will provide information on the chances of finding biomarkers. The discovery of biomarkers is a statistical process. The more different species are observed the more likely that one will find changes in one or more of the species and therefore the better the possibility of finding a biomarker in the sample. For this work, one generally chooses a sample that is easy to obtain and is very similar to the samples of interest. For example, if one is planning to work with human liver samples from biopsy it would be reasonable to optimize conditions with bovine liver samples and when methodologies have been worked out then move to the human samples. This will

optimize the use of valuable clinical samples and provide information on the size of the sample necessary and how the sample needs to be obtained, stored, subdivided, and processed.

7.4. Sample management

Two rules have been kept in mind when thinking about samples for biomarker discovery and primary validation. These rules are designed to reduce the complexity of discovery experiments by reducing the number of artifact and false markers that are the result of poor sample collection and handling processes, which often contribute to problems in statistical data analysis. These artifact markers can also mask real biomarkers and cause them to be ignored or missed in the early stages of the project.

Rule 1: All samples have been treated exactly the same. This means that a protocol has been developed for the entire process of sample collection, storage, and use. The defining of the sample collection protocol must be done with three main points in mind:

- What is reasonable in the collection environment, hospital, clinic, etc.?
- What can be done reproducibly?
- The collection process needs to be documented at each step.

In the development of the sample collection protocol, one must keep in mind what is reasonable in the real world. If blood samples are being collected by a nurse in a hospital (for example) it would be unreasonable to think that the samples could be aliquoted and frozen 20 min after collection.

Example: Sera proteomic study. Here is an example of a sample collection protocol that is being used for sera proteomics study in a hospital.

- Samples are collected three times in a 24 h period.
- After drawing, the blood samples (in vacutainer tubes for sera tiger-top) must be centrifuged a minimum of 2 h and a maximum of 2.5 h after collection with the sera remaining at room temperature between collection and centrifugation.
- After centrifugation, the samples must be aliquoted into ten 30 μl and four 200 μl fractions and frozen at −80°C within 30 min. Placing the samples in dry ice after aliquoting is acceptable.
- Each aliquot is to be bar-coded and the collection data are to be recorded in the patient database with cross-referencing to the clinical data.
- Five of the 30 μl and two of the 200 μl fractions are to be placed in freezer 7.
- Five of the 30 μl and two of the 200 μl fractions are to be placed in freezer 5.
- If the sample does not conform to this procedure, discrepancies must be noted in the database and the sample must be placed in storage with other nonconforming samples.
- Samples are only to be subjected to one freezing cycle. Any problems with the freezing or thawing and refreezing of a sample will require it to be reclassified as nonconforming.

Notes: Samples that do not fit this protocol are not to be used for discovery or primary validation experiments. The reason for this is that the treatment of samples can introduce artifacts into the sample: processes continue in biological samples after the sample has been removed from the body. In fact, biological processes are not even completely stopped by freezing to −20°C; thus, samples should be stored at −80°C especially for long-term storage. Samples should be divided into single-use portions and stored in different freezers for safety. A lot of effort and cost are associated with the collection and storage of samples and hence they need to be protected. Furthermore, if the integrity of the samples has been compromised, then all of the data resulting from these samples will also be compromised and the effort, money, and work that went into the discovery experiment could be wasted along with the sample. Compromised samples could have a use as bulk tissue for larger scale experiments of purification, characterization, and identification of prospective biomarkers.

Rule 2: Time is an enemy. The longer the samples remain at less than optimal storage conditions the more artifacts will be produced in the sample. A large number of biomolecules are unstable chemical compounds to start with. They have limited lifetimes in biological samples for several reasons. The first is chemically unstable biomolecules that are susceptible to oxidation, reduction, and hydrolysis, just to name a few of the many chemical reactions. The second issue is that there are enzymes present in biological samples that can accelerate these processes by many orders of magnitude by catalyzing these reactions. Thus to obtain a good picture of a biological system it is important to minimize these processes as soon as possible after the sample is obtained.

This is a problem for both the experimental process and data analysis. The primary problem is the degradation of the molecules of interest because at elevated temperatures (generally greater than −80°C) there are both enzymatic and chemical reactions that can occur in the sample resulting in the breakdown and/or modification of biomolecules. This can be compensated for in assays by using time courses to determine the rate of destruction of a known compound over time in a specific sample type. This is not the most desirable system to adopt since it is not perfect and introduces an increase in the level of error in the results. On the contrary, this is not a possibility when one is looking for biomarkers in discovery experiments, since time courses involve knowing what the target of interest is and during a biomarker discovery experiment by definition the target is not known.

The second issue is that some important biomarkers are present in low concentrations and are near the limit of detection; the loss of even a small amount of these molecules may cause them not to be observed in the sample and thus the information about them will be lost.

Notes: The use of confounding or clinically similar conditions to test biomarkers in the primary validation and validation process is vital to determining which prospective biomarkers are directly related to the disease or whether they are markers of other factors (age, sex, diet, etc.). With many disease states, there are

conditions that can make the diagnosis problematic or produce similar biomarkers to the disease under investigation. A good example of this can be seen in the diagnosis of prostate cancer where benign hyperplasia, age of the patient, and prostate cancer all cause an increase in the levels of prostate-specific antigen which is used as a biomarker (and a screening tool) for this disease; thus, it is important to include benign hyperplasia samples in the validation samples for any prospective prostate cancer biomarkers. The result of this type of experimental design is to find markers that can detect prostate cancer and distinguish it from other medical conditions that have different treatments or are not in fact disease conditions.

8. Technical issues in experimental design

The sample preparation methods and the separation techniques have to be developed carefully in order to facilitate the entire biomarker discovery process. Sample preparation methods should be compatible with the subsequent steps in the sample preparation process. The use of a sample preparation method that results in the sample being in a solution that is incompatible with the separation method to be used will cause problems such as increased experimental difficulty, sample loss, increased time, and experimental cost. A good sample preparation method will facilitate quality results and the collection of good data. Good-quality results are easier to analyze and less prone to experimental errors that require high numbers of replicates to control.

8.1. Sample preparation

It is important to make sure that the methods used for sample preparation are reproducible and robust. It is not important at this stage to develop a method that will be used during validation and any subsequent clinical method. The goal of discovery work is to see as much of the sample diversity as possible. Remember that biomarker discovery is a statistical game where the more choices that you have the better the chance that you will find a marker that can be validated, so it is important to see as much of the diversity in the chosen class of molecule as possible. One should try and find a method that will eliminate classes of biomolecules that are not of interest, will interfere with the analysis of molecules that are of interest, complicate the separation/detection of the species of interest, and could cause confusion during data analysis. Time and effort spent in the process of sample preparation will make life much easier in subsequent experimental processes.

The preparation of samples should focus on the type of molecule that one is interested in. Methods have been developed for the isolation of classes and subclasses of biomolecules; these methods can be found in the literature and it is

worth your time to look for a sample preparation method that is as specific as possible for the type of molecules that you are interested in. It is much better to focus your efforts on one class or subclass of molecule at a time, rather than to try to see everything from the same samples. If you are looking for biomarkers that are phospholipids, then select a sample preparation method that is specific for isolating phospholipids rather than complicate the situation by looking in a total lipid fraction for phospholipid biomarkers. The saving of time and effort in the early stages of sample preparation will result in a much more complicated set of problems during experimental design and especially during data analysis.

It is also worth some time to talk about the level of abundance of molecules. In general, the more abundant a molecule is in a system the less scientifically interesting the molecule tends to be. The reason for this is manifold, but one reason is that the tools for the analysis of compounds are all the same in one respect; it is easier to see the most abundant species and thus they have been studied the most and a lot of information is available. Thus, the rarer a species the more scientific interest it generates since less is known about the molecule. In the case of biomarkers for medical applications, it is the ability to detect and quantify a marker that should be of primary concern. This makes more common molecules of more interest since they are easier to detect and quantify than compounds that are extremely rare. Furthermore, one can use a smaller sample size and simpler sample preparation procedures; this becomes important when a large number of samples must be run for validation of a marker.

It is also advisable to prepare tissue samples in a logical way based on the biochemistry of the molecules of interest and use tools such as subcellular fractionation to assist in the separation of your molecules of interest. For example, if you are looking for changes that are occurring in the mitochondria, it is advisable to isolate these organelles and work with a pure mitochondrial sample rather than a sample contaminated with cytosol and other organelles since the simpler the sample the better your chances of seeing a wider spread of the diversity in the sample.

The simpler the sample the more you will be able to see in the sample; at first it might seem like a bit of an oxymoron but the truth in this statement becomes evident when we think about this issue from the other direction. If you have a pure compound as a sample, then you will be able to see 100% of the diversity in the sample during analysis. If the complexity of the mixture is increased to 100 compounds and one sees only 75 different species using an analysis method, then 25% of the diversity of the mixture is lost regardless of the reason (compounds of the same mass, overlapping peaks in chromatography, overlapping spots or bands on a gel, etc). This is a simple case compared to the levels of complexity of molecules in a biological sample where the presence of hundred thousands of molecules would not be considered unusual. Thus, it makes good sense to start with as pure a sample as possible and thus increases the possibility of seeing the maximum level of diversity in a class of molecules as possible.

8.2. Separation technique

Separation technology is sometimes the biggest problem in biomarker discovery experimentation. The reason for this is the complexity of the mixtures found in biological samples and the poor resolution of the separation tools that is available. If we use chromatography, for example, it is possible to see 150 peaks from an HPLC separation of a biological sample such as sera, which contains more than 100,000 different molecules. This represents less than 0.15% of the available diversity. To further aggravate the situation there is sample loss on all columns of 30–50% of the sample. This is the material that is lost and cannot be analyzed. Many years ago there was a general rule for enzyme purification, which stated that: "if you needed more that three columns to purify an enzyme then you would end up purifying the enzyme activity away from the protein" which means that the protein could not be seen on a SDS-PAGE, but you could still measure the activity and thus you needed to rethink the purification process. This rule is also applicable to biomarker discovery; one simply cannot add unlimited dimensions of separation to increase the level of observed diversity because after four different columns there will not be enough material eluted from the last column to be of practical use because one will only observe the most abundant molecules and very little of the diversity in the sample. This problem can partly be overcome by using larger samples but this quickly becomes limiting when dealing with clinical material (it is not possible to ask a patient for 50, 100, or 1000 ml of sera).

Multiple separation methods might be necessary to adequately simplify the biological sample to facilitate the identification of hundreds or thousands of components in a mixture. It is not unreasonable to think about two or three different separation methods used in sequence or in parallel to look at the diversity in a biological sample.

It is important to take the time to discuss the known problems that are associated with a particular sample type and find methods that are likely to reduce or eliminate these situations, and then test the system you plan to use to optimize the process and test the protocols at each stage of method development. This process will provide information on reproducibility at each step in the discovery process and could allow the prediction of the number of replicates that are necessary, the number of samples that are required, the amount of sample required, the type of data analysis that needs to be done, and the amount of data that will be generated.

8.3. Bioinformatics

The need for bioinformatics in the process of biomarker discovery becomes apparent when one starts to observe the amount of data that is produced by discovery and validation experiments. Bioinformatics really covers two areas of this work. The first is the management of the raw data output from the instrumentation. It is preferable

to use databases to manage this raw data so that it can be sorted, searched, and cross-referenced with other information about the sample. It is a wise idea to start this work early in the experiment process before the amount of information makes this an almost impossible task.

The second area of bioinformatics is the analysis of the data. This is where it is important to involve an expert in bioinformatics. Even if there are software packages available for the analysis of your data, it is important that these tools are used properly and that you understand the manipulations and output of the software. The choice of the type of analysis that one should use is also something best left to experts to decide since there are no right answers to the question of how data should be analyzed and what is the best method of analysis; there are only answers to what is your preferred method and whether a type of analysis is valid. Some of the most frequently used analysis tools are [17–21]:

- Principal component analysis
- Hierarchal clustering
- *p*-Values
- *t*-Test
- Regression analysis
- Decision tree analysis
- CART analysis

We will not elaborate more on this topic since it is complicated and it is best to interact with an expert in this area to find out what, how, and why certain types of analysis are best for the type of data that you will be producing.

8.4. Methods

The development of a biomarker discovery method is a process of putting together methods and processes that facilitate finding changes in a biological system that indicate specific changes in that system. This generally involves the separation, detection, quantitation, and identification of molecules that are different between the condition under study and a set of control samples. Each step in this process affects each subsequent step in the experimental pathway. With the high degree of complexity in biological samples, one must take the time to develop each step of the experimental pathway carefully. This involves looking at sample collection, sample preparation, sample separation, analyte detection, and identification as described in the previous section.

The development of each step helps in defining the requirements of the subsequent methods and the options that are possible to use. A mistake or a poor choice in the first steps can cause insurmountable problems at the end of the experiment. Thus, it is important to take the time and develop methods that will reduce as many foreseeable complications as possible because there are enough unpredictable complications that will occur during this process.

9. The discovery process overview

The steps in biomarker discovery are always the same. A summary of the process is as follows:

- Biomarker discovery
- Primary validation of the markers from the discovery experiment
- Purification of validated biomarkers
- Identification of validated biomarkers
- Characterization of validated biomarkers
- Assay development for biomarkers with good diagnostic potential
- Validation use of the biomarker assay that has been developed to test large numbers of samples

This process is simple and logical. Since the time, effort, and money required increase with each step in the process, it is important to focus these efforts on the biomarkers that have the best chance of success in completing the process. The analysis of the data at each step in this process will reduce the number of biomarkers that will be used in the final assay and allow the continuous refining of the diagnostic model, thus facilitating the development of the best possible diagnostic.

9.1. Tools for biomarker discovery

The improvement of tools for looking at biomolecules has accelerated over the past few years as technology has advanced. One of the most recent advancements in the detection and identification of biomarkers is the mass spectrometer. From the prospective of biomarker discovery, this tool has been the most interesting to date. The mass spectrometer in all of the configurations discussed in this book has played an increasingly important role in the detection and identification of biomolecules. Regardless of the type of mass spectrometer that one uses, they have been adapted, designed, and optimized for a wide range of biomolecules.

The only point that we hope to make in this chapter about mass spectrometry instrumentation is that this technology is rapidly becoming the method of choice for the detection, characterization, and identification of biomolecules.

From the available mass spectrometry instrument configurations, there are one or more instruments available to look at the type of molecule that you are interested in. The tools for biomarker discovery can seem complicated and diverse, but they accomplish one of these three things: separation, detection, and identification.

The discovery of biomarkers will require each of these three steps to be done in order to discover a biomarker from a biological sample. Biological samples are highly complex mixtures of molecules many of which will interfere with each other, and thus sample preparation methods are required to first isolate the class of molecule that one is interested in. The samples that have been optimized for a particular class of biomolecule (optimized samples) can then be more efficiently

separated and analyzed. It is much easier to work with a mixture of lipids from a tissue than it is to work with the entire tissue homogenate.

We will not discuss in detail each of these technologies in this section. The lists provided are not complete; since very specific separation, detection, and identification tools exist; we will attempt to briefly mention the most general and often-used tools.

The development of mass spectrometry technology [22,23] including instruments, informatics, and methodologies has made this instrumentation the preferred tool for the detection and identification of biomolecules. The true power of mass spectrometry in the field of biomarker discovery is the large amount of information that this technique provides as well as the speed at which answers can be generated. An example of this is in the sequencing and identification of proteins; Edman chemistry requires 90 min to sequence a single amino acid, whereas an entire peptide can be sequenced with a mass spectrometer in a few minutes. For the detection of eluents from chromatographic separations, mass spectrometry can provide detection as well as identification with the additional benefit of separating each peak further by mass to determine if one or more components are eluting at once (Table 1).

In many cases a number of technologies are used together to produce biomarker data during the discovery process. Techniques are combined to overcome the weaknesses and exploit the strengths of each technique. The most important methods in the process of biomarker discovery are the design of the experiment and the analysis of the data; this is especially true in the medical field. As we have seen in the discussion of the tool of biomarker discovery, it is simple to generate large quantities of data that become difficult or impossible to analyze. Thus, it is vitally important that the discovery experiments are designed to focus all aspects of the experiment on the initial clinical question. One must also design an experiment to produce quantities of data that one is able to analyze with the tools available. There is little value in producing thousands of mass spectra without the informatics tools available to facilitate the analysis of the data. This type of experiment consumes valuable sample and does not produce results that can be used.

9.2. Biomarker identification and validation

Identification of prospective biomarkers after primary validation is an important step in the process of doing a good discovery experiment. There are some researchers who believe that patterns of biomarkers are adequate and there is no need to know what these markers are to diagnose a condition. This theory is losing favor as problems with this theory increase. First, these types of biomarker pattern diagnostics are failing to be accepted by the literature and regulatory agencies. Second, without knowledge of the identity of the markers there are no options in assay development or the ability to validate the biomarkers by a second methodology. Third, the identification of biomarkers or patterns of biomarkers provides significant advantages for the researcher in the validation and assay development aspects of prospective biomarkers.

Table 1
Most common techniques available for biomarker discovery, characterization, separation, and identification

Method	Strengths	Weaknesses
Separation tools		
SDS-PAGE	Good for proteins, ease of use, good analysis tools, rapid, very robust, sample can be recovered	Low resolution, cannot be used for small molecules
2D gel	High resolution, well-established method, good analysis tools, sample can be recovered	Time-consuming, large sample size, experience required, problems with very basic/acidic proteins, not good for small molecules/peptides
Capillary electrophoresis	High resolution, small sample size, rapid, not affected by size of the analyte, reproducible	Few methods, difficult to recover sample for further analysis
Liquid chromatography	High resolution, possible to combine column types to produce 2D separations, not affected by size of the analyte, wide range of separation methods available, very robust, reproducible, sample can be recovered, can be coupled to a wide variety of detector types	Sample loss, large sample size, separations can be time-consuming
Gas chromatography	Very high resolution, can be coupled to mass spectrometry for detection and identification, robust, good for small volatile molecules, reproducible, very small sample size, rapid	Only good for volatile compounds, may require derivatization, generally the sample cannot be recovered
ProteinChip arrays	Rapid, reproducible, 2D separations standard, wide range of separation types available, robust, small sample size	Designed for protein/peptides, limited capacity
Detection tools		
Staining	Very sensitive, large number of stains available, robust, simple, inexpensive, permits sample recovery, very widely used	Mostly used for SDS and 2D gels, quantitation can be problematic, minimal information about the sample

Table 1
Continued

Method	Strengths	Weaknesses
Absorbance	Very sensitive, robust, reproducible, medium cost, permits sample recovery, widely used, can be coupled to a number of separation techniques	Provides minimal information about the analyte
Mass spectrometry	Sensitive, produces further information about the analyte, adds a dimension of separation (by mass), reproducible, robust, sample identification possible	Sample recovery not possible, less sensitive than stains or absorbance
Identification tools		
Edman chemistry (N-terminal sequencing)	The gold standard for protein/peptide identification, moderate cost, widely used, can identify modified amino acids	Time-consuming, experience required, will only identify proteins/peptides, large sample size, requires purified proteins, sample consumed
Mass spectrometry	Can identify a wide range of molecules, can work with mixtures of samples	High cost, expert required
Biological interaction (antibodies)	Very sensitive, can be analytical, reproducible, widely used, low cost	Must have an antibody to the molecule, cross-reactivity, will not identify modifications to a protein, must know exactly what you are looking for
IR spectroscopy	Sensitive, provides structural information, good for small molecules, identifies organic functionality, use of the fingerprint region can provide compound ID for small molecules, moderate cost	Not useful for protein ID
NMR	Provides structural information, identification, good for ID of small organic molecules, sample recovery possible, can find and identify modifications	Very expensive, expertise required, works best for small molecules, not generally useful for proteins, generally requires purified samples, large sample size

Highlighted in this list are some of the strengths and weaknesses of each method.

The identification of a prospective biomarker can provide a wealth of information. The following is a list of a few of the advantages of identifying the biomarkers that are discovered:

- Provides a way to look into the literature to provide evidence that this marker has been linked to some aspect of disease pathology
- Facilitates the use of other tools to validate a biomarker by an independent method(s)
- Points to an alteration in a biochemical pathway that will facilitate further discovery experiments and basic research
- Provides other research tools/methods to characterize the prospective biomarker
- Increases confidence in the marker being a result of the disease and not an artifact
- Provides further understanding of the disease process/pathology

10. Conclusions

The important messages to remember from this chapter are to think carefully about the discovery experiment first. Put together a solid team with the skills necessary to accomplish the project. Think carefully about the question that is being asked: Is it reasonable and achievable? Plan the project well and listen to potential problems and challenges so that the question can be answered satisfactorily. Focus on the question being asked and do not get sidetracked until there is an answer to the initial question; there is always an opportunity to investigate the data further after the question has been answered. Simplify each step in the biomarker discovery experiment as much as possible so that the data analysis results are as clear as possible.

The order of work for biomarker discovery is as follows:

- Think carefully about the question that you would like to ask.
- Check to make sure that you have the samples required by the question that you are posing.
- Do you have enough samples for validation? Can more samples be obtained?
- Is there enough (quantity) of sample to facilitate purification and identification?
- What method will be selected for separation and detection of the molecules of interest?
- Do the optimization experiments so that some sample data can be produced and protocols worked out.
- Work with a biostatistician to determine the number of samples that could be required for discovery, primary validation, and validation of the biomarkers with your proposed methodology.
- Perform the discovery experiment.

- Carry out primary validation experiment.
- Identify and characterize the potential biomarkers.
- Develop a robust assay for the biomarkers.
- Validate the biomarkers.

References

1. Srivastava, S. (Ed.), *Cancer Biomarkers* (Book Edition of *Disease Markers*), IOS Press, Amsterdam, The Netherlands, 2004, 94 pp.
2. Pusch, W., Flocco, M.T., Leung, S.M., Thiele, H. and Kostrzewa, M., Mass spectrometry-based clinical proteomics (review). *Pharmacogenomics*, **4**(**4**), 463–476 (2003).
3. Chang, P.L., Clinical bioinformatics (review). *Chang Gung Med. J.*, **28**(**4**), 201–211 (2005).
4. Seibert, V., Ebert, M.P. and Buschmann, T., Advances in clinical cancer proteomics: SELDI-ToF-mass spectrometry and biomarker discovery (review). *Brief Funct. Genomic Proteomic*, **4**(**1**), 16–26 (2005).
5. Righetti, P.G., Castagna, A., Antonucci, F., Piubelli, C., Cecconi, D., Campostrini, N., Rustichelli, C., Antonioli, P., Zanusso, G., Monaco, S., Lomas, L. and Boschetti, E., Proteome analysis in the clinical chemistry laboratory: myth or reality? *Clin. Chim. Acta*, **357**(**2**), 123–139 (2005).
6. Alaiya, A., Al-Mohanna, M. and Linder, S., Clinical cancer proteomics: promises and pitfalls. *J. Proteome Res.*, **4**(**4**), 1213–1222 (2005).
7. Weiss, J.N., Qu, Z. and Garfinkel, A., Understanding biological complexity: lessons from the past (review). *FASEB J.*, **17**(**1**), 1–6 (2003).
8. Jotti, G.S. and Bombardieri, E., Circulating tumor markers in breast cancer (review). *Anticancer Res.*, **10**(**1**), 253–258 (1990).
9. Herbst, A.L., The epidemiology of ovarian carcinoma and the current status of tumor markers to detect disease (review). *Am. J. Obstet. Gynecol.*, **170**(**4**), 1099–1105 (1994) (discussion 1105-7).
10. Bischof, P., What do we know about the origin of CA 125? (review). *Eur. J. Obstet. Gynecol. Reprod. Biol.*, **49**(**1**–**2**), 93–98 (1993).
11. Sevinc, A., Camci, C., Turk, H.M. and Buyukberber, S., How to interpret serum CA 125 levels in patients with serosal involvement? A clinical dilemma (review). *Oncology.*, **65**(**1**), 1–6 (2003).
12. Albertsen, P.C., Prostate-specific antigen: how to advise patients as the screening debate continues (review). *Cleve. Clin. J. Med.*, **72**(**6**), 521–527 (2005).
13. Berlin, B., Prostate cancer: is the PSA test the answer? (review). *N. J. Med.*, **95**(**3**), 53–55 (1998).
14. Pfister, C. and Basuyau, J.P., Current usefulness of free/total PSA ratio in the diagnosis of prostate cancer at an early stage. *World J. Urol.*, **23**(**4**), 236–242 (2005).
15. Mitas, M., Almeida, J.S., Mikhitarian, K., Gillanders, W.E., Lewin, D.N., Spyropoulos, D.D., Hoover, L., Graham, A., Glenn, T., King, P., Cole, D.J., Hawes, R., Reed, C.E. and Hoffman, B.J., Accurate discrimination of Barrett's esophagus and esophageal adenocarcinoma using a quantitative three-tiered algorithm and multimarker real-time reverse transcription-PCR. *Clin. Cancer Res.*, **11**(**6**), 2205–2214 (2005).
16. Eric, T., Fung strategies in clinical proteomics: from discovery to assay. *Preclinica*, **4**, 253–258 (2004).
17. Coste, J., Bouee, S., Ecosse, E., Leplege, A. and Pouchot, J., Methodological issues in determining the dimensionality of composite health measures using principal component analysis: case illustration and suggestions for practice (review). *Qual. Life Res.*, **14**(**3**), 641–654 (2005).
18. McLachlan, G.J., Cluster analysis and related techniques in medical research (review). *Stat. Methods Med. Res.*, **1**(**1**), 27–48 (1992).

19. Livingston, E.H. and Cassidy, L., Statistical power and estimation of the number of required subjects for a study based on the t-test: a surgeon's primer. *J. Surg. Res.*, **126**(**2**), 149–159 (2005).
20. Ludwig, D.A., Use and misuse of p-values in designed and observational studies: guide for researchers and reviewers. *Aviat. Space Environ. Med.*, **76**(**7**), 675–680 (2005).
21. Huland, H., Graefen, M., Haese, A., Hammerer, P.G., Palisaar, J., Pichlmeier, U., Henke, R.P., Erbersdobler, A., Huland, E. and Lilja, H., Prediction of tumor heterogeneity in localized prostate cancer (review). *Urol. Clin. N. Am.*, **29**(**1**), 213–222 (2002).
22. Merchant, M. and Weinberger, S.R., Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry (review). *Electrophoresis*, **21**(**6**), 1164–1177 (2000).
23. Chaurand, P., Schwartz, S.A., Reyzer, M.L. and Caprioli, R.M., Imaging mass spectrometry: principles and potentials (review). *Toxicol. Pathol.*, **33**(**1**), 92–101 (2005).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 23

Molecular imaging by mass spectrometry

SARAH A. SCHWARTZ[a] and RICHARD M. CAPRIOLI[b,*]

[a]*Midwest Research Institute, Kansas City, MO 64110-2241, USA*
[b]*Department of Biochemistry and the Mass Spectrometry Research Center, Vanderbilt University School of Medicine, Nashville, TN 37232-8575, USA*

1. Introduction

The potential windfall of information for molecular-based clinical diagnostics from genomic and proteomic studies involves discovery of disease-specific biomarkers. With respect to proteomics, mass spectrometry (MS) improvements in instrumentation and ionization techniques have dramatically impacted protein biomarker identification. Specifically, the introduction of matrix-assisted laser desorption/ionization (MALDI) [1,2] and instrumental improvements in time-of-flight (TOF) mass spectrometers [3,4] have greatly enhanced signal resolution and mass accuracy for intact high mass molecules. Using this technology, protein detection at femtomole to attomole levels and low parts-per-million mass accuracies have been achieved [5]. The development of tandem MALDI TOF mass spectrometers also

*Corresponding author. Tel.: +1-615-322-4336; Fax: +1-615-343-8372.
E-mail: r.caprioli@vanderbilt.edu (R.M. Caprioli).

allows for rapid and accurate peptide identification from >96 samples on a single instrument target plate. As a result, MALDI TOF technology has become a major tool for peptide and protein detection, identification, and characterization [6,7].

Over the past 10 years, methods have been optimized for the direct analysis of individual cells, groups of cells, and small tissue sections [8–27]. Peptide profiling of complex mixtures by MALDI TOF MS has been performed without previous molecular separation to show differences between cell types and physiological conditions, identify novel peptides, assess post-translational processing, and demonstrate peptide localization within the tissues. Additional experiments have demonstrated sub-cellular protein localization within intact cells [23,28]. Efforts have also been made to characterize bacterial strains based on intact peptide and protein profiles [12,15,18]. Most of this early work involved analysis of peptides and low-molecular-weight proteins, typically from established cell lines.

Several MALDI TOF MS peptide and protein profiling studies have focused on the direct analysis of tissue sections [9,10,20,22,24–26,29–32]. This work aimed to retain protein spatial information with validation by histology. MALDI TOF MS technology has been used to characterize human disease based on the protein patterns from biopsy tissue, localize specific biomarkers within the tissue samples, and monitor proteomic changes due to disease progression or drug therapy. Several studies have demonstrated that protein profiles can be obtained directly from pre-specified regions on a tissue sample and that these profiles can potentially be used for human disease diagnosis and patient prognosis [30,31]. Protein localization within thin tissue sections has also been demonstrated through MS imaging [9,24,25]. In this approach, spectra are collected across the tissue section in an array of spots or pixels. Molecular ion images are then reconstructed from this high-resolution analysis by plotting the intensity of one (or more) signals relative to other pixels in the array, yielding localization information on hundreds of protein signals in two-dimensional space. The relative protein abundance can then be displayed in terms of the x, y coordinates of the original tissue section.

These techniques serve as a valuable discovery tool since no prior knowledge of the proteins of interest is required and no target-specific reagents (e.g., antibodies) are required. Comparison of the protein patterns derived from mass spectrometric analyses of different tissues has identified potential disease-specific molecular biomarkers and possible drug targets. Imaging experiments have been performed to monitor protein changes across both normal and diseased organs, monitor drug localization within treated tissues, and determine protein changes as a response to a given treatment. This technology has wide applications including the areas of chemistry, biochemistry, biology, and clinical research. The purpose of this chapter is to summarize the current state of profiling and imaging MALDI–MS as applied to tissue analysis, focusing on sample preparation, and illustrating the technological capabilities with several applications.

2. Methods

2.1. Sample preparation

Two general modes of data acquisition are used for direct tissue analysis, termed profiling and imaging, as shown in Fig. 1. Profiling proteins in a tissue section involves depositing matrix droplets (typically between 50 nL and 0.5 μL) on selected areas of a tissue (Fig. 1A). Each of the droplets is independently analyzed, yielding a protein profile for a relatively large, discrete region within the tissue. Typically, between 5 and 10 droplets are deposited on a given tissue section. This approach is commonly performed to quickly determine the protein content from a specific morphological region within a tissue section. Using these data, protein profile comparisons between tissues (i.e., disease and normal or treated and untreated) can be performed to determine specific molecular changes. Profiling experiments can be accomplished using most commercially available MALDI TOF instruments without additional software.

Imaging involves collecting MS data in a regular pattern or array across a matrix-coated tissue section (Fig. 1B). This approach typically provides a higher resolution molecular picture of the tissue from data collected from hundreds or thousands of pixels across the tissue surface. An image displaying the distribution

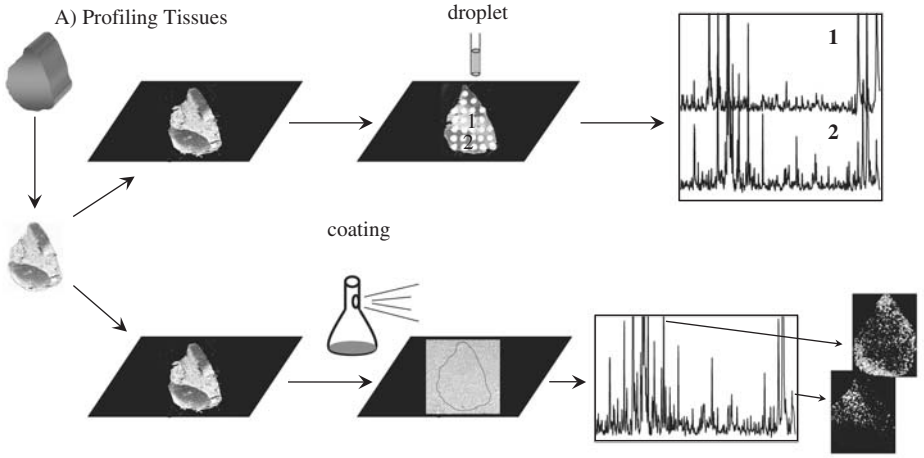


Fig. 1. Direct tissue sample analysis by mass spectrometry. Tissue samples are sectioned and mounted onto MALDI target plates. Matrix solution is applied to the tissue surface by (A) manually depositing droplets in specific regions on the tissue or (B) coating the entire tissue section with either a thin matrix film or small (pL) matrix droplets. The resulting spectra demonstrate the protein expression complexity from discrete tissue regions. (A) Spectra collected by profiling experiments can be compared to determining region- or morphology-specific biomarkers. (B) Data collected by imaging can be interrogated to produce an ion density map, or image, for each *m*/*z* signal detected.

of a specific *m*/*z* signal within the tissue can then be reconstructed. High-resolution images are automatically acquired using custom imaging software that controls the data acquisition parameters [33,34] or through an automated acquisition using commercially available software supplied with the instrument. During image acquisition, a virtual grid is created over the tissue and the MS instrument is directed to move the sample stage and trigger the laser at each point on the grid. While the distance between acquisition points is determined by the user, the image resolution is limited by the laser beam diameter at the target and, for very high-resolution images (<20–30 μm), by the matrix crystal size. An ion density map, or image, is created by integrating the signal intensity for a selected *m*/*z* window at each acquisition point across the grid. An image can be generated for each ion signal detected within the section; therefore, the localization of several hundred proteins can be monitored using this approach.

Applications presented in this chapter have employed a Voyager DE-STR MALDI TOF mass spectrometer and a QStar Pulsar *i* QqTOF mass spectrometer (Applied Biosystems, Foster City, CA). The Voyager instrument utilizes a 337 nm nitrogen laser with a 2.5 ns pulse and a repetition rate of 2 Hz. The instrument was operated in linear mode under delayed extraction conditions. The laser spot size on target is ~50 μm in diameter. The QStar Pulsar instrument is equipped with a MALDI source and a 337 nm nitrogen laser operating at a 20 Hz repetition rate. The laser spot size is ~200 μm in diameter on target.

2.1.1. Tissue treatment

Careful sample preparation is important in order to obtain high quality protein and peptide data from intact tissue samples [35]. Surgically removed tissues are loosely wrapped in aluminum foil, frozen in liquid nitrogen, and stored at −80ºC. Frozen tissues are cut into 5–20 μm thick sections with a cryostat at below freezing temperatures (Fig. 1). Commonly used histological sectioning procedures involve embedding the tissue in cutting polymers such as optimal cutting temperature (OCT) polymer, wax, or agar in order to stabilize the tissue and hold it in place. These procedures are not optimal for MS analysis since, during sectioning, the polymer is spread over the surface of the tissue and can suppress desorption of surface compounds. However, small amounts of polymer can be used at the base of the sample, away from the surface to be cut. This protects the sample from contamination while attaching the tissue to the cryostat probe. Sections are then mounted onto MALDI target plates. Additional sections can be collected for immunohistochemical staining and histology to define regions of interested identified by the cellular morphology. Prior to analysis, sections are dried in a vacuum desiccator for at least 1 h to remove excess water from the sample. This step can be performed either before or after matrix deposition with minor protein signal differences. Washing protocols, typically as an ethanol gradient for 15–30s at each

step, have also been used to remove salts and other contaminants from the sections before matrix deposition [35,36]. This procedure enhances the protein profile quality by increasing the signal-to-noise ratio for many ions, but it should be used with caution as some molecules may be removed during this process.

Several types of MALDI plates are available for sample analysis. Stainless steel or gold-coated target plates manufactured for the instrument are traditionally used for tissue analysis. We have found that the gold-coated surface is preferred over stainless steel since the contrast from the gold enhances visualization of the tissue on the target plate as well as morphology changes within the tissue [35]. Conductive glass slides can also be cut to the dimensions of the manufactured plates and serve as an alternative to the traditional target plate. Using this medium, the same section can be stained with MALDI-compatible tissue stains and analyzed [36]. This allows for the direct selection of morphological regions of interest for analysis without the uncertainty of alignment accuracy between a stained section and the section to be analyzed. Several nuclear-specific stains have been demonstrated to yield quality protein profiles with minimal stain-specific distortions, while maintaining visualization of the cell nucleus for cell identification and cancer diagnosis.

2.1.2. Matrix deposition

Perhaps the most important aspect of sample preparation is matrix application. Matrix and matrix solvent selection for tissue section analysis is key to obtaining quality protein profiles [35]. Sinapinic acid (SA, 3,5-dimethoxy-4-hydroxycinnamic acid) is routinely used in most tissue profiling and imaging experiments, although other matrices can be used, such as α-*cyano*-4-hydroxycinnamic acid (HCCA) [37], 2,5-dihydroxybenzoic acid (DHB), or combinations of these. SA matrix concentrations typically range from 20 mg/mL to a saturated matrix solution. Comparative studies have shown that a 50:50:0.1 organic:water:trifluoroacetic acid (TFA) matrix solution, where the organic component is ethanol or acetonitrile, yields the best general protein profile. TFA concentrations ranging from 0.3 to 1% have been used to enhance the maximum number of proteins analyzed. For improved sensitivity, these solvent parameters should be optimized for each experimental tissue.

Depending on the experimental goal, different methods for matrix deposition may be applied including depositing discrete large (nL) droplets regionally and coating the entire tissue section with matrix. Fig. 2 presents a visual comparison of three matrix deposition methods (described below in more detail).

For regional profiling experiments, relatively large matrix spots can be deposited directly on the sample by the use of either a low volume automatic pipette or an automatic syringe pump attached to a small capillary (Fig. 2B). Matrix solution is typically deposited in volumes ranging from 50 nL to 1 μL. Stained histology

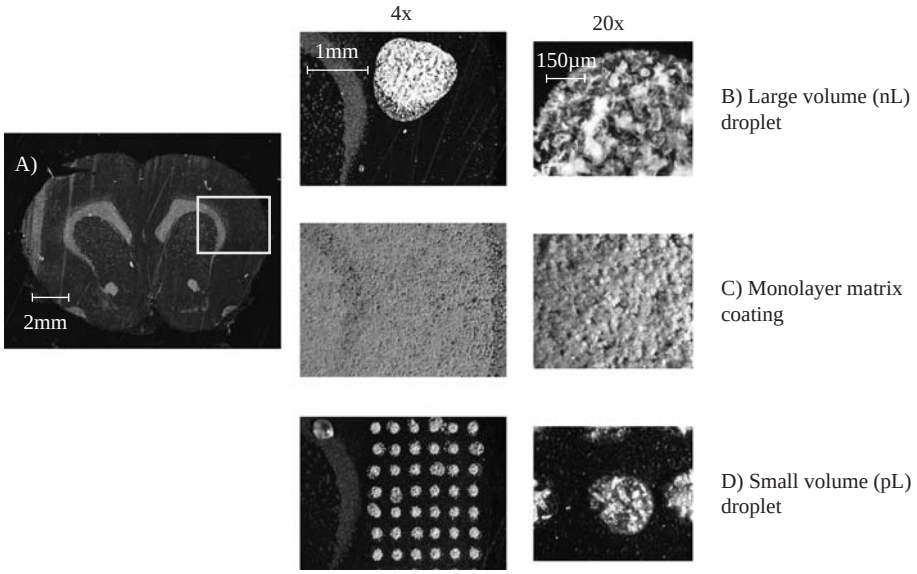


Fig. 2. Approaches to matrix deposition. (A) Photomicrograph of a mouse brain section, 12 μm thick, prior to matrix deposition. The yellow rectangle represents the tissue region coated with matrix and expanded in the 4× panel. Three serial mouse brain sections were collected for matrix deposition. (B) Matrix droplet created by depositing two sequential 0.1 μL droplets of matrix onto the same region of the tissue surface. (C) Thin film of matrix created by 10 spray cycles using a glass nebulizer. (D) Array of pL droplets deposited using a robotic ejector. Panels from a 4× and a 20× magnification are shown for each example.

slides of adjacent tissue sections can be used to guide matrix deposition. Matrix droplets deposited in this fashion incorporate proteins from a fairly large cell population. A 100 nL droplet produces a matrix spot ~1 mm in diameter on the surface of the tissue section, incorporating proteins from thousands of cells. MS analysis is then performed on each individual droplet. Data collected across the droplet surface from 100 or more laser shots can be averaged to yield a profile reflecting the protein content from cells within the droplet surface.

For high-resolution images, tissues may be coated with matrix. Two primary approaches have been applied to tissue coating: applying a homogenous layer of matrix crystals using spray deposition techniques (Fig. 2C) and coating the tissue with an array of small (pL) matrix droplets (Fig. 2D). The goal of coating the tissue is to achieve a homogeneous field of small matrix crystals at the resolution limit of the instrument. A critical step in this process is to ensure that the tissue remains wet with the matrix solution so that proteins can be incorporated during crystallization, while minimizing protein delocalization. Matrix crystal concentration is also important. A matrix coating that is too light will produce too few crystals, resulting in poor incorporation of peptides and proteins, low intensity MS signals, and poor image resolution.

Various mechanisms have been applied to deposit matrix uniformly onto the tissue surface. One technique utilizes a deactivated glass spray venturi nebulizer (VWR Scientific Products, USA) to spray a fine matrix mist over the sample (Fig. 2C) [35]. In this approach, the solvent mixture is only in contact with deactivated glass, thus eliminating any corrosive reaction found in other metal-based technologies. For complete coverage of the tissue, a cycle of matrix spray coatings is performed. Typically, small volumes of matrix are sprayed over the plate surface until the entire tissue surface is damp. Care is taken to not overwet the tissue as proteins can delocalize. The sample plate is usually held vertically ~20–30 cm from the sprayer. During matrix application, the sprayer is moved back and forth, parallel to the target, to evenly apply matrix over the entire sample surface. Following each coating cycle, the sample may be allowed to dry briefly before the next coating cycle is performed. Typically, 8–10 cycles of coating and drying are applied to achieve an even, dense homogeneous crystal field. Since different tissue types can exhibit different surface properties, the number of coating cycles can vary. Surface properties may also affect the final crystal size. Another technique applied in several of the examples discussed in this chapter requires immersing the tissue in matrix solution containing a high percentage (70–95%) of organic solvent for a short period of time (several minutes) and allowing it to dry. Following immersion, several spray-coating cycles of matrix may also be applied to the tissue surface to increase the crystal coverage and protein incorporation. This technique can result in a thin homogeneous field of small matrix crystals. However, this approach may be sensitive to temperature, humidity, or other environmental factors.

Another approach to tissue coating utilizes robotics to deposit an array of small matrix droplets across the surface of the tissue (Fig. 2D) [38]. Matrix droplets, ranging from 0.1 pL to 10 μL, can be ejected onto the surface to coat the sample; droplets less than 100 μm in diameter have been generated with this technology. The deposition of multiple droplets onto the same position on the tissue enhances protein incorporation while increasing the total number of matrix molecules deposited. Protein delocalization is minimized to within the droplet area. Since the introduction of this technology, several matrix deposition systems have become commercially available. Precisely targeting regions of interest by direct integration of histology with tissue profiling has advanced these capabilities for seamless profiling analysis [39]. In this approach, a pathologist selects tissue regions of interest on a digitally scanned, stained tissue section image. These selected regions are transferred to the robotic spotter for matrix deposition on the target tissue section. Digital images of the prepared MALDI plates are then analyzed to automatically locate all MALDI spots. This information is transferred to generate a custom plate geometry file for automated MS analysis of each matrix droplet.

Robotic droplet ejection holds several advantages over spray coating for matrix deposition including quality sample preparation and data reproducibility for mass spectrometric imaging. The small discrete droplets can be sampled hundreds of

times with the laser, improving signal quality compared to a thin matrix layer in which the crystalline layer is quickly depleted. However, image resolution for robotic deposition is limited to droplet size. Generally, ~200 μm resolution is obtained with this device. Image resolution for thin matrix layer deposition is limited by the laser diameter at the target surface (30–100 μm or more) and crystal size, which can be 1–50 μm in diameter depending on the application method.

During experimental design, the type of information desired should help guide the choice of the method of matrix application. As mentioned above, deposition of relatively large matrix droplets for tissue profiling should be used when the goal is to simply compare large regions of tissue. Data files resulting from these experiments are relatively small, 4–5 MB for 30 profiles. Such experiments can be analyzed without complex computer programs and the entire procedure can be performed in less than 1 h. High-resolution images yield data files on the order of 1–2 GB and can take several hours for acquisition, depending on the image resolution (distance between pixels) and instrument spot-to-spot speed. Additionally, image analysis requires additional software viewing programs. However, the data collected through imaging results in a more detailed molecular picture of protein distribution and have been used in many applications including drug distribution within a tissue and protein changes due to tumor infiltration.

2.1.3. Protein identification

Identification of molecular weight markers of interest has been performed in many studies. Tissues are homogenized and proteins separated by one or more dimensions of liquid chromatography, such as ion exchange or reverse-phase chromatography. Collected fractions are screened by MALDI–MS and the fractions containing the intact molecular weight of interest are digested with trypsin or another suitable protease. The resulting peptides are analyzed by peptide mass mapping and sequenced by tandem MS. The data collected are compared to current protein databases for protein identification. This approach has led to the identification of many tumor or tumor-stage-specific markers determined in profiling or imaging experiments [29,40,41].

2.2. Biological applications

2.2.1. Glioma profiling and imaging

The molecular protein patterns present within defined histological regions can be obtained by direct mass spectrometric profiling of tissue sections for biomarker discovery applications. Comparison of the protein profiles collected from different disease stages, such as tumor and non-tumor tissues, can reveal stage-specific proteins based on signal expression changes between groups. These results provide potential diagnostic biomarkers or therapeutic targets. Mass spectrometric protein

profiling has been used to determine diagnostic markers for several disease states including brain tumors [41], lung cancer [31], colon cancer [29], prostate cancer, and breast cancer.

As an example, gliomas, representing ~25,000 new cases per year, are the most common primary brain tumors and one of the more fatal human malignancies. Accurate clinical diagnosis for these tumors is critical since diagnosis and treatment decisions are based almost exclusively on tissue histology. Research therefore has focused on developing a protein profiling tool for accurate glioma diagnosis. In this work, samples from over 120 patients were collected and analyzed by direct tissue protein profiling. Spectra from these samples were acquired in the *m*/*z* range of 2000–70,000 to avoid detector saturation from matrix signals. Typically, between 200 and 500 ion signals were detected in each spectrum. For example, analysis of a matrix droplet (100 nL) deposited directly on a 12 μm thick section of a human glioma resulted in the complex spectrum shown in Fig. 3. The *m*/*z* range from 4300 to 12,500 is shown in expanded intensity to emphasize low intensity

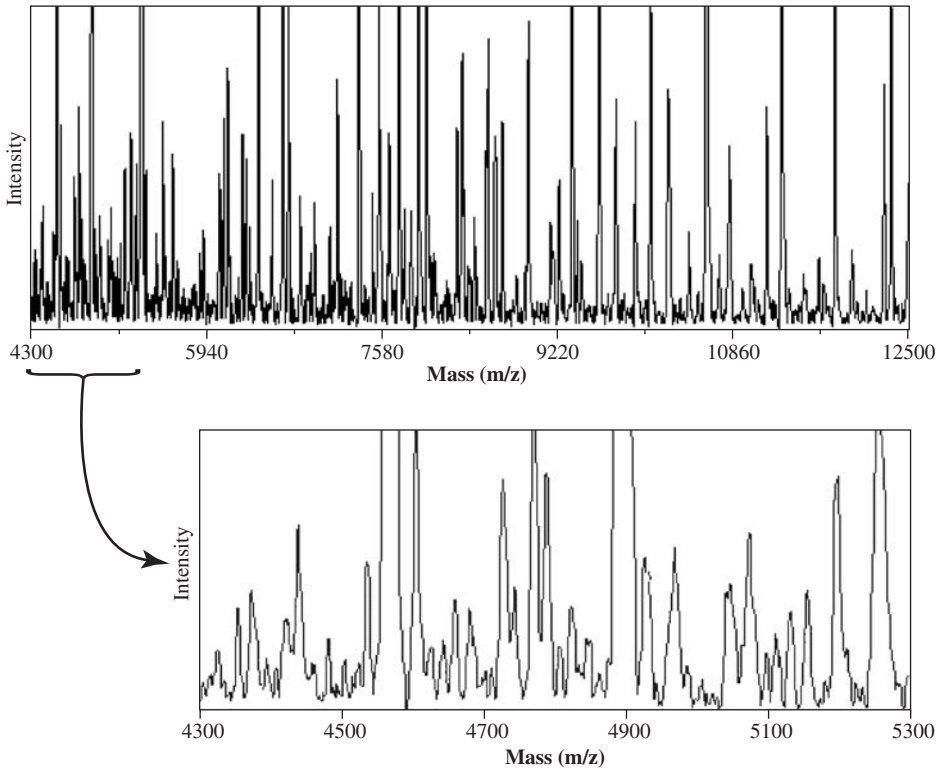


Fig. 3. Protein profile generated from direct MS analysis of a matrix droplet deposited on a 12 μm human glioma section. The intensity scale has been expanded to display low intensity ion signals. The inset, displaying the *m*/*z* range 4300–5300, demonstrates the complexity of the data collected from tissue samples. Over 50 ion signals can be recognized in the inset alone; over 500 signals were observed across the entire spectrum.

signals in the spectrum. Additionally, the range from 4300 to 5300 Da has been expanded on the *m*/*z* scale to show the complexity of the signal pattern. Over 500 signals were detected across the entire mass range recorded for this sample.

Profiles collected from different tissue regions yield information on the cellular content within the tissue. Spectra from morphologically similar tissue regions demonstrate the general reproducibility of this technique for complex biological samples. For example, in Fig. 4, two mass spectra (1 and 2) collected from a high-grade, aggressive glioma region, and two (3 and 4) from the surrounding non-tumor tissue are presented. The original tissue section (Fig. 4A) and corresponding spectra (Fig. 4B) are shown. The spectra collected from similar cellular regions are comparable, reflecting the similarity in protein content within this tumor sample and the reproducibility of the mass spectrometric analysis from tissue sections. On the other hand, spectra collected from different histological regions of the tissue highlight the changes in protein expression between tumor and non-tumor areas. Statistical analysis of the protein patterns collected from these samples has identified a suite of signals that could distinguish tumor from non-tumor tissues as well as stages of tumor progression (grade 2, 3, and 4 gliomas). These data allow one to segregate disease stages based on the protein pattern themselves. Additional analysis suggests that specific mass spectrometric

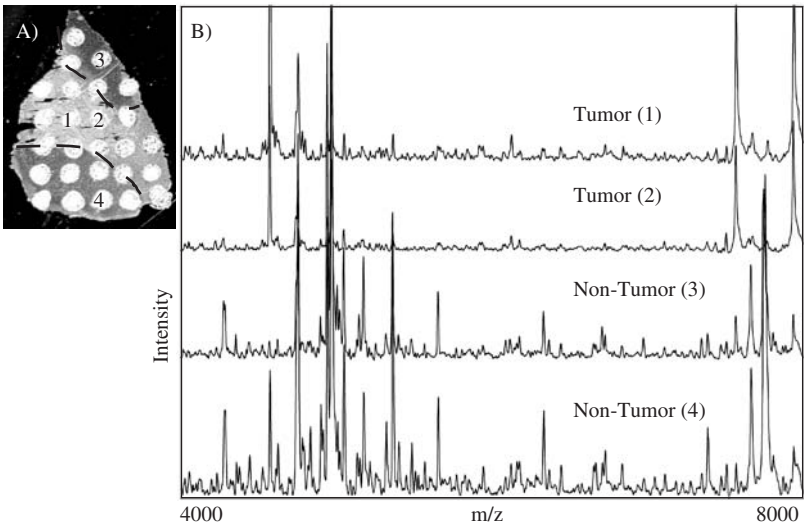


Fig. 4. Protein profiles generated from direct MS analysis of a human glioma biopsy. (A) Histology identified two distinct regions within the presented tissue, a region of high-grade glioma tissue (outlined by two dashed, - - -, lines) surrounded by non-tumor tissue. (B) Two mass spectra (1 and 2) from the tumor region and two (3 and 4) from the surrounding non-tumor tissue are presented. Spectra collected from similar cellular regions are comparable, reflecting the similarity in protein content within this tumor sample. Spectra collected from different histological regions of the tissue are markedly dissimilar, indicating the changes in protein expression between the tumor and the non-tumor areas.

protein patterns can be used to accurately indicate patient prognosis. Such studies suggest that these tools, in combination with current techniques, may be useful for a more accurate molecular diagnosis of disease states.

MS imaging has also been performed on many of these tissues to serve as a visual comparative analysis of differential protein expression. For example, Fig. 5

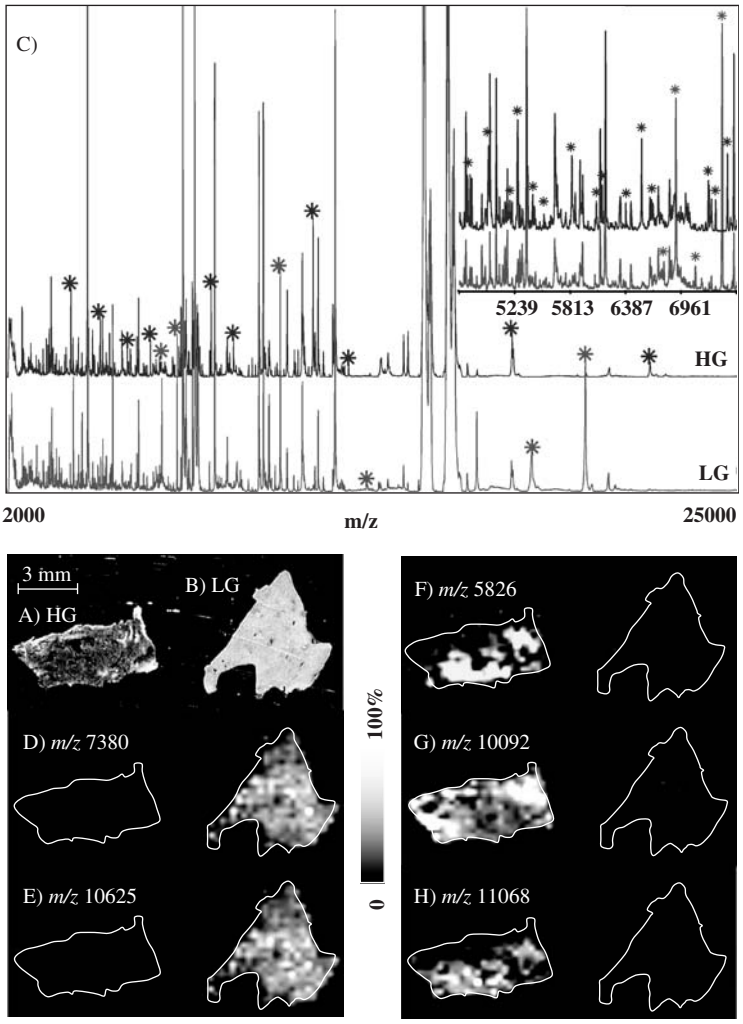


Fig. 5. Imaging MS analysis of a high- and low-grade glioma. (A and B) Initial tumor sections, thaw-mounted on the MALDI sample plate prior to matrix deposition. Samples were coated with matrix by robotic ejection and imaged at 250 μm resolution. (C) Averaged protein profiles obtained for each section after image analysis. Stars indicate signals with differential signal intensity observed between the two profiles. (D–H) Ion density maps selected from different *m*/*z* values. The maps are depicted as gray scale images with white representing the highest signal intensity and black the lowest. (Reprinted with permission from Toxicologic Pathology, Volume 33:1, pp. 97, 2005.)

presents a simultaneous analysis of two glioma biopsies, a low-grade, grade II, tumor and a high-grade, grade IV, tumor. The tumor sections analyzed are shown in Fig. 5A and B. A robotic instrument was used to deposit matrix droplets across the tissue surface, which was then imaged with a resolution of 250 μm. Fig. 5C presents averaged protein profiles obtained from the low- and high-grade glioma samples. Many signals expressed with differing intensities between the low- and high-grade tumors as described above. Fig. 5D–H presents ion images for several of these signals. The mass signals shown were found to be statistically significant in distinguishing these disease states.

2.2.2. Brain and brain tumor imaging

Discrete, unique features exist within the brain, many of which have complex functions. This complexity is shown in the mouse brain imaged by MALDI–MS in the top panel of Fig. 6. A flash frozen mouse brain was sectioned (12 μm) on a cryostat and mounted onto a gold-coated stainless-steel sample plate. A sequential section was collected on a glass slide and stained with hematoxylin and eosin for feature recognition (Fig. 6A). The tissue section was immersed in matrix solution (20 mg/mL SA in 90/10/0.1 ethanol/water/TFA), dried, and coated with several spray cycles of matrix solution (20 mg/mL SA in 50/50/0.1 acetonitrile/water/TFA) as described above. The tissue section was then imaged with a 50 μm spot-to-spot resolution; each spectrum per spot was an average of 40 laser shots. Several of the hundreds of mass signals observed were selected and an ion density map for each signal was constructed (Fig. 6B–F). These images show the distinct localization of proteins within the tissue section as measured by MS. For example, *m/z* 18,412 is localized primarily to the corpus callosum while *m/z* 6720 is abundant in the striatum.

Tumor models serve as useful tools in measuring protein differences distinctive for growing tumors as well as studying protein changes across tumor margins. As an example, a tumor, resulting from the injection of GL261 brain cancer cells into a mouse brain, and the surrounding brain tissue were imaged by MS (bottom panel of Fig. 6). The tumor developed in the left lateral ventricle of the brain with evidence of tumor migration distinguishable in the right lateral ventricle. The mouse brain was sectioned and coated with matrix. A photomicrograph of the section prior to imaging is presented in Fig. 6G. The sample was imaged with 20 laser shots per spectrum at an imaging resolution of 110 μm. Fig. 6H–R presents several ion images reconstructed following sample analysis. Some of the presented patterns reflect proteins specific to the growing tumor including those at *m/z* 6924 and 11,307 while others are localized to non-tumor regions such as *m/z* 18,412.

2.2.3. Drug imaging and drug response profiling

Imaging MS can also be used to map the location of administered drugs from various organs and monitor protein changes as a response to drug treatment. One

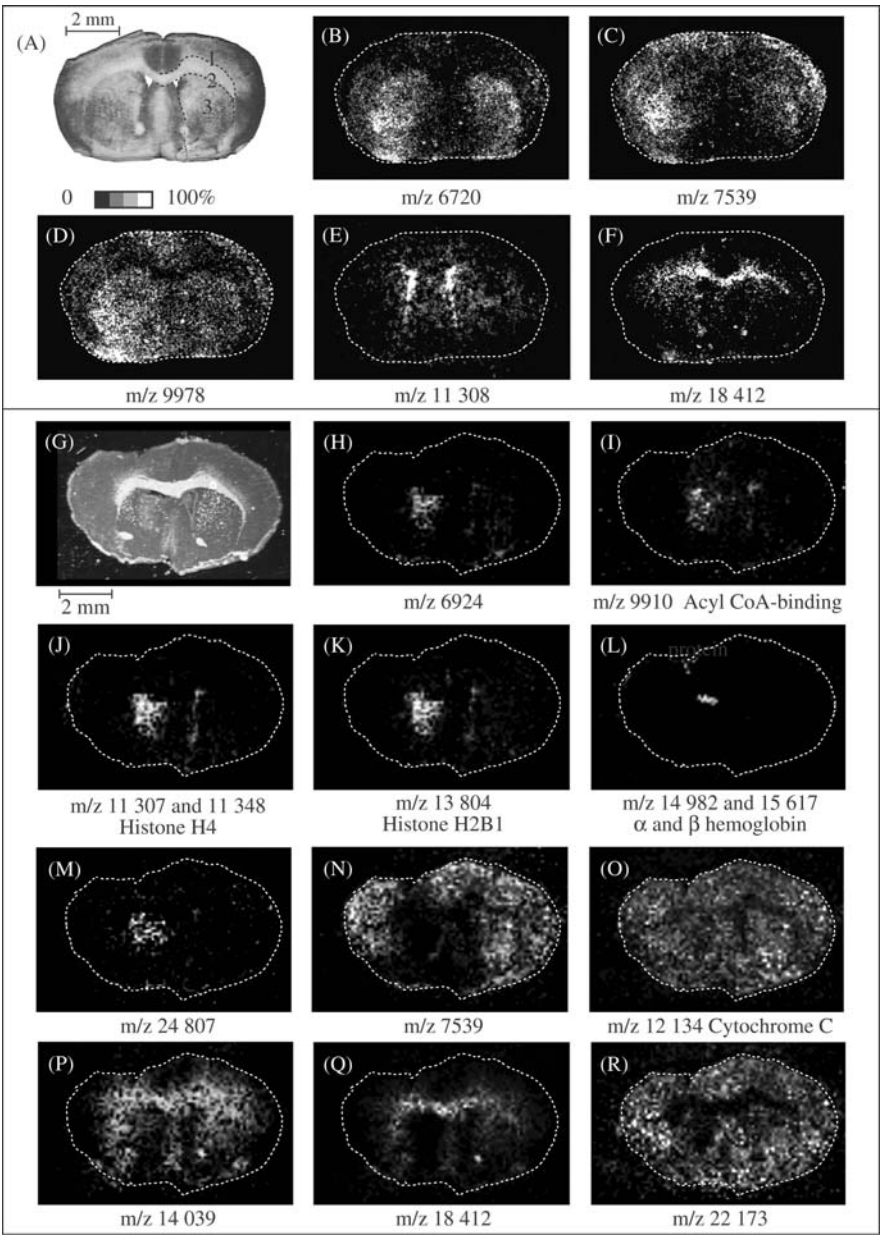


Fig. 6. Imaging MS of a healthy and a diseased mouse brain section. Top panel: (A) Photomicrograph of a 12 μm hematoxylin and eosin stained brain tissue section at bregma +0.75 mm. Differences in anatomic brain substructures can be distinguished including (1) cerebral cortex, (2) corpus callosum, and (3) striatum. A sequential section was collected, coated with matrix, and imaged at 50 μm. (B–F) Ion density maps obtained at different *m/z* ratios are displayed. Bottom panel: (G) Photomicrograph of a brain section containing a tumor, 12 μm thick, prior to matrix coating. The tissue region containing the tumor is outlined in gray. The sample was coated with matrix and imaged at 110 μm. (H–R) Ion density maps obtained at different *m/z* ratios are displayed. The ion density maps are depicted as pseudo-color images with white representing the highest protein concentration and black the lowest. Images represented in the top and bottom panels were independently normalized by intensity. (Reprinted with permission from the authors: Richard Caprioli, Pierre Chaurand and Sarah Schwartz, Anal Chem 76: 87A–93A, 2004.)

advantage to this approach over other methods, such as the use of radiolabeled compounds or fluorescent markers, is that both intact drug as well as drug metabolites can be simultaneously analyzed. Additionally, drug-induced protein changes that are dose and time dependent can be determined. These results yield a potential tool for predicting clinical efficacy. Tentative studies also suggest that MALDI–MS analysis of tissue samples prior to drug treatment can predict drug resistance.

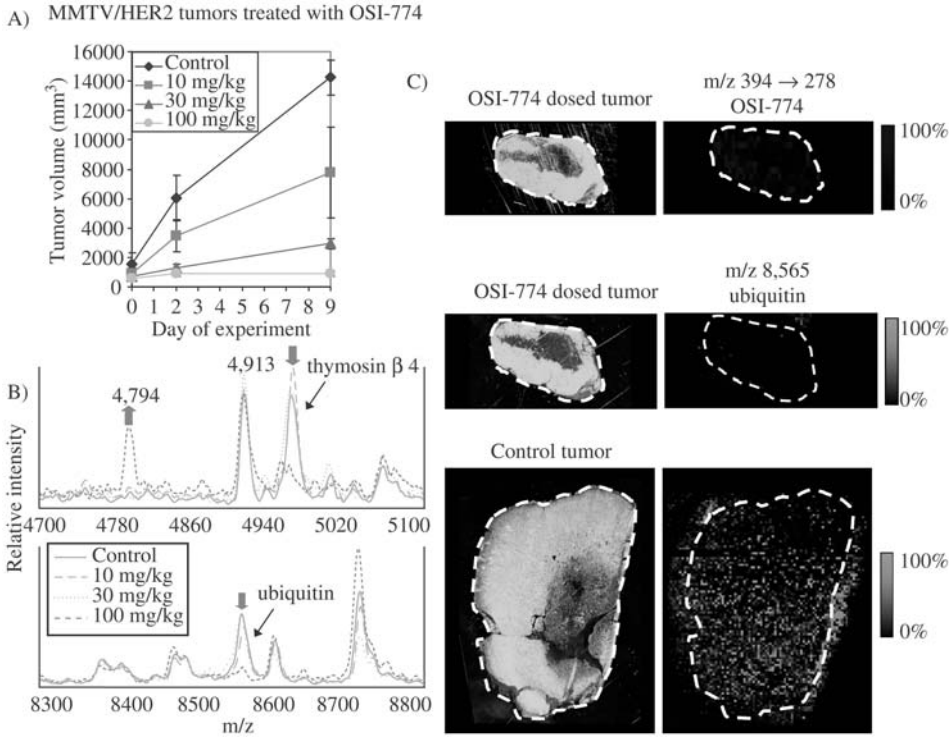


Fig. 7. OSI-774 localization in mouse mammary tumors and dose-dependent therapy-induced proteomic changes. (A) Two (approximately) 1 mm^3 pellets from an MMTV/HER2 mammary tumor were implanted via small surgical incisions in the s.c. space of the right and left dorsum of wild-type female FVB mice, respectively. Once they reached a volume of >250 mm^3, treatment with OSI-774 at the indicated doses p.o. was started. Twenty hours after the first dose, all right-sided tumors were harvested, and the mice continued on daily therapy for the next 9 days. Left-sided tumor volumes are shown; bars, $3 \pm$ SD. (B) Mass spectral analysis of MMTV/HER2 sections from tumors harvested 20 h after a variable single dose of OSI-774 reveals several proteomic changes induced from the 100 mg/kg dose. (C) Wild-type FVB mice bearing MMTV/HER2 tumors measuring $\sim$200 mm^3 were treated with 100 mg/kg OSI-774 p.o. and harvested after 16 h. Two serial sections of the treated tumor and one section of the untreated tumor were analyzed by imaging MS. A mass spectral image of OSI-774 performed on the first treated section demonstrates that OSI-774 is distributed throughout the tumor section, but is less evident in the necrotic center. Selected protein images for ubiquitin, performed on the second treated tumor section and the untreated tumor section, demonstrate that ubiquitin is markedly down-regulated in the treated tumor. (Reprinted with permission from the American Association for Cancer Research from Cancer Research 64: 9093–9100, 2004.)

An example of this work focused on analyzing proteomic changes following tumor treatment with the therapeutic agent, OSI-774 [22]. OSI-774 suppresses tumor growth by inhibiting the EGFR tyrosine kinase. Dose-dependent studies on mouse models bearing mammary tumors demonstrated that treatments of 100 mg/kg were necessary for tumor growth arrest (Fig. 7A). MS profiling of tissues treated with increasing OSI-774 doses (10, 30, and 100 mg/kg) demonstrated that specific protein pattern changes were dose dependent. For example, the Tβ4 (m/z 4965) and ubiquitin (m/z 8565) ion signals were significantly decreased while a signal corresponding to an E-cadherin fragment (m/z 4794) was increased in the tissue samples treated with 100 mg/kg OSI-774 (Fig. 7B). Imaging experiments were also used to confirm these protein changes and correlate the protein localization with the presence of OSI-774. As shown in Fig. 7C, OSI-774, as measured through selected reaction monitoring (described previously in ref. [22]) of the transition m/z 394.2 $\rightarrow$ 278.1, is distributed throughout the tumor section but is less predominant in the necrotic center. Imaging experiments performed on a sequential section demonstrate that m/z 8565, corresponding to ubiquitin, is expressed at lower levels in the treated sample when compared to a control, untreated tumor.

Further studies were used to demonstrate that profiles generated by MALDI–MS analysis of tumor samples may be indicative of drug resistance. Two types of tumors, originating from the same transgenic founder cell line, were found to have a differing response to Herceptin, a therapeutic agent that binds the HER2 receptor and inhibits tumor growth. The growth of F2-1282 tumors is inhibited by Herceptin while Fo5 tumors are resistant (Fig. 8A), even though the tumor cells express similar levels of HER2. Analysis of these tumor lines following Herceptin treatment showed differential protein expression including an increase in m/z 9212 in the Herceptin sensitive tumors that was not present in the resistant tumors (Fig. 8B). These studies suggest that biomarker changes specific for tumor response to treatment can be monitored.

3. Discussion

There are many potential clinical applications for direct molecular analysis of tissue samples. Profiling and imaging technologies using MALDI–MS have the advantage of providing molecular weight-specific protein profiles and high-resolution (30–50 μm resolution) protein images. Results from these experiments detail the molecular complexities of the samples as well as precise protein localization within a tissue. Cellular processes occurring within healthy and diseased states can therefore be mapped with high sensitivity and specificity. In this approach, tissue analysis by MS serves as an extraordinary discovery tool since many hundreds of proteins, the identities of which do not need to be known in advance, can be monitored in a comparative study. The relative concentrations of the markers can

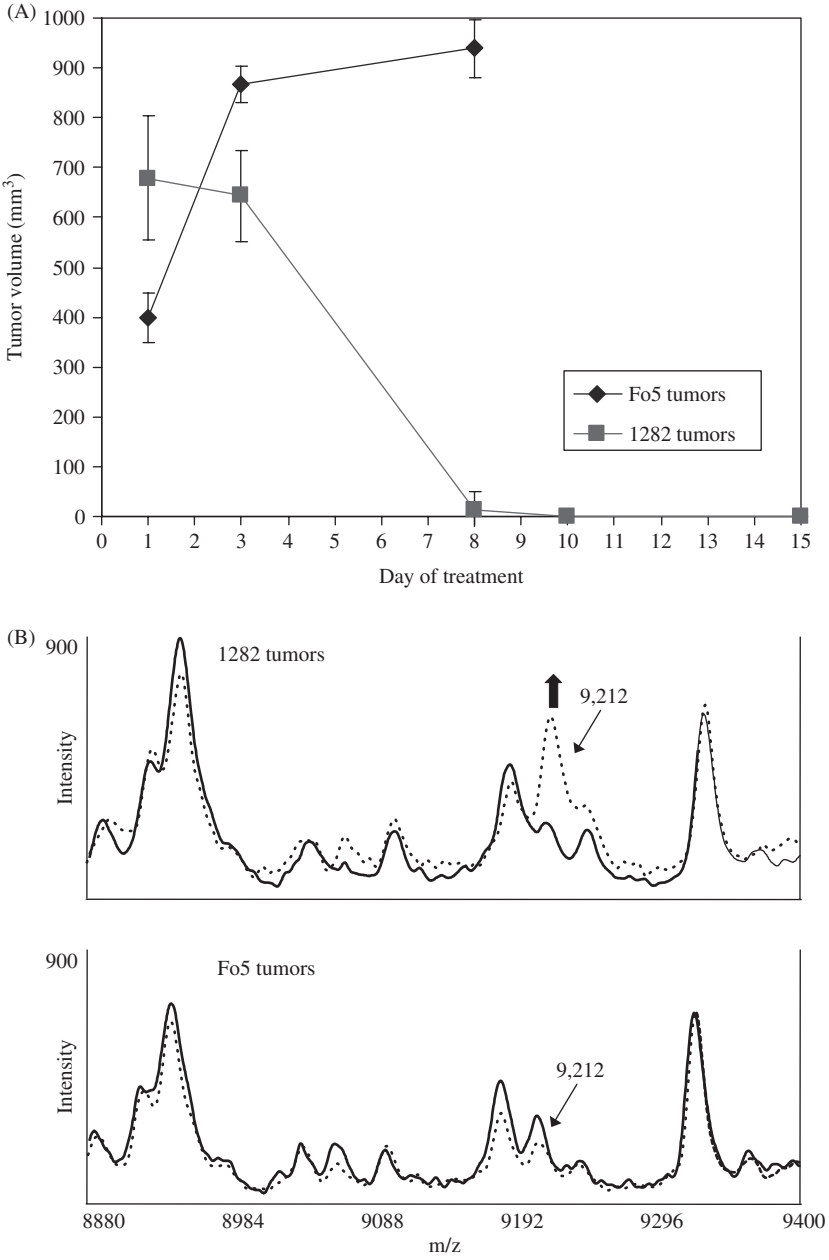


Fig. 8. Drug-induced proteome changes predict for therapeutic resistance. (A) Mice bearing established >300 mm^3 Fo5 (Herceptin-resistant) and 1282 (Herceptin-sensitive) tumors were treated with Herceptin 30 mg/kg i.p. twice a week. Each data point represents mean tumor volume ± SD (Fo5, $n = 3$; 1282, $n = 6$). (B) Fo5 and 1282 tumors of equivalent size were harvested 24 and 48 h after a single dose of Herceptin i.p. and subjected to mass spectral proteomic profiling analysis. An example of a statistically significant change observed after Herceptin treatment in the 1282 tumors not observed in the Fo5 tumors is shown. The solid line trace (—) represents control, untreated tumors while the dotted line trace (· · ·) represents Herceptin-treated tumors. (Reprinted with permission from the American Association for Cancer Research from Cancer Research 64: 9093–9100, 2004.)

be assessed either across a tissue section or between sample groups. Once a potential biomarker is identified, its sub-cellular localization, concentration, regulation mechanism, and function can be further investigated to increase the understanding of a specific disease or disease state at the molecular level. It is also important to note that the biomarker changes discussed here were detected in unfixed, frozen tissues that did not require additional preparation before analysis. Therefore, tissue analysis can be performed in a high-throughput, accurate manner.

Clinically, profiling and imaging MS may provide a molecular assessment of disease states, including tumor diagnosis and progression, and aid in patient prognosis and designing treatment strategies. These tools can work in concert with classifying disease based on pathology to augment clinical diagnosis and ultimately, patient treatment and outcome. Furthermore, predicting therapeutic response through biomarker discovery is of increasing importance. The assessment of drug treatment efficacy and identification of early signs of treatment resistance through mass spectrometric proteomic studies can serve as an invaluable tool to eventually improve the clinical outcome of patients. Studies correlating protein expression changes with drug distribution illustrate the ability to evaluate the effects of targeted therapeutics within the disease microenvironment. In combination with other traditional assays, an improved understanding of the effects of new therapeutics can be attained, thereby enhancing drug development and the success of clinical trials.

4. Future trends

One of the features of profiling and imaging MS is its potential as a discovery tool in research areas involving biomarker identification as well as protein or drug localization without the use of antibodies or fluorescent markers. Little *a priori* information is necessary for differences in protein expression patterns to be identified. Future developments are required to make this technology more useful and routinely accessible including increasing the number of proteins detected, improving detection sensitivity at higher masses, decreasing the analysis time for protein identification, and instrumental improvements to allow for faster data acquisition and processing and higher resolution images.

Currently, mass spectrometric analysis directly off tissue detects hundreds of proteins in the mass range of 2000–100,000 Da, although beyond about m/z 30,000 ion detection efficiency and resolution decrease. Improved instrumentation suggests the potential for ion detection at higher molecular weight analysis with enhanced signal resolution. In addition, since matrix application is necessary for protein ionization, detection is limited to primarily hydrophilic proteins or proteins bound through non-covalent interactions. Methods to increase hydrophobic protein solubilization under MS-compatible conditions should increase the number of biological molecules analyzed by MALDI–MS. Membrane-bound proteins are for the most

part inaccessible to the matrix solvent and, as a result, are not incorporated into the matrix crystals. Recent efforts have therefore focused on applying MS-compatible surfactants to disrupt cell membranes and solubilize membrane-bound proteins [42].

Efforts have also focused on developing faster and higher throughput methods for protein identification. Although MALDI–MS profiling and imaging approaches precisely determine molecular weights for specific ions, these data are usually not sufficient for protein identification. Many existing protein databases do not take into account protein post-translational modifications in molecular weight calculations, thus limiting the utility of database searches for proteins based solely on molecular weight. Other approaches to identify proteins of interest, such as using HPLC coupled to tandem mass spectrometers, are more robust and better utilize protein database search engines, but the process can be very time consuming. Exploring avenues including on-tissue digestions, soft landing approaches [43–45], and high energy CID [46,47] may lead to faster protein identification while utilizing existing protein databases.

Finally, in order for tissue analysis by MS to become a routine technology, instrumental modifications and software development are important. Tissue imaging requires optimization of the laser repetition rate as well as data downloading and processing times. Lasers with faster repetition rates ($>$1 kHz) and improved electronics should reduce profile and image acquisition times. Acquisition algorithms that can record high-throughput data are also being developed. Data processing and mining tools are being designed that enhance biomarker selection and accuracy. Although the MS techniques discussed here do not allow for sub-cellular analysis, new developments may allow for this application in the future. Demands for higher resolution images have also led to the development of smaller laser spot diameters (on the order of 1–10 μm). Improved matrix application to generate smaller crystals will aid these studies.

5. Conclusion

Utilizing MS to profile and image tissue samples combines the advantages of identifying molecular differences between tissue samples and maintaining analyte spatial information. Protein changes measured with this technology suggest the application of MALDI–MS to identify molecular mechanisms associated with tumor development or tumor response to therapeutic treatments. Protein patterns, protein localization, and protein–protein co-localization can be monitored as well as how these patterns change in compromised tissues. A single acquisition yields hundreds of signals with relatively high mass accuracy measurements. This large quantity of data not only produces information related to tissue morphology, but also identifies potential molecular biomarkers for diseased cells or targets for drug development. The discovery aspect of this research combined with the quantity of

information obtained for each tissue sample provides a new, innovative tool for use in biological research.

Acknowledgments

The authors would like to acknowledge Pierre Chaurand, Michelle Reyzer, Hans-Rudolf Aerni, and Malin Andersson (Mass Spectrometry Research Center, Vanderbilt University) for generating some of the data presented here. We would also like to thank Robert Weil, Steven Toms (Cleveland Clinic Foundation), and Reid Thompson (Vanderbilt University) for providing the brain tumors, Mahlon Johnson (Department of Pathology, Vanderbilt University) for histological analysis of the brain tumors, Claudine Laurent and Pat Levitt (Department of Pharmacology, Vanderbilt University) for providing the normal mouse brains, and Dennis Hallahan (Department of Radiation Oncology, Vanderbilt University) for providing the mouse brain containing the tumor. We would also like to acknowledge funding by the National Institutes of Health (grant GM 58008) and the National Cancer Institute (grant CA86243-02).

References

1. Hillenkamp, F., Karas, M., Beavis, R.C. and Chait, B.T., Matrix-assisted laser desorption ionization mass-spectrometry of biopolymers. *Anal. Chem.*, **63**, A1193–A1202 (1991).
2. Karas, M. and Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10000 Daltons. *Anal. Chem.*, **60**, 2299–2301 (1988).
3. Bahr, U., StahlZeng, J., Gleitsmann, E. and Karas, M., Delayed extraction time-of-flight MALDI mass spectrometry of proteins above 25,000 Da. *J. Mass Spectrom.*, **32**, 1111–1116 (1997).
4. Brown, R.S. and Lennon, J.J., Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser-desorption ionization linear time-of-flight mass-spectrometer. *Anal. Chem.*, **67**, 1998–2003 (1995).
5. Guilhaus, M., Mlynski, V. and Selby, D., Perfect timing: time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, **11**, 951–962 (1997).
6. Aebersold, R. and Goodlett, D.R., Mass spectrometry in proteomics. *Chem. Rev.*, **101**, 269–295 (2001).
7. Pandey, A. and Mann, M., Proteomics to study genes and genomes. *Nature*, **405**, 837–846 (2000).
8. Bergquist, J., Cells on the target—matrix-assisted laser-desorption/ionization time-of-flight mass-spectrometric analysis of mammalian cells grown on the target. *Chromatographia*, **49**, S41–S48 (1999).
9. Caprioli, R.M., Farmer, T.B. and Gile, J., Molecular imaging of biological samples: localization of peptides and proteins using MALDI–TOF MS. *Anal. Chem.*, **69**, 4751–4760 (1997).
10. Chaurand, P., Stoeckli, M. and Caprioli, R.M., Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. *Anal. Chem.*, **71**, 5263–5270 (1999).
11. Garden, R.W., Moroz, L.L., Moroz, T.P., Shippy, S.A. and Sweedler, J.V., Excess salt removal with matrix rinsing: direct peptide profiling of neurons from marine invertebrates using

matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *J. Mass Spectrom.*, **31**, 1126–1130 (1996).
12. Holland, R.D., Wilkes, J.G., Rafii, F., Sutherland, J.B., Persons, C.C., Voorhees, K.J. and Lay Jr. J.O., Rapid identification of intact whole bacteria based on spectral patterns using matrix-assisted laser desorption/ionization with time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, **10**, 1227–1232 (1996).
13. Jimenez, C.R. and Burlingame, A.L., Ultramicroanalysis of peptide profiles in biological samples using MALDI mass spectrometry. *Exp. Nephrol.*, **6**, 421–428 (1998).
14. Jiménez, C.R., Li, K.W., Dreisewerd, K., Spijker, S., Kingston, R., Bateman, R.H., Burlingame, A.L., Smit, A.B., van Minnen, J. and Geraerts, W.P.M., Direct mass spectrometric peptide profiling and sequencing of single neurons reveals differential peptide patterns in a small neuronal network. *Biochemistry*, **37**, 2070–2076 (1998).
15. Krishnamurthy, T. and Ross, P.L., Rapid identification of bacteria by direct matrix-assisted laser desorption/ionization mass spectrometric analysis of whole cells. *Rapid Commun. Mass Spectrom.*, **10**, 1992–1996 (1996).
16. Li, K.W., Hoek, R.M., Smith, F., Jimenez, C.R., van der Schors, R.C., van Veelen, P.A., Chen, S., van der Greef, J., Parish, D.C. and Benjamin, P.R., Direct peptide profiling by mass-spectrometry of single identified neurons reveals complex neuropeptide-processing pattern. *J. Biol. Chem.*, **269**, 30288–30292 (1994).
17. Li, L.J., Garden, R.W. and Sweedler, J.V., Single-cell MALDI: a new tool for direct peptide profiling. *Trends Biotechnol.*, **18**, 151–160 (2000).
18. Lynn, E.C., Chung, M.C., Tsai, W.C. and Han, C.C., Identification of Enterobacteriaceae bacteria by direct matrix-assisted laser desorption/ionization mass spectrometric analysis of whole cells. *Rapid Commun. Mass Spectrom.*, **13**, 2022–2027 (1999).
19. Moroz, L.L., Gillette, R. and Sweedler, J.V., Single-cell analyses of nitrergic neurons in simple nervous systems. *J. Exp. Biol.*, **202**, 333–341 (1999).
20. Palmer-Toy, D.E., Sarracino, D.A., Sgroi, D., LeVangie, R. and Leopold, P.E., Direct acquisition of matrix-assisted laser desorption/ionization time-of-flight mass spectra from laser capture microdissected tissues. *Clin. Chem.*, **46**, 1513–1516 (2000).
21. Redeker, V., Toullec, J.Y., Vinh, J., Rossier, J. and Soyez, D., Combination of peptide profiling by matrix-assisted laser desorption/ionization time of flight mass spectrometry and immunodetection on single glands or cells. *Anal. Chem.*, **70**, 1805–1811 (1998).
22. Reyzer, M.L., Hsieh, Y.S., Ng, K., Korfmacher, W.A. and Caprioli, R.M., Direct analysis of drug candidates in tissue by matrix-assisted laser desorption/ionization mass spectrometry. *J. Mass Spectrom.*, **38**, 1081–1092 (2003).
23. Rubakhin, S.S., Garden, R.W., Fuller, R.R. and Sweedler, J.V., Measuring the peptides in individual organelles with mass spectrometry. *Nat. Biotechnol.*, **18**, 172–175 (2000).
24. Stoeckli, M., Chaurand, P., Hallahan, D.E. and Caprioli, R.M., Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nat. Med.*, **7**, 493–496 (2001).
25. Todd, P.J., Schaaff, T.G., Chaurand, P. and Caprioli, R.M., Organic ion imaging of biological tissue with secondary ion mass spectrometry and matrix-assisted laser desorption/ionization. *J. Mass Spectrom.*, **36**, 355–369 (2001).
26. Troendle, F.J., Reddick, C.D. and Yost, R.A., Detection of pharmaceutical compounds in tissue by matrix-assisted laser desorption/ionization and laser desorption/chemical ionization tandem mass spectrometry with a quadrupole ion trap. *J. Am. Soc. Mass Spectrom.*, **10**, 1315–1321 (1999).
27. van Veelen, P.A., Jimenez, C.R., Li, K.W., Wildering, W.C., Geraerts, W.P.M., Tjaden, U.R. and van der Greef, J., Direct peptide profiling of single neurons by matrix-assisted laser-desorption ionization mass-spectrometry. *Org. Mass Spectrom.*, **28**, 1542–1546 (1993).

28. Dreisewerd, K., Kingston, R., Geraerts, W.P.M. and Li, K.W., Direct mass spectrometric peptide profiling and sequencing of nervous tissues to identify peptides involved in male copulatory behavior in *Lymnaea stagnalis*. *Int. J. Mass Spectrom.*, **169**, 291–299 (1997).
29. Chaurand, P., DaGue, B.B., Pearsall, R.S., Threadgill, D.W. and Caprioli, R.M., Profiling proteins from azoxymethane-induced colon tumors at the molecular level by matrix-assisted laser desorption/ionization mass spectrometry. *Proteomics*, **1**, 1320–1326 (2001).
30. Schwartz, S.A., Weil, R.J., Johnson, M.D., Toms, S.A. and Caprioli, R.M., Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression. *Clin. Cancer Res.*, **10**, 981–987 (2004).
31. Yanagisawa, K., Shyr, Y., Xu, B.J., Massion, P.P., Larsen, P.H., White, B.C., Roberts, J.R., Edgerton, M., Gonzalez, A., Nadaf, S., Moore, J.H., Caprioli, R.M. and Carbone, D.P., Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet*, **362**, 433–439 (2003).
32. Pierson, J., Norris, J.L., Aerni, H.-R., Svenningsson, P., Caprioli, R.M. and Andrén, P.E., Molecular profiling of experimental Parkinson's disease: direct analysis of peptides and proteins on brain tissue sections by MALDI mass spectrometry. *J. Proteome Res.*, **3**, 289–295 (2004).
33. Spengler, B. and Hubert, M., Scanning microprobe matrix-assisted laser desorption ionization (SMALDI) mass spectrometry: instrumentation for sub-micrometer resolved LDI and MALDI surface analysis. *J. Am. Soc. Mass Spectrom.*, **13**, 735–748 (2002).
34. Stoeckli, M., Farmer, T.B. and Caprioli, R.M., Automated mass spectrometry imaging with a matrix-assisted laser desorption ionization time-of-flight instrument. *J. Am. Soc. Mass Spectrom.*, **10**, 67–71 (1999).
35. Schwartz, S., Reyzer, M. and Caprioli, R., Direct tissue analysis using matrix-assisted laser desorption/ionization mass spectrometry: practical aspects of sample preparation. *J. Mass Spectrom.*, **38**, 699–708 (2003).
36. Chaurand, P., Schwartz, S.A., Billheimer, D., Xu, B.J., Crecelius, A. and Caprioli, R.M., Integrating histology and imaging mass spectrometry. *Anal. Chem.*, **76**, 1145–1155 (2004).
37. Kruse, R. and Sweedler, J.V., Spatial profiling invertebrate ganglia using MALDI MS. *J. Am. Soc. Mass Spectrom.*, **14**, 752–759 (2003).
38. Aerni, H.R., Cornett, D.S. and Caprioli, R.M., Automated acoustic matrix deposition for MALDI sample preparation. *Anal. Chem.*, **78**, 827–834 (2006).
39. Cornett, D.S., Mobley, J.A., Dias, E.C., Andersson, M., Arteaga, C.L., Sanders, M.E. and Caprioli, R.M., A novel histology directed strategy for MALDI–MS tissue profiling that improves throughput and cellular specificity in human breast cancer. *Mol. Cell. Proteomics*, **5**, 1975–1983 (2006).
40. Xu, J., Stolk, J.A., Zhang, X., Silva, S.J., Houghton, R.L., Matsumura, M., Vedvick, T.S., Leslie, K.B., Badaro, R. and Reed, S.G., Identification of differentially expressed genes in human prostate cancer using subtraction and microarray. *Cancer Res.*, **60**, 1677–1682 (2000).
41. Schwartz, S.A., Weil, R.J., Thompson, R.C., Shyr, Y., Moore, J.H., Toms, S.A., Johnson, M.D. and Caprioli, R.M., Proteomic-based prognosis of brain tumor patients using direct-tissue matrix-assisted laser desorption ionization mass spectrometry. *Cancer Res.*, **65**, 7674–7681 (2005).
42. Norris, J.L., Porter, N.A. and Caprioli, R.M., Mass spectrometry of intracellular and membrane proteins using cleavable detergents. *Anal. Chem.*, **75**, 6642–6647 (2003).
43. Gologan, B., Takáts, Z., Alvarez, J., Wiseman, J.M., Talaty, N., Ouyang, Z. and Cooks, R.G., Ion soft-landing into liquids: protein identification, separation, and purification with retention of biological activity. *J. Am. Soc. Mass Spectrom.*, **15**, 1874–1884 (2004).
44. Ouyang, Z., Takáts, Z., Blake, T.A., Gologan, B., Guymon, A.J., Wiseman, J.M., Oliver, J.C., Davisson, V.J. and Cooks, R.G., Preparing protein microarrays by soft-landing of mass-selected ions. *Science*, **301**, 1351–1354 (2003).
45. Washburn, M.P., Soft landing for protein chips. *Nat. Biotechnol.*, **21**, 1156–1157 (2003).

46. Lin, M., Campbell, J.M., Mueller, D.R. and Wirth, U., Intact protein analysis by matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, **17**, 1809–1814 (2003).
47. Medzihradszky, K.F., Campbell, J.M., Baldwin, M.A., Falick, A.M., Juhasz, P., Vestal, M.L. and Burlingame, A.L., The characteristics of peptide collision-induced dissociation using a high-performance MALDI–TOF/TOF tandem mass spectrometer. *Anal. Chem.*, **72**, 552–558 (2000).
48. Chaurand, P., Schwartz, S.A., Reyzer, M.L. and Caprioli, R.M., Imaging mass spectrometry: principles and potentials. *Toxicol. Pathol.*, **33**, 92–101 (2005).
49. Chaurand, P., Schwartz, S.A. and Caprioli, R.M., Profiling and imaging proteins in tissue sections by MS. *Anal. Chem.*, **76**, 87A–93A (2004).
50. Reyzer, M.L., Caldwell, R.L., Dugger, T.C., Forbes, J.T., Ritter, C.A., Guix, M., Arteaga, C.L. and Caprioli, R.M., Early changes in protein expression detected by mass spectrometry predict tumor response to molecular therapeutics. *Cancer Res.*, **64**, 9093–9100 (2004).

Medical Applications of Mass Spectrometry
K. Vékey, A. Telekes and A. Vertes (editors)

Chapter 24

Brief outlook

KÁROLY VÉKEY[a,*], ANDRÁS TELEKES[b,1], and AKOS VERTES[c,2]

[a]*Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary*
[b]*National Institute of Oncology, Budapest, Hungary*
[c]*W.M. Keck Institute for Proteomics Technology and Applications, Department of Chemistry, The George Washington University, Washington, DC, USA*

Mass spectrometry is widely used in analytical chemistry and biochemistry, and is just making a spectacular entry into the biomedical and clinical fields. Together with genetic techniques, it signals a new era in medical treatment: the arrival of advanced instrumentation to the bedside, the characterization of the human organism at the molecular level, and the use of this information to improve healthcare.

Mass spectrometry has proven itself in biomedical research; several chapters in this book are a testament to its applications in various disciplines. It is also accurate and robust enough to be widely used, e.g., in the pharmaceutical industry, and it is well adapted to routine applications. These features allow mass spectrometry to enter into the clinical field, where its potential is enormous. It can be used for a variety of purposes, from diagnosis to prognosis and to help selecting the optimal treatment for a patient. It is suitable for high-throughput analysis, requires a small amount of sample (e.g., a drop of blood suffices), produces accurate results, and can be used according to good clinical practice. In spite of these advantages, mass spectrometry is just emerging as a tool for clinical work. To some degree this may be the consequence of a communication barrier between physicians and

*Corresponding author. Tel.: +361-438-1158; Fax: +261-438-1157. E-mail: vekey@chemres.hu (K. Vékey).
[1]E-mail: telekes@oncol.hu (A. Telekes).
[2]E-mail: vertes@gwu.du (A. Vertes).

analytical chemists. The complex technical nature of mass spectrometry and the data analysis involved makes it difficult to enter the field without formal training. The primary purpose of this book is to break down this communication barrier, and to illustrate to medical professionals the powerful contribution mass spectrometry can provide to various aspects of medicine.

Looking into the near future, i.e., the coming 5–10 years, several trends can be predicted. The availability of mass spectrometric techniques is beginning to reframe our thinking about the medical questions that can be addressed. This might pave the way for the emergence of new disciplines and widen the horizon of medical science. Evidence-based medicine clearly requires all available, objective data, which helps to select the best possible treatment for a given patient. Modern analytical instrumentation is increasingly well suited for high-throughput sample handling and straightforward operation. The capabilities of mass spectrometry and the needs of clinical research and clinical laboratories are converging. For these reasons, the editors firmly believe that we are at the beginning of a new revolution, when mass spectrometry will have an important role in clinics and in everyday medical laboratory practice. We hope this book will help to facilitate this revolution, by surmounting the communication barrier between analytical chemists and medical professionals.

1. Trends in instrumentation

The figures of merit for modern mass spectrometry are so excellent that listing them here might give the impression we are citing a promotional brochure. This technique provides structural information (e.g., the molecular mass and protein sequence); it can be used for quantitation; inorganics, small organics, and macromolecules can all be studied; and it can identify trace impurities and a given class of target compounds in the presence of a complex matrix (e.g., plasma). Mass spectrometry is very sensitive; in some cases even a few molecules (zeptomoles, i.e., 10^{-21} mol) might be detected. It can be integrated with separation techniques (e.g., GC and HPLC), it can be automated, and it is well suited for high-throughput applications (typically needed for screening). The less favorable aspects of mass spectrometry include operational complexity (although in recent years the human interface of the instrumentation has been significantly simplified) and the need for highly qualified personnel. Sample preparation can be time consuming (similar to most other analytical methods), results are not always straightforward to interpret, and mass spectrometers are quite expensive.

There are major efforts underway in instrument development to overcome these disadvantages and to lower the cost of ownership for mass spectrometers. We foresee major improvements during the next decade in the following areas:

(a) *Robotics, automation, and high throughput*. These technologies are aimed at dramatically increasing the efficiency of sample manipulation and analysis.

They are currently available but, as yet, overpriced and underutilized. Fraction collectors, sample spotters, and gel cutters are prime examples of the robots available today. Other aspects of sample preparation can also be performed by dedicated robots. These may involve simple dilution of plasma, adding solvents, reagents, or standards, or taking on the complex series of sample preparation steps and integrating them with mass spectrometry. Proteomics is probably the first area where robots will become widespread. There are systems to manipulate and digest samples but there are also specialized ones designed primarily for mass spectrometry. The operation of most commercial mass spectrometers can be automated. Following sample preparation, samples are often loaded onto standardized well-plates, introduced into and measured by the mass spectrometer automatically.

Throughput depends significantly on the type of service required and whether chromatography is needed. Typical research laboratories might run ~10 samples/day on an instrument. In contrast, service laboratories might average ~100 samples/day and those equipped for high throughput can reach ~1000 samples/day. When working with high throughput, simple and accurate sample labeling (represented by barcodes) and documentation are essential. The typical time requirement for obtaining a mass spectrum is less than a second; therefore, samples can be introduced into the mass spectrometer every minute or so. Chromatography, however, requires 10–100 min. Not surprisingly, therefore, major efforts are directed to eliminating or speeding up chromatographic separation. As chromatography is extremely powerful, leaving it out to gain time involves a significant compromise. This may be alleviated by using more advanced mass spectrometry, such as tandem instrumentation. As described in Chapter 6, the separating power of the first mass spectrometer stage can substitute chromatography, whereas the second stage provides identification/quantitation. This can be considered a main reason for the success of tandem mass spectrometry-based neonatal screening methods (see Chapters 12 and 16). A recently emerging different approach is ion mobility separation in combination with mass spectrometry. The sample components can be sorted based on their molecular size by ion mobility in less than a second. This can be rapidly followed by MS or tandem MS providing a total analysis time of a few seconds.

(b) *Integration with bioinformatics, automated workflow, and documentation*. Mass spectrometry has the disadvantage that it produces an intimidating amount of data that is difficult to interpret. Usually the services of a specialist are needed but humans do not perform well in a high-throughput setting. There are advanced bioinformatics tools to help interpreting mass spectral data, especially in proteomics (see Chapter 8). In fact, it is practically impossible to manually evaluate the gigabytes of data generated by proteomics instrumentation each day without using bioinformatics. However, using these tools also requires expertise. We foresee a fast development in automatic data evaluation, especially, but not exclusively, in proteomics. Mass spectrometers, data evaluation, bioinformatics, and documentation need to

be (and will be) much better integrated in the near future. Mass spectrometry-based "total solutions" are needed for particular medical applications probably integrated into expert systems. These are likely to involve artificial intelligence to improve diagnostic accuracy. The development of such expert systems is a prerequisite to make mass spectrometry ubiquitous in the clinical field.

(c) *Placing of mass spectrometers in hospitals*. Presently, mass spectrometers working in the biomedical field are usually located in chemistry, biochemistry, or pharmaceutical departments, core facilities, or research institutes, rarely in hospitals or clinics. This is mostly the consequence of the relatively complex nature of the instrumentation and of the required expertise. In the coming decade this is likely to change, and more hospitals will acquire mass spectrometers for their own use, i.e., for "routine" analysis or screening. For this to become reality, the streamlining of the analytical process should progress to the point of "black box" operation. Ideally, the sample (e.g., 100 μL of plasma) is loaded into the sample holder of the instrument (that comprises both mass spectrometry and data evaluation), and in a few minutes a report is generated. Current trends of the instrumentation market support the notion of such systems emerging within the next decade.

(d) *Advanced sampling interfaces and imaging*. New developments in imaging mass spectrometry can provide molecular distributions from biomedical samples in the form of 2D images (see, for example, Chapter 23). The combination of these advances with the traditional methods of histology might be among the first areas of direct medical applications. Tissue sections, biopsies, and even *in vivo* studies of, e.g., skin or tissue exposed in surgery are amenable for this technology. Special probes or sampling interfaces of the mass spectrometer can be inserted into body cavities to collect the most relevant sample for *in situ* investigations.

2. Emerging systems approach

A few decades ago, most of our knowledge on biological systems was based on studying small organic molecules, mostly metabolites. Due to advances in analytical capabilities and high-throughput technologies, our understanding of genomics, and later proteomics, tremendously expanded during the last decade. Many chapters in the present book illustrate this growing understanding and the significance of proteomics in the biomedical field. At present, it is common to consider the information gathered by genetics, proteomics, metabolomics, etc., separately. However, our rapidly expanding knowledge of biomedical systems is ready for integration. This integration in the field of biology is termed systems biology, and it is rapidly being embraced by the scientific community. Related institutes are

spawned around the world and leading universities establish systems biology departments (e.g., Harvard Medical School). The premise of this new discipline is that we need to look at the organism as a whole, understand how it works, and (in the clinical context) base interventions on this knowledge.

Mass spectrometry will provide key information in the field of systems biology. It is well suited for high-throughput studies and is equally well adapted for studying small organics (e.g., metabolomics and metabonomics), proteins (proteomics and peptidomics), and various other macromolecules (lipidomics, etc.). Analytical instrumentation will be integrated with bioinformatics tools, to a higher degree than it is at present, to extract useful information from the plethora of data obtained.

3. Mass spectrometry in translational medicine

Various chapters in this book illustrate active research in the biomedical field using mass spectrometry, but as yet there are few clinical applications. This is likely to change in the near future. It is being recognized that additional efforts are needed to translate basic research results to bedside practice. As outlined above, further development of instrumentation is also needed to realize the full potential of mass spectrometry in a clinical environment.

In the long term we expect the appearance of expert systems, which utilize various types of analytical tools (mass spectrometry being a prime component), base their assessment on utilizing the systems biology approach, and give advice to the clinician about the state of patient and possible courses of treatment. In the short and medium terms a diversity of trends can be discerned where mass spectrometry has an impact on translational medicine.

(a) *Collaboration* between mass spectrometry facilities and clinics will further improve. Integration of mass spectrometry equipment with other types of clinical laboratory techniques will eventually be common, but this is likely to require more time.

(b) *Improved diagnostics* will be the primary entry point for mass spectrometry into the clinical field. This is already in place in pediatrics, oncology seems to be the next main target and other fields (e.g., infectious and autoimmune diseases) are likely to follow. Diagnostics is closely related to finding (and validating) new biomarkers, which is (and likely to remain) a very active field. However, the use of biomarkers is not limited to diagnostics.

(c) *Selecting optimal treatment* is also important. There are very few examples as yet, but early studies indicate that mass spectrometry will have a significant contribution. It can be used to monitor the course of a disease (e.g., the progression of cancer), which may help to select if, when, and which treatment is necessary. Clinically the same disease may or may not respond to a given treatment. Mass spectrometry (with the help of appropriate

biomarkers) may define subgroups of a disease. These subgroups could be treated differently (e.g., by using different drugs or deciding what other treatment option is the best course of action). The early distinction of subgroups (e.g., the identification of responders for a given treatment) may be life saving. For example, if cancer is treated by an inefficient drug, the patient might be weakened due to the side effects, while the cancer continues to advance. Less importantly, more accurate targeting may reduce hospital expenses as well by eliminating an expensive, but for the given purpose inefficient, drug. Mass spectrometry may also help to identify novel therapeutic targets.

(d) *Prognostic* application of mass spectrometry is also a possibility. Preliminary research indicates that biomarkers may be found that correlate with the likely disease progression.

(e) *Personalized medicine* means taking into account genetic and proteomic information for a given patient in medical decision making. Ongoing efforts in pharmacogenomics to tailor medication to the genetic makeup of an individual are in the exploratory phase. Corresponding advances are underway based on proteomic information for an individual gathered by mass spectrometry. Optimal treatment may be decided in light of the level of a given biomarker for a particular patient. This approach is not yet in place, but it is expected to become a reality in the medium term. Therapeutic drug monitoring is an early version of this concept (see Chapter 13). It is already used at various medical centers, but currently it is far from being widespread. Here the idea is to determine and control optimum drug dosage for a given patient, based on the measurement of drug concentrations, e.g., in the blood or urine. These approaches are likely to become more and more common and complex in the near future.

The promise of translating the results of research with the aid of mass spectrometry into clinical reality is exciting. In our view, this prospect hinges on two basic conditions. First, as we have explained above (p. 557–558), the human interface of these devices has to become user friendly to the point where a nurse can walk up to them, insert a sample, and collect the results. Reading and interpreting the results still might require a specialist, but this is also the case with other sophisticated equipment in, e.g., imaging radiology. Second, for mass spectrometers to become ubiquitous in medicine, the cost of ownership of the instrumentation has to drop to levels comparable to separation equipment. With current advances in instrumentation both of these objectives might fast become a reality.

Index

This page intentionally left blank