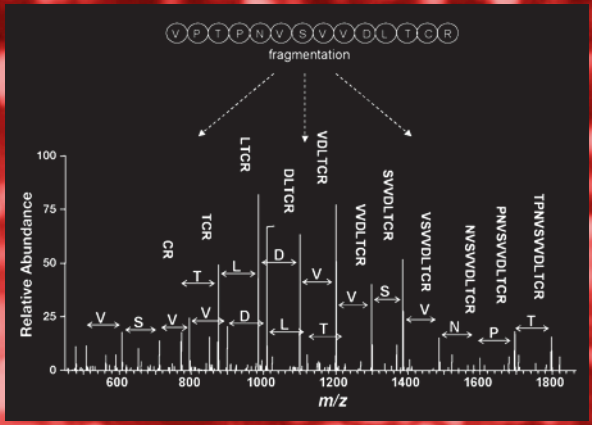
METHODS IN MOLECULAR BIOLOGY™ 367

Mass Spectrometry Data Analysis in Proteomics

Edited by

Rune Matthiesen



HUMANA PRESS

Mass Spectrometry Data Analysis in Proteomics

METHODS IN MOLECULAR BIOLOGY™

John M. Walker, SERIES EDITOR

387. Serial Analysis of Gene Expression (SAGE): *Digital Gene Expression Profiling*, edited by *Kare Lehmann Nielsen, 2007*

386. Peptide Characterization and Application Protocols, edited by *Gregg B. Fields, 2007*

385. Microchip-Based Assay Systems: *Methods and Applications*, edited by *Pierre N. Floriano, 2007*

384. Capillary Electrophoresis: *Methods and Protocols*, edited by *Philippe Schmitt-Kopplin, 2007*

383. Cancer Genomics and Proteomics: *Methods and Protocols*, edited by *Paul B. Fisher, 2007*

382. Microarrays, Second Edition: *Volume 2, Applications and Data Analysis*, edited by *Jang B. Rampal, 2007*

381. Microarrays, Second Edition: *Volume 1, Synthesis Methods*, edited by *Jang B. Rampal, 2007*

380. Immunological Tolerance: *Methods and Protocols*, edited by *Paul J. Fairchild, 2007*

379. Glycovirology Protocols, edited by *Richard J. Sugrue, 2007*

378. Monoclonal Antibodies: *Methods and Protocols*, edited by *Maher Albitar, 2007*

377. Microarray Data Analysis: *Methods and Applications*, edited by *Michael J. Korenberg, 2007*

376. Linkage Disequilibrium and Association Mapping: *Analysis and Application*, edited by *Andrew R. Collins, 2007*

375. In Vitro Transcription and Translation Protocols: *Second Edition*, edited by *Guido Grandi, 2007*

374. Biological Applications of Quantum Dots, edited by *Marcel Bruchez and Charles Z. Hotz, 2007*

373. Pyrosequencing® Protocols, edited by *Sharon Marsh, 2007*

372. Mitochondrial Genomics and Proteomics Protocols, edited by *Dario Leister and Johannes Herrmann, 2007*

371. Biological Aging: *Methods and Protocols*, edited by *Trygve O. Tollefsbol, 2007*

370. Adhesion Protein Protocols, *Second Edition*, edited by *Amanda S. Coutts, 2007*

369. Electron Microscopy*: Methods and Protocols, Second Edition*, edited by *John Kuo, 2007*

368. Cryopreservation and Freeze-Drying Protocols, *Second Edition*, edited by *John G. Day and Glyn Stacey, 2007*

367. Mass Spectrometry Data Analysis in Proteomics, edited by *Rune Matthiesen, 2007*

366. Cardiac Gene Expression: *Methods and Protocols*, edited by *Jun Zhang and Gregg Rokosh, 2007*

365. Protein Phosphatase Protocols: edited by *Greg Moorhead, 2007*

364. Macromolecular Crystallography Protocols: *Volume 2, Structure Determination*, edited by *Sylvie Doublie, 2007*

363. Macromolecular Crystallography Protocols: *Volume 1, Preparation and Crystallization of Macromolecules*, edited by *Sylvie Doublie, 2007*

362. Circadian Rhythms: *Methods and Protocols*, edited by *Ezio Rosato, 2007*

361. Target Discovery and Validation Reviews and Protocols: *Emerging Molecular Targets and Treatment Options, Volume 2*, edited by *Mouldy Sioud, 2007*

360. Target Discovery and Validation Reviews and Protocols: *Emerging Strategies for Targets and Biomarker Discovery, Volume 1*, edited by *Mouldy Sioud, 2007*

359. Quantitative Proteomics, edited by *Salvatore Sechi, 2007*

358. Metabolomics: *Methods and Protocols*, edited by *Wolfram Weckworth, 2007*

357. Cardiovascular Proteomics: *Methods and Protocols*, edited by *Fernando Vivanco, 2006*

356. High-Content Screening: *A Powerful Approach to Systems Cell Biology and Drug Discovery*, edited by *Ken Guiliano, D. Lansing Taylor, and Jeffrey Haskins, 2006*

355. Plant Proteomics: *Methods and Protocols*, edited by *Hervé Thiellement, Michel Zivy, Catherine Damerval, and Valerie Mechin, 2006*

354. Plant–Pathogen Interactions: *Methods and Protocols*, edited by *Pamela C. Ronald, 2006*

353. DNA Analysis by Nonradioactive Probes: *Methods and Protocols*, edited by *Elena Hilario and John. F. MacKay, 2006*

352. Protein Engineering Protocols, edited by *Kristian Müller and Katja Arndt, 2006*

351. *C. elegans:* *Methods and Applications*, edited by *Kevin Strange, 2006*

350. Protein Folding Protocols, edited by *Yawen Bai and Ruth Nussinov 2006*

349. YAC Protocols, *Second Edition*, edited by *Alasdair MacKenzie, 2006*

348. Nuclear Transfer Protocols: *Cell Reprogramming and Transgenesis*, edited by *Paul J. Verma and Alan Trounson, 2006*

347. Glycobiology Protocols, edited by *Inka Brockhausen-Schutzbach, 2006*

346. *Dictyostelium discoideum* Protocols, edited by *Ludwig Eichinger and Francisco Rivero-Crespo, 2006*

345. Diagnostic Bacteriology Protocols, *Second Edition*, edited by *Louise O'Connor, 2006*

344. Agrobacterium Protocols, *Second Edition: Volume 2*, edited by *Kan Wang, 2006*

343. Agrobacterium Protocols, *Second Edition: Volume 1*, edited by *Kan Wang, 2006*

342. MicroRNA Protocols, edited by *Shao-Yao Ying, 2006*

341. Cell–Cell Interactions: *Methods and Protocols*, edited by *Sean P. Colgan, 2006*

340. Protein Design: *Methods and Applications*, edited by *Raphael Guerois and Manuela López de la Paz, 2006*

339. Microchip Capillary Electrophoresis: *Methods and Protocols*, edited by *Charles S. Henry, 2006*

METHODS IN MOLECULAR BIOLOGY™

Mass Spectrometry Data Analysis in Proteomics

Edited by

Rune Matthiesen

Department of Biochemistry and Molecular Biology
University of Southern Denmark, Odense M, Denmark

HUMANA PRESS TOTOWA, NEW JERSEY

999 Riverview Drive, Suite 208
Totowa, New Jersey 07512

www.humanapress.com

This publication is printed on acid-free paper. ♾
ANSI Z39.48-1984 (American Standards Institute) Permanence of Paper for Printed Library Materials.

Cover illustrations: Fig. 2 from Chapter 6 (*foreground*), "Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching" by Alexey I. Nesvizhskii, and Fig. 1 from Chapter 13 (*background*), "Quantitative Proteomics for Two-Dimensional Gels Using Difference Gel Electrophoresis" by David B. Friedman.

.
Production Editor: Rhukea J. Hussain

Cover design by Patricia F. Cleary

For additional copies, pricing for bulk purchases, and/or information about other Humana titles, contact Humana at the above address or at any of the following numbers: Tel.: 973-256-1699; Fax: 973-256-8341; E-mail: orders@humanapr.com; or visit our Website: www.humanapress.com

Printed in the United States of America. 10 9 8 7 6 5 4 3 2 1
1-59745-275-0 (e-book)
ISSN 1064-3745

Library of Congress Cataloging-in-Publication Data
Mass spectrometry data analysis in proteomics / edited by Rune Matthiesen.
p. ; cm. -- (Methods in molecular biology, ISSN 1064-3745 ; 367)
Includes bibliographical references and index.
ISBN-13: 978-1-58829-563-7
ISBN-10: 1-58829-563-X (alk. paper)
1. Proteomics--Methodology. 2. Mass spectrometry. I. Matthiesen, Rune. II. Series: Methods in molecular biology (Clifton, N.J.) ; v. 367.
[DNLM: 1. Proteomics--methods. 2. Spectrum Analysis, Mass--methods. 3. Expressed Sequence Tags. W1 ME9616J v.367 2007 / QU 58.5 M414 2007]
QP519.9.M3M37 2007
572'.6--dc22 2006015500

-

Preface

Currently, a number of books cover the experimental side of proteomics and only briefly describe the theory and practice of data analysis. Additionally, the generation of mass spectrometry (MS) data already has become a high-throughput technique, which calls for efficient high-quality algorithms for data analysis. The intention with this volume is to support researchers in deciding which programs to use in various tasks related to analysis of MS data in proteomics. *Mass Spectrometry Data Analysis in Proteomics* gives a precise description of the theoretical background of each topic followed by accurate descriptions of programs and the parameters best suited for different cases. The focus has been on covering the most common steps in analyzing MS data.

First, different types of MS data and the data format are introduced, followed by a description of the best way to convert raw data into peak list, which can be input to various search engines. For searching databases with MS data, it is important to use databases that are as complete as possible. Sequences can be gathered from several resources, i.e., predicted genes from genomic data, expressed sequence tags (ESTs), and protein sequences. When searching predicted genes from genomic data it is important to consider the accuracy of the predicted exons, for ESTs possible frame-shift is a central issue, and for protein sequences potential signal peptides are worth considering. *Mass Spectrometry Data Analysis in Proteomics* not only gives a report of the available sequence databases, but also covers how to assemble ESTs into nonredundant databases and to further process the sequences into a format suitable for searching with MS data.

In the proteomics field there is a figure of speech, "100% sequence coverage is not enough." The proteomics field has to deal with more than 200 possible modifications of amino acids. When looking for modifications, it is important to have high mass accuracy and it is also an advantage to use other experimental techniques or consider information stated in the literature, which can help to limit the number of possible modifications.

Quantification is an important issue in proteome projects because the dynamic range of protein concentrations is thought to be around 105–106 for eukaryotic cells. Relative quantification of proteins has been available by densitometry of protein spots in two-dimensional electrophoresis (2-DE) gels for many years. Recently, relative quantification of stable isotope-labeled peptides analyzed by liquid chromatography (LC)-MS/MS has drawn great attention. This interest is mainly due to easy automation of the LC-MS/MS runs, whereas

the preparations of 2-DE gels are quite tedious. However, the data analysis is more complex especially if the isotopic peaks from the nonstable isotope-labeled and the stable isotope-labeled version of the peptides are overlapping, as will be discussed further.

Mass Spectrometry Data Analysis in Proteomics mainly describes publicly available programs. However, for computations where no publicly accessible programs are available, commercial programs have been described. The choice of programs in proteomics is, unfortunately, often limited by the data format. The Proteomics Standards Initiative has been established to define community standards for data representation in proteomics. The XML format has recently been suggested as a good tool for interchanging data between applications and is used already by several public and commercial applications.

There are some similarities between the proteomics field of today and the situation in the structural proteomics community in the late 1960s with the lack of public databases containing results and detailed experimental procedures. In the last chapter, strategies for creating such databases are discussed.

Rune Matthiesen

Contents

Contributors

RUEDI AEBERSOLD • *The Institute for Molecular Systems Biology, Swiss Federal Institute of Technology, Hoenggerberg, Zurich, Switzerland*
RIKARD ALM • *Department of Biochemistry, Center for Chemistry and Chemical Engineering, Lund University, Lund, Sweden*
ROLF APWEILER • *EMBL Outstation-Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
JAKOB BUNKENBORG • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*
JEPPE EMMERSEN • *Department of Life Sciences, Aalborg University, Aalborg, Denmark*
DAVID B. FRIEDMAN • *Department of Biochemistry, Mass Spectrometry Research Center and Proteomics Laboratory, Vanderbilt University Medical Center, Nashville, TN*
PER GÄRDÉN • *Complex Systems Division and Lund Swegene Bioinformatics Facility, Department of Theoretical Physics, Lund University, Lund, Sweden*
ALBRECHT GRUHLER • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*
HENNING HERMJAKOB • *EMBL Outstation-Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
KARIN HJERNØ • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*
PETER HØJRUP • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*
OLE NØRREGAARD JENSEN • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*
PHILIP JONES • *EMBL Outstation-Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
RANDALL K. JULIAN, JR. • *EMBL Outstation-Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
KRIS LAUKENS • *CEPROMA, Center for Proteome Analysis and Mass Spectrometry, University of Antwerp, Antwerp, Belgium*
RUNE MATTHIESEN • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*

KUDZAI E. MUTENDA • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*
ALEXEY I. NESVIZHSKII • *Department of Pathology, University of Michigan, Ann Arbor, MI*
SANDRA ORCHARD • *EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
MAGNUS PALMBLAD • *The BioCentre, The University of Reading, Reading, UK*
SHENG PAN • *The Institute for System Biology, Seattle, WA*
CHRISTIAN RAVNSBORG INGRELL • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*
CHRIS TAYLOR • *EMBL Outstation-Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
ALLAN L. THOMSEN • *Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark*
WEIMIN ZHU • *EMBL Outstation-Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*

1

Introduction to Proteomics

Rune Matthiesen and Kudzai E. Mutenda

Summary

Mass spectrometry (MS) has recently become one of the most informative methods for studying proteins. Albiet, MS cannot compete with the detailed structural information obtained by methods such as nuclear magnetic resonance and X-ray crystallography. However, MS is much easier to automate and use as a large-scale technique. Large-scale proteomic methods are valuable for studying the dynamics of the proteins and their posttranslational modification in living cells. Despite the great potential of mass spectrometers, many laboratories are struggling with data analysis and data storage. The complexity of the data analysis stems from the large number of experiments that can be performed by various mass spectrometers. In addition, many mass spectrometers have their own data formats. Performing data analysis on MS data, therefore, requires a rather extensive setup of algorithms and data parsers. In recent years it has become evident that the proteomics society needs standard formats for storing and exchanging data. This has triggered a new problem, which is the invention of several different standard formats. In this chapter, an overview of the most common proteomics experiments with MS, together with an overview of data formats, is presented.

Key Words: Proteomics; mass spectrometry; data formats.

1. Introduction

1.1. The Basic Principles of Proteomics

The term proteomics covers the analysis of expressed proteins in organisms. One of the first tools used in proteomics was two-dimensional gel electrophoresis (2D-GE) introduced in 1975 ***(1)***. Recently, mass spectrometry (MS) techniques have been combined with 2D-GE for direct and systematic identification of polypeptides. Proteins resolved by 2D-GE can be enzymatically digested in-gel, or digested during blotting onto membranes containing immobilized trypsin ***(2)***. Trypsin cleaves specifically after the basic residues arginine (Arg) and lysine (Lys), if not followed by proline (Pro), and is by far the most commonly used

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

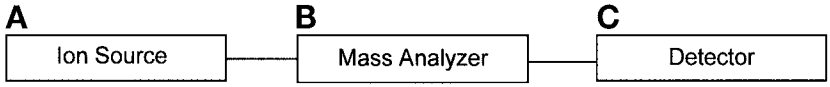


Fig. 1. Outline of a mass spectrometer. **(A)** The ion source for electrospray ionization is at atmospheric pressure, and the source for MALDI is under vacuum. **(B)** The mass analyzer (commonly used in proteomics) can be a time-of-flight, an ion trap, a quadrupole, an FTICR, or a hybrid of the aforementioned analyzers. **(C)** The detector is normally an electron multiplier. The mass analyzer and detector are always within the high-vacuum region ***(10)***.

enzyme for proteomic studies. However, cyanogen bromide, which cleaves after methionine, and endoproteinases such as chymotrypsin (cleaves after large hydrophobic amino acids), Lys-C (cleaves after Lys), and Asp-N (cleaves before aspartate) are also used. A good cleavage method must be compatible with the subsequent mass spectrometric analyses, have close to 100% cleavage efficiency, and a high specificity. The limitation of possible cleavage methods has led to the development of engineered proteases with new specificity ***(3)***. Mass spectrometric analysis of peptide fragments can lead to the identification and partial sequencing of a protein and identification of modifications. Such strategies are referred to as bottom-up sequencing, and are the most common techniques used in MS. An alternative approach, referred to as top-down sequencing, is starting to emerge where intact proteins are fragmented directly in the mass spectrometer. Top-down sequencing will not be discussed further in this chapter.

All mass spectrometers consist of three main parts: an ion source, a mass analyzer, and a detector (*see* **Fig. 1**). Analyte ions are produced in the ion source. Several ionization methods exist, but the most commonly used methods in proteomics are electrospray ionization (ESI) and matrix-assisted laser desorption and ionization (MALDI). The ions that are produced in the ion source are then transferred to the mass analyzer where they are separated according to their mass-to-charge ratio (*m/z*). Ion sources can be combined with different mass analyzers giving mass spectrometers, such as MALDI-time-of-flight (TOF), ESI-ion trap (IT), and ESI-Fourier transform ion cyclotron resonance (FTICR). The physical entity measured by all mass analyzers is the *m/z* value of the ions. The output, which is recorded at the detector, is ion intensity at different *m/z* values. The result is visualized by an *m/z* vs intensity plot, or a mass spectrum.

It has been shown that traditional 2D-polyacrylamide gel electrophoresis (PAGE) can resolve up to 1000 protein spots in a single gel ***(4)***. This is an impressive number, but compared to the number of expressed genes in various organisms, which range typically from 5000 to 40,000, it is clear that it is not good enough. Recently, efforts have been made to optimize the standard 2D-PAGE technique by making larger 2D-PAGE ***(5)***. The technique uses multiple, narrow-range

isoelectric focusing gels to improve separation in the first dimension. In the second dimension, multiple long sodium dodecyl sulfate-PAGE gels of different polyacrylamide concentrations are used. The large 2D-PAGE was claimed to resolve more than 11,000 protein spots. In general, it is the low abundance and hydrophobic proteins that are difficult to identify by the 2D-PAGE-based method ***(6)***. In addition, proteins with extremes in isoelectric point and molecular mass will not be retained in the gel and 2D-PAGE has a low throughput of samples ***(7)***.

As an alternative to the 2D-GE approach, multidimensional protein identification technology (MudPIT) is a method that has proven to be very efficient for identification of proteins in complex mixtures. This type of approach is also referred to as shotgun proteomics. A study on *Saccharomyces cerevisiae* using this method identified 1484 proteins ***(8)***. However, many of the identified proteins in this original study were not highly confident.

The MudPIT technology has some problems when it comes to quantification and significance of the peptides identified. One way to quantify is to integrate the absorbance or ion counts of a peptide during the chromatographic step. This method requires high reproducibility when two samples are compared. Reproducibility is especially difficult to achieve if nano-liquid chromatography columns are used. This is because partial blocking of columns may occur to different extents during consecutive runs. The problem can be reduced by adjusting the backpressure to assure a continuous flow. A more precise method of quantification of peptides in liquid chromatography–tandem mass spectrometry (LC–MS/MS) experiments is to use stable isotope labeling. The principle is that two or more samples are labeled with different stable isotopes. The differential labeling can occur during the synthesis of the proteins in cultured cells (SILAC; *see* Chapters 8 and 18), by reacting residues with labels containing different stable isotopes (chemical labeling; *see* Chapter 12), or by enzyme-catalyzed incorporation of ^{18}O from ^{18}O water during proteolysis ***(9)***.

The MudPIT method gives an enormous amount of data that requires automatic processing (*see* Chapters 6 and 7). Automatic computer-based interpretation of data calls for high-quality statistical testing to evaluate the quality of the interpretation. In the traditional 2D-PAGE, several peptides from the same protein normally confirm the identification, whereas the MudPIT method often claims identification of proteins from two peptide sequence tags. This is problematic because the same tryptic peptides can occur in rather diverse protein sequences (*see* Chapter 6). Therefore, the MudPIT method requires that the significance of protein assignment must be more precisely evaluated than for the 2D-PAGE method.

Protein sequences can be predicted from the expressed sequence tags (*see* Chapter 5) and gene sequences. However, the transcriptome and proteome are highly dynamic over time in different cells, in contrast to the genome. Therefore,

the identification and quantification of proteins and mRNAs are still major challenges. The identification of mRNA and protein over time in different cells works well, whereas the quantification has proven difficult for both mRNA and protein. The proteome is further complicated by the fact that the proteins can be modified to different extents, which is also a dynamic process over time in different cells.

1.2. Sample Preparation for MS

A proteomic project starts by generating a protein extract from tissues or from a homogenous cell culture. It is an advantage to chemically modify the reactive cysteines at the earliest possible stage to prevent mixtures of different cysteine modifications. The reactive cysteines can, for example, become oxidized or react with nonpolymerized acrylamide during electrophoresis ***(10)***. A discussion of the advantages of different cysteine modifications is presented in Matthiesen et al. ***(11)***.

A protein for identification needs to be purified before the proteolytic cleavage and analyzed by MALDI-TOF MS. The preferred method for the separation of proteins is 2D-PAGE.

The proteolytic cleavage is most often done with trypsin. Trypsin is a serine protease that specifically cleaves at the carboxylic side of Lys and Arg residues if these are not followed by Pro (*see* **Note 1**). The abundance and distribution of Lys and Arg residues in proteins are such that trypsin digestion yields peptides of molecular weights that can be analyzed, for example, by MALDI-TOF MS. The specificity of trypsin is of extreme importance. Native trypsin is subject to autolysis, generating pseudotrypsin, which exhibits a broadened specificity including chymotrypsin-like activity ***(12)***. Additionally, trypsin is often contaminated with chymotrypsin. For these reasons, trypsin, which has reductively methylated Lys and has been treated with *N*-tosyl-L-phenyl chloromethyl ketone, a chymotrypsin inhibitor, should be used. Such trypsin preparations can be bought under the trade name “Sequencing Grade Trypsin” (Sigma) or “Trypsin Gold” (Promega). After digestion, it is advantageous to concentrate and remove buffer contaminants by reverse-phase (RP) microcolumns to increase signal-to-noise ratios and sensitivity before mass spectrometric analysis. However, small and hydrophilic peptides can be lost by using the RP microcolumns. This problem can be solved by combining the RP microcolumns with purification using graphite powder ***(13)***. Additionally, the graphite powder can be used to remove some types of undefined biopolymers.

Recently, new methods have been developed to enrich for peptides that have a specific modification. The enrichment is in many cases important because it lowers the ion suppression from nonmodified peptides and thereby increases the signal of peptides having the specific modification. For example, hydrophilic interaction LC is used for the enrichment of glycosylated peptides ***(14)***. Titanium oxide has proven effective for enrichment of phosphopeptides ***(15,16)***.

1.3. Ionization Methods

Formation of ions and transition into the gas phase are required before the molecular masses of a sample can be measured by the mass analyzer. The generation of intact gas-phase ions is in general more difficult for higher molecular mass molecules. Early ionization methods for peptides and proteins, like fast-atom bombardment and ^{252}Cf plasma desorption, were successful and can be given credit for directing the attention of the biochemists toward MS ***(17)***. The breakthrough in MS of proteins and peptides came with the introduction of ESI–MS ***(18)*** and MALDI MS ***(19)***. The advantage of these soft ionization techniques is that intact gas-phase ions are efficiently created from large biomolecules with minimum fragmentation. MALDI and ESI are further discussed in the next sections. **Subheadings 1.4.–1.8.** discuss a common instrument setup using MALDI for ionization and a TOF mass analyzer that comprises the cheapest instrument setup. **Subheadings 1.9.–1.11.** discuss instrument setups using ESI ionization with either an ion trap or a Q-TOF mass analyzer.

1.4. MALDI

MALDI is an improvement of the laser desorption ionization (LDI) technique. In LDI, a soluble analyte is air-dried on a metal surface and the ionization is achieved by irradiating with an ultraviolet laser. The disadvantage of LDI is that, in general, it has low sensitivity, the ionization method causes ion fragmentation, and the signal is very dependant on the ultraviolet-absorbing characteristics of the analyte ***(20)***. This is solved with MALDI by decoupling the energy needed for desorption and ionization of the analyte. In MALDI, the analyte is mixed with a compound, the matrix, which absorbs the energy from the laser. The sample is cocrystallized with an excess amount of the matrix. A variety of matrices, small aromatic acids, can be used. The aromatic group absorbs at the wavelength of the laser light, while the acid supports the ionization of the analyte. Irradiation with a short-pulsed laser, often a 337-nm N_2 laser, causes mainly ionization of the matrix followed by energy and proton transfer to the analyte ***(21)***. The MALDI technique is good for ionizing peptides and proteins. Contaminants frequently encountered in protein and peptide samples, such as salts, urea, glycerol, and Tween-20, which are normally ionization suppressing, can be present at low concentration without a major effect on the ionization. It is believed that these compounds are excluded from the matrix and peptide/protein crystal ***(10)***. However, the MALDI technique is sensitive to low concentrations of sodium dodecyl sulfate ***(22)***.

The matrix preparation has a major effect on the quality of MALDI-MS spectra. Several different matrix–sample preparations have been developed. The earliest and most frequently used technique is the dried droplet method

(19,23). An improvement in this method was made by applying a pure matrix surface by fast evaporation before applying the matrix–analyte mixture *(24)*. Fast evaporation can be achieved by dissolving the matrix in volatile solvents like acetone. The method gives a more homogenous layer of small crystals. Nonuniform crystals formed under slow evaporation will give lower resolution, low correlation between intensity and analyte concentration, and lower reproducibility. A number of other preparation methods have been studied by Kussmann et al. *(25)* using a variety of matrices: α-cyano-4-hydroxy cinnamic acid (HCCA), sinapinic acid (SA), 2,5-dihydroxybenzoic acid (DHB), and 2,4,6-trihydroxyacetophenone (THAP). Some of the matrix preparation methods allow cleanup from buffer contaminants by a short wash with ice-cold 0.1% trifluoroacetic acid. There is no universal matrix preparation procedure that gives good results for all peptides and proteins. Only general guidelines have been found, such as the requirement of the sample–matrix mixture to be adjusted to pH less than 2.0 for optimal signal-to-noise ratio. Therefore, the previously mentioned studies pointed to the conclusion that several combinations should be tested for each sample and that optimization in some cases will be necessary.

In general, when MALDI ionization of tryptic peptides is used, protein sequence coverages of 30–40% are obtained. The limited sequence coverage can be explained by the competition for the protons in the matrix plum and limited mass range of the mass spectrometer. The more basic Arg-containing peptides have higher affinity for the protons than the Lys-containing peptide. The higher proton affinity and the higher stability of the Arg-containing peptides are therefore more frequently observed in MALDI-MS spectra. In addition, some peptides may not be able to co-crystallize with the used matrix.

1.5. The TOF Mass Analyzer

In TOF–MS a population of ions, for example derived by MALDI, is accelerated by an electrical potential as shown in **Fig. 2**. After acceleration, the ions pass through a field-free region where each ion is traveling with a speed characteristic of their *m/z* value. At the end of the field-free region a detector measures the TOF. The recorded TOF spectrum is a sum of the following times: TOF = $t_a + t_D + t_d$, where t_a is the flight time in the acceleration region, t_D is the flight time in the field-free region, and t_d is the detection time. Because the acceleration region is much smaller than the field-free region, the flight time can be approximated by the drift time, t_D:

$$t_D = D\sqrt{\frac{m}{2zeV_{ac}}} \tag{1}$$

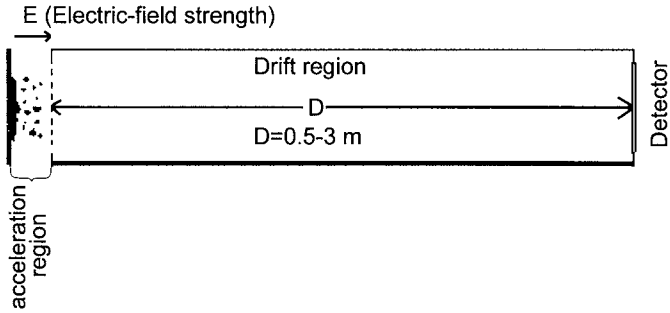


Fig. 2. Schematic representation of a linear time-of-flight mass spectrometer.

where D is the drift distance, z is the number of charges of the ion, e is the elementary charge, V_{ac} is the acceleration voltage, and m the mass of the ion. Therefore, t_D for an ion will be proportional to $(m/z)^{1/2}$ ***(26)***.

The mass resolution for TOF analyzers is reported as $m/\Delta m$ (Δm is the width of the peak at half maximal height). The ion production time, initial velocity distribution, and the ion extraction time all contribute to reduced resolution. The initial kinetic energy differences can be compensated by adding an ion reflector (also called ion mirror) to the linear TOF instrument. The reflector is an electric field that returns the ions in an opposite direction at an angle to the incoming ions at the end of the drift region. The more energetic ions will penetrate more deeply into the reflecting field and, therefore, have a longer flight path than ions with the same mass but less kinetic energy. The resolution of MALDI-TOF MS can also be considerably improved by using a delayed ion extraction technique ***(27)***. In this system, ions formed by MALDI are produced in a weak electric field, and after a predetermined time delay, are extracted by a high-voltage pulse.

1.6. The MALDI-TOF MS Spectrum

In a TOF spectrum it is the TOF of ions that is recorded. However, the spectrum is normally reported as *m/z* vs relative intensity. **Figure 3** shows the spectrum of tryptic peptides of bovine serum albumin analyzed on a Bruker Reflex III mass spectrometer using HCCA as matrix. The peak annotation was done automatically using a cut-off for a signal-to-noise ratio of 4.

Peptide ions generated by MALDI generally occur in a low-charge state, and normally only the singly charged state of a peptide is observed ***(10)***. For proteins it is common to observe higher charge states in addition to the singly charged.

Often some mass peaks in a spectrum cannot be matched with a theoretical mass of a tryptic peptide from a single protein. Typical reasons are tryptic miscleavage, modifications of tryptic peptides, tryptic peptides from other proteins,

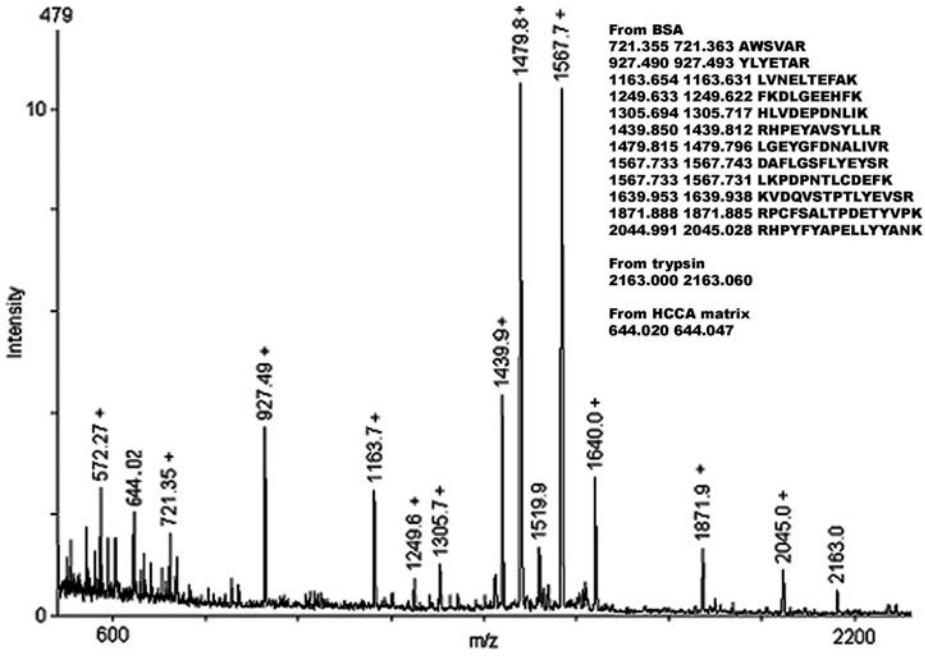


Fig. 3. Typical MALDI-TOF MS spectrum. (+) Indicates mass peaks which matched the masses of theoretical tryptic peptides from bovine serum albumin (BSA). The data analysis was done in VEMSmaldi v2.0. The first column in the upper right corner shows the experimental masses, the second column the theoretical tryptic peptide masses, and the third column shows the tryptic peptide sequences from BSA.

ions from matrix clusters, and fragmentation of ions during MALDI-TOF MS. The observed miscleavages are mainly because of pseudotrypsin, which is generated by autolysis of trypsin. Pseudotrypsin has a broader specificity than trypsin (*see* **Subheading 1.2.**).

The modifications of the tryptic peptides have different origins, in vivo post-translational modification, uncontrolled chemical modifications during sample preparation, and intentional modifications such as cysteine and cystine modifications.

The most frequently observed contaminating tryptic peptides come from trypsin (**Fig. 3**) and human keratins from hair and skin. Proteins purified by chromatography or 2D-electrophoresis are not always 100% free from other proteins in the original sample. For samples with high concentration of peptides and low salt, ions from matrix clusters are normally not observed in the mass region of interesting tryptic peptides (*m/z* above 500). For dilute samples, and samples containing a high concentration of salt, mass peaks from matrix clusters might be observed in the mass region up to 2000 Da for HCCA ***(28)***. In general, peaks

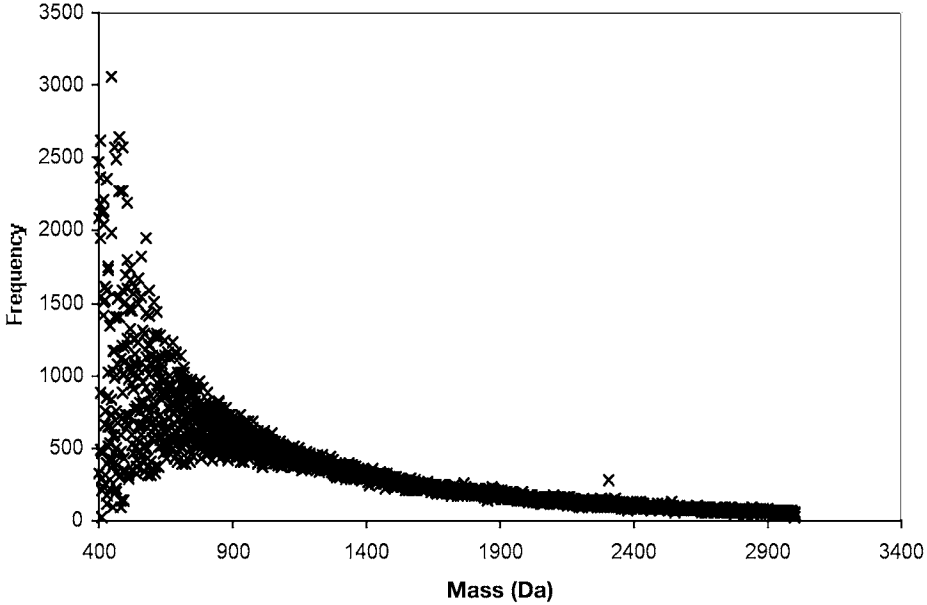


Fig. 4. Distribution of the theoretical masses between 400 and 3000 Da of tryptic peptides for all *Arabidopsis thaliana* proteins. The mass scale was divided in 1-Da intervals, and the numbers of tryptic peptides within each interval were counted. The distribution was made in VEMS v3.0.

of matrix clusters can be observed up to *m/z* 1000. Erroneous protein matches can be the result if these masses are not removed prior to the peptide mass fingerprinting (PMF) search. One solution to this problem could be to remove all mass peaks below *m/z* 800, as some shorter tryptic peptides are less unique, as shown in **Fig. 4**.

However, the low mass region still contains valuable information and some shorter peptides are actually more unique than longer peptides. A better option would be to try to identify the mass peaks from the matrix clusters. It has been reported that the mass of most of the matrix peaks can be calculated using the following equations ***(28)***:

$$M_{Cluster} = nM - xH + yK + zNa \tag{2}$$

where

$$y + z = x + 1 \tag{3}$$

and

$$y + z \leq n + 1 \tag{4}$$

Table 1
Masses of the Elements Used in the Calculation of Ion Adducts[a]

	Mass (Da)
H	1.007825032
Na	22.98976967
K	38.9637069
HCCA	189.0425931

[a]The elemental masses are from http://www.ionsource.com, and the mass for HCCA was calculated using the elemental masses.

n, *y*, *z*, and *x* are integer values. $M_{cluster}$ is the observed mass of the ionized matrix cluster; *M*, *H*, *K*, and *Na* stand for the masses of HCCA, hydrogen, potassium, and sodium, respectively. For the DHB matrix a similar equation where also the loss of water is taken into account, have been established ***(28)***. **Table 1** gives the masses used to calculate mass of the different matrix clusters.

The mass peak 644.02 *m/z* in the MALDI-TOF spectrum in **Fig. 3** was assigned as a matrix peak ($n = 3$, $x = 0$, $y = 2$, and $z = 0$) using **Eqs. 2–4**. Thereafter, only one annotated mass peak of **Fig. 3** has not been assigned. Zooming in on the low mass region some additional low-intensity peaks were annotated (**Fig. 5**).

It is not recommended to annotate low-intensity peaks prior to the first database search, as these include a higher proportion of peaks that cannot be interpreted and, therefore, lead to erroneous protein identification. Further annotation of the low-intensity mass peaks identified one additional mass peak matching bovine serum albumin, and four matching matrix clusters. The result of the final data analysis is summarized in **Table 2**. The identification of trypsin and matrix cluster peaks gives the opportunity to use these mass peaks for internal calibration (*see* **Appendix B;** ***29***).

Matrix clusters might be suppressed either by purifying samples on RP microcolumns, by using cetrimonium bromide in the matrix solution ***(30)***, or by recrystallizing the matrix in ethanol ***(28)***. However, some peptides might get lost on RP columns and cetrimonium bromide lowers the intensity of sample ions although it is very efficient in suppressing matrix clusters.

Fragmentation of sample ions may occur from collision with matrix molecules during ionization. However, in MALDI only minimal fragmentation occurs ***(22)***.

1.7. Annotation of MALDI-TOF MS

In order to annotate MS spectra it is important to understand the nature of the isotopic distribution. Biological molecules are mainly composed of carbon (C),

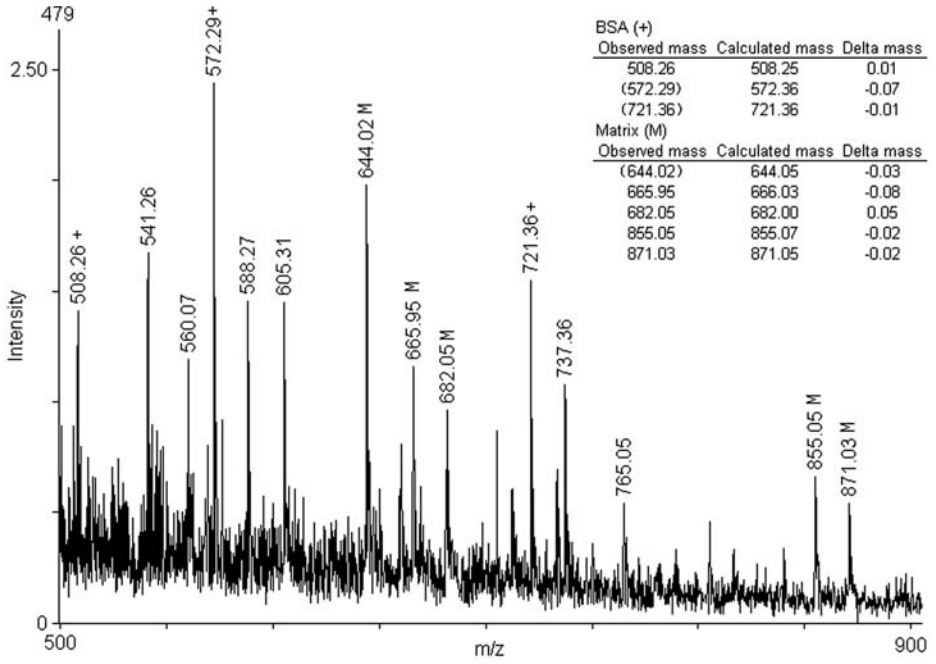


Fig. 5. Zooming in on the low mass region of the MALDI-TOF spectrum shown in **Fig. 3**. Peptide mass peaks from bovine serum albumin are indicated by (+), and (M) indicates mass peaks that correspond to matrix clusters. The observed mass peaks in parentheses were already annotated in the first annotation round. The annotation in the second round was done manually. The data analysis was made in VEMSmaldi v2.0.

hydrogen (H), nitrogen (N), oxygen (O), and sulfur (S). Some biological molecules also bind metal ions and these include proteins and DNA. Natural isotopes of these elements occur at almost constant relative abundance (**Table 3**). A more extensive list of biological relevant isotopes can be found at http://www.ionsource.com/Card/Mass/mass.htm.

From the values given in **Table 3** one can calculate relative isotopic abundance of different biological molecules. The relative abundance of the monoisotopic mass of a molecule with the composition $C_xH_yN_zO_vS_w$ can be calculated using the following expression ***(32)***.

$$P_M = P_C^x * P_H^y * P_N^z * P_O^v * P_S^w \quad (5)$$

P_M is the relative abundance of the monoisotopic peak for the molecule. P_C, P_H, P_N, P_O, and P_S are the abundance of the monoisotopic masses of the C, H, N, O, and S elements, and x, y, z, v, w are positive integer values. The expression is simply the probability that all the elements in the molecule have the

Table 2
Interpretation of Low-Intensity Peaks in the Low Mass Region of a MALDI Spectrum of a Bovine Serum Albumin Tryptic Digest[a]

Observed mass (Da)	Calculated mass (Da)	Delta mass (Da)	Match	n	x	y	z
499.204			?				
508.257	508.252	0.005	BSA				
541.262			?				
560.071			?				
(572.291)	572.363	–0.072	BSA				
588.265			?				
605.314			?				
(644.022)	644.047	–0.025	matrix	3	1	2	0
665.954	666.029	–0.075	matrix	3	2	2	1
682.054	682.003	0.051	matrix	3	2	3	0
(721.355)	721.363	–0.008	BSA				
737.361			?				
765.047			?				
855.047	855.072	–0.025	matrix	4	2	2	1
871.026	871.046	–0.02	matrix	4	2	3	0

[a]The data analysis was done in the program VEMSmaldi. The observed mass peaks in parentheses were already annotated in the first annotation round.

Table 3
Masses and Abundance of Biologically Relevant Isotopes (*28*,*31*)[a]

Isotope	**A**	**%**	Isotope	**A + 1**	**%**
^{12}C	12	98.93(8)	**^{13}C**	13.0033548378(1)	1.07(8)
^{1}H	1.0078250321(4)	99.9885(7)	**^{2}H**	2.0141017780(4)	0.0115(7)
^{14}N	14.0030740052(9)	99.632(7)	**^{15}N**	15.0001088984(9)	0.368(7)
^{16}O	15.9949146221(15)	99.757(2)	**^{17}O**	16.99913150(2)	0.038(1)
^{32}S	31.97207069(12)	94.93(3)	**^{33}S**	32.97145850(1)	0.76(2)
Isotope	**A + 2**	**%**	Isotope	**A + 4**	**%**
^{14}C	14.003241988(4)	–	–	–	–
^{3}H	3.0160492675(11)	–	–	–	–
^{18}O	17.9991604(9)	0.205(1)	–	–	–
^{34}S	33.96786683(11)	4.29(3)	**^{36}S**	35.96708088(3)	0.02(1)

[a]Uncertain digits are shown in parenthesis.

monoisotopic mass. A similar expression can be made for the monoisotopic mass plus one,

$$P_{M+1} = \binom{x}{1}P_C^{x-1}P_{C+1}P_H^y P_N^z P_O^v P_S^w + P_C^x\binom{y}{1}P_H^{y-1}P_{H+1}P_N^z P_O^v P_S^w + \ldots + P_C^x P_H^y P_N^z P_O^v \binom{w}{1}P_S^{w-1}P_{S+1} \quad (6)$$

where $P_{C+1,}$ $P_{H+1,}$ $P_{N+1,}$ P_{O+1}, and P_{S+1} are the abundance of the monoisotopic mass plus approx 1 Da of the elements. Again, the expression is the probability that one atom in the molecule is the monoisotopic mass plus one. It is important to note that this way of calculating the isotopic distribution is an approximation, which works well when comparing with observed isotopic distribution from mass spectrometers that are unable to resolve the different element's contribution to the M+1 ion. In a similar way one can make an expression for the abundance of the M+2 isotope peak. How to perform such calculations in practice is described in Chapter 2. Zooming in on the mass peaks 927,49 and 2045,00 *m/z* in **Fig. 3** reveals the isotopic distributions in **Fig. 6**.

For PMF searches it is obligatory to annotate the monoisotopic mass peak for each isotopic distribution because a list of monoisotopic masses can be submitted directly to most search engines. If the isotopic peaks are not resolved, which is sometimes observed for peptides above 2500 Da, then the average mass can be annotated and used for the search. The annotation of peptide masses is straight forward because often there is only single-charged ions present in the MALDI-TOF MS spectrum. The charge states can be determined from the distance between two peaks in the isotopic mass distribution of an ion (*see* **Fig. 6**). Isotopic peaks are separated by approx 1 Da for single-charged ions, approx 0.5 Da for double charged, and so on. In general, the charge state can be calculated by:

$$\Delta m = \frac{m}{z} - \frac{m+1}{z} = \frac{1}{z} \Leftrightarrow z = \frac{1}{\Delta m} \quad (7)$$

where Δm is the mass difference between the isotopic peaks.

1.8. The Future for MALDI-TOF MS

MALDI-TOF MS is playing a major role in proteomics. The main advantage with MALDI ionization is high sensitivity. MALDI-TOF MS has been used extensively together with 2D-electrophoresis ***(33)***. One of the biggest problems with the MALDI-TOF MS for identification of proteins is the difficulty in obtaining highly significant search result. This is owing to the fact that not all the expected tryptic peptides show up in the experimental MS spectrum. It is additionally complicated by unknown mass peaks. The PMF by MALDI-TOF MS is limited to the analysis of purified proteins. Obtaining statistically significant PMF results for protein mixtures have proven difficult. However, there have been reports on strategies to identify protein in mixtures containing few proteins by

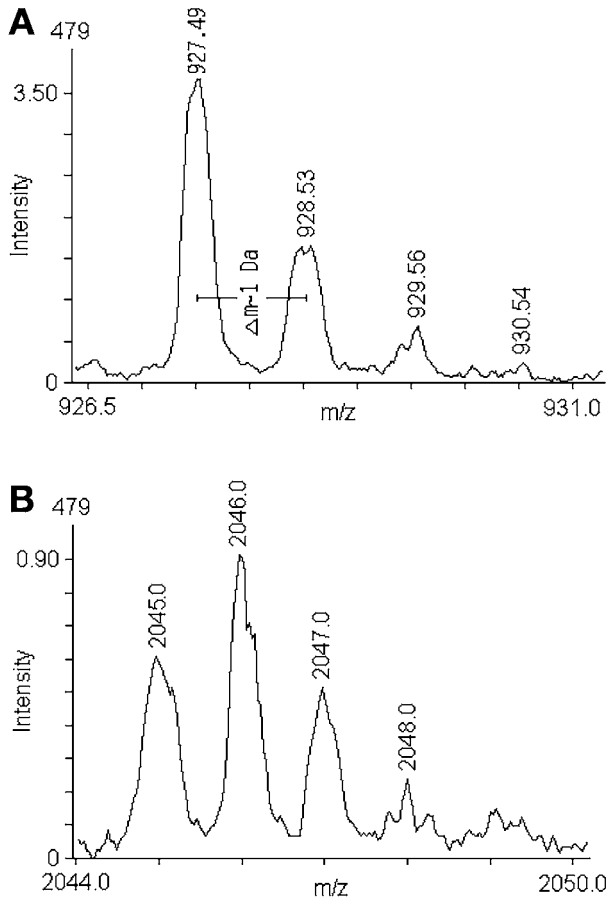


Fig. 6. Zoom in on two annotated masses from the spectrum in **Fig. 3**. **(A)** At low masses the monoisotopic peak is the most intense peak in the isotopic cluster. **(B)** At higher masses it is no longer the monoisotopic peak that is the most intense peak in the isotopic distribution.

PMF searches ***(34,35)***. The significance of the PMF results can be improved by using more cleavage methods in addition to trypsin, and the search algorithms can be improved by using both average and monoisotopic mass for PMF searches.

By using MALDI-post source decay (PSD) it is possible to obtain sequence information with the MALDI method ***(36)***. However, the PSD fragmentation is rather complex. With the introduction of a collision cell in a MALDI-TOF-TOF MS it is now possible to obtain low- and high-energy collision-induced dissociation (CID) fragmentation with the MALDI method ***(9)***, which can provide more significant MALDI data and extend the MALDI method to more complex mixtures containing tryptic peptides from several proteins.

Another problem related to the MALDI method is the heterogeneity in matrix–sample crystal formation, which makes it difficult to quantify and reproduce results. Using fast evaporation methods when applying the matrix seem to give more homogenous matrix–sample crystals. However, it is still often necessary to search for a good sample spot within a target.

1.9. ESI

Electrospray has found many applications, such as spray painting, fuel atomization in combustion systems, and crop spraying ***(37)***.

In ESI–MS, the ions are formed at atmospheric pressure followed by droplet evaporation. A solution containing peptides or proteins is passed through a fine needle at high potential in order to generate ions ***(10)***. The potential difference between the capillary and the counter electrode, located 0.3–2.0 cm away from the capillary, is typically 3–6 kV ***(38)***.

The electric potential is responsible for charge accumulation on the liquid surface, leading to the production of charged droplets from the liquid cone at the capillary tip. The initial drops formed by ESI range from few micrometers to 60 μm in diameter ***(39)***. These drops shrink by evaporation, which results in increased charge density. The increased charge density creates a Coulomb repulsion force that eventually will exceed the surface tension causing drop explosion (Coulomb explosion) into smaller drops. This process continues until the drops are small enough to desorb analyte ions into the gas phase. A good sample spray is dependant on flow rate, liquid conductivity, and surface tension. The addition of organic solvent to aqueous analyte solution results in lower surface tension, heat capacity, and dielectric constant, all of which facilitate formation of fine droplets by Coulomb explosion.

To assist in the formation of droplets a nebulizer gas can be used. The nebulizing nitrogen flows through the outside of the needle. As the liquid exits, the nebulizer gas helps break the liquid into droplets. In addition, a N_2 drying gas is applied on the entrance of the dielectric capillary to help droplet evaporation.

ESI has proven effective in producing gas-phase ions of proteins and peptides ***(40)***. In general, the longer the polypeptide chain the higher the charge state as more groups can ionize. Therefore, multiple charge states are observed for proteins and peptides. The distribution of these charge states will depend on the equilibrium between the different protein folds in solution prior to the electrospraying and events in the gas phase ***(41)***. The net charge in solution will depend on intrinsic polypeptide properties as well as extrinsic factors. The intrinsic factors include the number, distribution, and pK_as of ionizable amino acid residues, which depend on the initial three-dimensional conformation. The extrinsic factors are solvent composition, pH, ionic strength, and temperature ***(41)***.

MS/MS spectra of polypeptides are most often done in positive ionization mode. The negative ionization mode can be applied in a scanning mode to detect sulphated and phosphorylated peptides ***(42)***.

For analysis in positive mode the sample solution is often acidified, which also leads to relatively high amounts of anions in addition to the protons. The distribution of $(M+nH)^{n+}$ shifts toward larger *n* and decreases the *m/z* values of the ions formed when the pH is lowered ***(43)***. However, the different types of anions have been observed to have different effects on the net average charge of peptides and proteins ions in the spectra. The propensity for neutralization follows the order: $CCl_3COO^- > CF_3COO^- > CH_3COO^- \sim Cl^-$ ***(41)***. The charge reduction is proposed to occur in two steps. The first occurs in the solution where the anion can pair with a basic group on the peptide or protein. The second occurs in the gas phase during desolvation where the protons dissociate from the peptide to form the neutral acid and the peptide in a reduced charge state ***(41)***. It has been found that acetic acid and formic acid produce better ESI results than trifluoroacetic acid. Additionally, detergents should be avoided. Detergents can lead to signal reduction and can add complexity to the mass spectra because of their polymeric nature ***(10)***.

Cations also affect the quality of ESI mass spectra. Cation impurities from buffers can cause a reduction of analyte signal owing to spreading of the signal over multiple *m/z* values; it also adds complexity to the spectra. The presence of alkali metals in ESI mass spectra indicates that desalting of the sample is required. In some cases, analysis of metalloproteins can lead to determination of metal-binding constants ***(44)***.

ESI is performed at atmospheric pressure, which allows online coupling of high-performance liquid chromatography and capillary electrophoresis to mass spectrometers ***(45,46)***. ESI can be coupled to mass analyzers, such as triple quadrupoles, ion traps, and quadrupole-TOF (Q-TOF). A powerful and common setup is RP-LC coupled to ESI–MS/MS. Increasingly nano- and microbore RP-columns are being coupled to electrospray-MS/MS. The typical flow rates in commonly used LC systems are 100–300 nL/min and 1–100 µL/min for the nano- and micro-LC for ESI systems, respectively. However, values of 10–15 nL/min and 1–10 µL/min have been obtained ***(35,44)***. The low flow rate, high ionization efficiency, and high absolute sensitivity make nano-ESI–MS/MS suitable for identification or sequencing of gel-isolated proteins available in subpicomole amounts ***(48)***.

1.10. Ion Trap Mass Analyzer

An LC-ion trap system produces a continuous source of ions by ESI that are guided into an ion trap mass analyzer by a combination of electrostatic lenses and a radio frequency octapole ion guide. First, the incoming ions are focused

toward the center of the ion trap, which is composed of three electrodes, a ring electrode, and two end caps. This is accomplished by slowing down the incoming ions with helium (He) gas (typically 1–5 mbar He) and trapping them in a three-dimensional quadrupole field ***(49,50)***. The quadrupole field, which establishes a parabolic potential induces an oscillatory harmonic motion of the ions at a frequency known as the secular frequency. The oscillation can be described by solutions to the Mathieu equations ***(51)***.

The range of ion masses that can be trapped in the ion trap simultaneously is limited in the low *m/z* range, and also in the high *m/z* range in practice. The stored ions can be ejected according to their *m/z* value by applying a dipolar field of frequency that is proportional to the secular frequency of an ion. By detecting the ejected ions at different dipolar field frequency, a MS spectrum can be obtained.

Ion traps have the ability to store a precursor ion of interest and eject all other ions simultaneously. The stored ion can be fragmented to produce sequence information. This is done by increasing the energy of the isolated precursor ion by excitation with the dipolar field. The amplitude for excitation is less than that used for ejection. The increased energy of the precursor ion leads to harder and more frequent collisions with the He gas causing fragmentation of the precursor ion ***(49)***. The fragmentation is termed CID. The product ions will have a different secular frequency than the precursor ion, preventing further fragmentation. The fragment ions can be ejected at different dipolar field frequencies and detected producing the MS/MS spectrum.

1.11. Q-TOF Mass Analyzer

The Q-TOF mass analyzers are often coupled to an ESI ion source. A typical Q-TOF configuration consists of three quadrupoles, Q1, Q2, and Q3, followed by a reflectron TOF mass analyzer ***(51)***. In some instruments some of the quadrupoles are replaced by hexapoles. However, it is the same principle. Q1 is used as an ion guide and for collisional cooling of the ions entering the instrument. For separating ions for MS spectra, Q1 and Q2 serve as transmission elements, whereas the reflector TOF separates ions according to the *m/z* values.

For recording MS/MS, Q1 is operated in mass filter mode to transmit only precursor ions of interest. The width of the mass window of ions allowed to pass Q1 determines the range of the isotopic cluster ***(51)***. The ion is then accelerated to Q2, where it undergoes CID. The collision gas is usually He or argon. The fragment ions are collisionally cooled, and focused in Q3. Finally, their *m/z* values are separated in the reflector TOF.

1.12. Comparison of Q-TOF and Ion Trap Instruments

The different methods of fragmentation in Q-TOF and ion trap mass analyzers results in differences in the MS/MS spectra. In the Q-TOF the ions are fragmented

by accelerating them and passing them through a collision cell filled with gas. The fragments from the precursor ion can still contain enough kinetic energy so further fragmentation occurs on collision with the collision gas. This means that only the most stable ion fragments will be observed in an MS/MS spectrum from a Q-TOF mass analyzer.

In the ion trap, fragmentation occurs by increasing the secular frequency of the precursor ion making the precursor ion collisions with the collision gas stronger and, thereby, causing fragmentation. The fragment ions will have a different secular frequency than the precursor ion and will therefore not be further excited by the applied dipolar field frequency, which prevents further fragmentation. This difference means that b-, a-, and y-ions frequently can be observed in data from an ion trap, whereas Q-TOF data mainly contains y-ions (ion types are explained in **Subheading 1.13.**). This means that Q-TOF data is easier to interpret as only one ion series is mainly present. On the other hand, the found sequence tag in the ion trap would be more significant because more ions would be present to confirm the derived sequence tag. However, the higher mass accuracy and resolution of Q-TOF data is likely to compensate for the lower significance caused by the low frequency of a- and b-ions.

In Q-TOF, MS spectra immonium ions can be observed. The *m/z* value for these ions is normally below the mass cutoff used for ion traps and, therefore, is not stable in the ion trap (the mass cutoff is normally one-third of the *m/z* value for the precursor ion ***[49]***). The immonium ions can be used as an indicator of the presence of the corresponding amino acid in the peptide sequence ***(52)***.

1.13. The MS/MS Spectrum

The MS/MS spectrum of peptides contains sequence information. To obtain MS/MS spectra one needs a tandem mass spectrometer. MS/MS can be one of the two types: tandem-in-space and tandem-in-time.

In tandem-in-space MS, the first mass analyzer separates and isolates the precursor ion of interest. The isolated ion is transmitted to the collision cell where it is fragmented. The fragments are transmitted and analyzed in the second mass analyzer (tandem in space). In tandem in time, the precursor ion selection, fragmentation, and the separation of the fragments all occur in one mass analyzer at different time-points. The precursor ions with the highest intensity and with a charge state of +2 or more are generally preferred because they give the best fragmentation spectra. The isolated precursor ion is then fragmented by CID, infrared multiphoton dissociation (IRMPD), electron capture dissociation (ECD), blackbody infrared dissociation (BIRD), surface-induced dissociation (SID), or electron transfer dissociation (ETD). In most cases, CID is used for fragmentation. The resulting spectrum is called an MS/MS spectrum. The precursor ion can have different charge states. Singly

charged tryptic peptide ions generally do not fragment as easily as tryptic peptide ions with higher charges.

A number of computer programs such as Probid ***(35)***, Mascot ***(53)***, Lutefisk ***(54)***, and Virtual Expert Mass Spectrometrist (VEMS) ***(55)*** exist for automatic interpretation of MS/MS spectra. However, all of these programs have limitations. To obtain an extended interpretation of data it is necessary to understand the structure of protonated peptides and their fragmentation pathways.

If a peptide sample is ionized by ESI in the positive mode, then positive peptide ions will be generated. The protons can attach to all the strongly basic sites of the peptides. Examples of these sites are the N-terminal amine, Lys, Arg, and histidine residues ***(52)***. Protons attached to the N-terminal site may move to any of the amide linkages, whereas the protons associated with Arg, Lys, and histidine are bound more strongly.

This model is termed the mobile proton model and explains many observed phenomena in MS/MS spectra of peptides ***(56)***. The precise location of the proton after transfer to the backbone is not fully established. However, the carbonyl oxygen of a peptide bond has been suggested to be the most likely candidate ***(56)***.

According to the mobile proton model, a peptide ionized by ESI is best viewed as a heterogeneous population with different location of the charges. The population of peptides is then focused into the collision cell where it is accelerated to induce fragmentation. The kinetic energy from the collisions with the collision gas is converted to vibrational energy in the peptide ions. The peptides then release the vibrational energy by fragmentation. For most peptides fragmentation can be described as a charge-directed cleavage ***(56,57)***. In charge-directed cleavages, fragmentation is guided by the site of the protonated peptide bonds. Charge-directed cleavage is a complicated reaction where the different chemical bonds along the peptide backbone are cleaved with different probability ***(57)***. The result is mainly b- and/or y-type sequence ions (*see* **Fig. 7**).

If the number of Lys, Arg, and histidine in a peptide equals the number of positive charges of the peptide then there are no mobile protons. These stronger bound protons can be mobilized if the kinetic energy of the peptides is increased in the collision cell. However, the increased energy might lead to charge remote fragmentation where no proton is involved in the fragmentation ***(56)***.

Normally, MS/MS experiments are made on tryptic peptides and the experimental settings are such that most of the tryptic peptides have charge states higher than one. The major low-energy CID pathway by proton-directed cleavages of doubly charged peptides gives mainly b- and y-ion fragments. The most favored reaction leads to the formation of singly charged b- and y-ions. Because the mobile proton can be located at different amide bonds, a whole series of b- and y-ions is generated for tryptic peptides with more than one charge. The

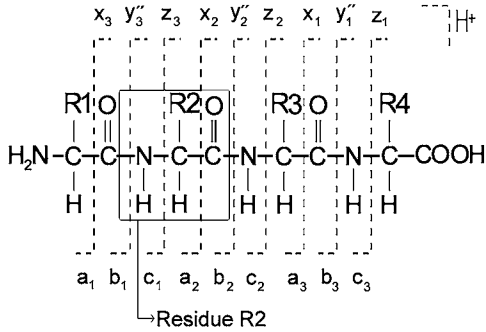


Fig. 7. The nomenclature used for peptide fragment ions. a-, b-, and c-ions are charged fragments containing the N-terminal part of the fragment and x, y, and z the C-terminal part ***(58)***. The box indicates the composition of residue R2. Residue mass for R2 is given by summing the masses of the elements of residue R2.

mass difference between two neighboring b_n and b_{n+1}-ion mass peaks corresponds to the residue mass of the most C-terminal residue of the b_{n+1}-ion (*see* **Fig. 7**). Similarly, the mass difference between two neighboring y_n and y_{n+1}-ion mass peaks corresponds to the residue mass of the most N-terminal residue of the y_{n+1}-ion (*see* **Fig. 7**). Calculating mass differences between mass peaks in a MS/MS spectrum and relating these mass differences with a list of residue masses can reveal the peptide sequence, or part of it, i.e., a sequence tag. In **Appendix C** a table of residue masses is provided.

From **Fig. 7** the following simple, but important, relation can be deduced:

$$m_{p2+} = m_{b_i} + m_{y_{n-i}} \qquad i = 1,2,...,n-1 \tag{8}$$

where m_{p2+} is the mass of the doubly charged parent ion, n is the length of the peptide, and the right side is the sum of the masses of any pair of complementary b- and y-ions. This equation can be used to verify that the complementary ion series is present in the experimental spectrum.

The obtained sequence tag can be validated by calculating a theoretical a-, b-, and y-ion series for the sequence tag and comparing these with the experimental spectrum. Consider a peptide of n amino acids $AA_1,\ldots,AA_n$ with masses $m(AA_j)$. The mass of the doubly charge peptide can be calculated as:

$$m_{p2+} = m(H_2O) + 2 * m(H) - 2 * m(e) + \sum_{j=1}^{n} m(AA_j) = 20.025118 + \sum_{j=1}^{n} m(AA_j) \tag{9}$$

This is the mass of a doubly charged peptide precursor ion and is sometimes observed in the MS/MS spectrum at m/z $m_{p2+}/2$. In practice it is not necessary to take the mass of the electron m(e) into account because the mass accuracy of

common MS/MS mass spectrometers is in the range 0.01 to 0.2 Da and a low number of missing electrons will only have an effect on the third decimal. The mass of the electron here is taken into account to be precise.

If the b_i-ion is composed of the amino acids $AA_1,\ldots, AA_i$; its mass is given by

$$m_{b_i} = m(H) - m(e) + \sum_{j=1}^{i} m(AA_j) = 1.007276 + \sum_{j=1}^{i} m(AA_j) \quad (10)$$

The expression is the sum of all the residue masses of residues in b_i plus the mass of a proton ***(59)***.

The masses of the y-ion series can be computed in a similar way ***(59)***:

$$m_{y_{n-i}} = m(OH) + 2 * m(H) - m(e) + \sum_{j=i+1}^{n} m(AA_j) = 19.017841 + \sum_{j=i+1}^{n} m(AA_j) \quad (11)$$

The masses in the a-ion series (*see* **Fig. 7**) can be calculated as

$$m_{a_i} = -m(CO) + m(H) - m(e) + \sum_{j=1}^{i} m(AA_j) = -26.987638 + \sum_{j=1}^{i} m(AA_j) \quad (12)$$

The frequency of the different fragmentation ions observed in ESI low-energy CID can be ordered as y>b>a for tryptic peptides with charge state higher than one, whereas the fragment ions c, x, and z cannot be observed to any significant extent ***(59)***. The b_1-ions are not observed in the MS/MS spectrum because of the lack of a carbonyl group to initiate the nucleophilic attack on the carbonyl carbon between the first two amino acids; without this carbonyl the cyclic intermediate cannot be formed ***(60)***.

Fragmentation of a doubly charged peptide with one proton on a basic amino acid and one mobile proton will lead to formation of a b_i- and y_{n-i}-ion. The b_i-ion would still have a mobile proton and can therefore undergo further fragmentation in a quadrupole collision cell, and to a much lesser extent in an ion trap (*see* **Subheading 1.12.**). The b-ions preferentially fragment to smaller b-ions than to a-ions ***(60)***. Because the b_2-ion is the last b-ion fragment, the corresponding mass peaks often have high intensity in MS/MS spectra generated in a quadrupole collision cell (*see* **Fig. 8**).

The b_2-ions can fragment further to a_2-ions. This means that high-intensity a_2-ion mass peaks can often be observed (*see* **Fig. 8**). The presence of a_2- and b_2-ion mass peaks with high intensity and with a mass difference of approx 28 Da is very useful in manual interpretation of MS/MS spectra. The mass of the observed b_2-ion can be compared with the mass of b-ions of all combinations of two amino acids. This often gives a limited set of possibilities. The assignment of the precise ordering of the two amino acids depends on observing the

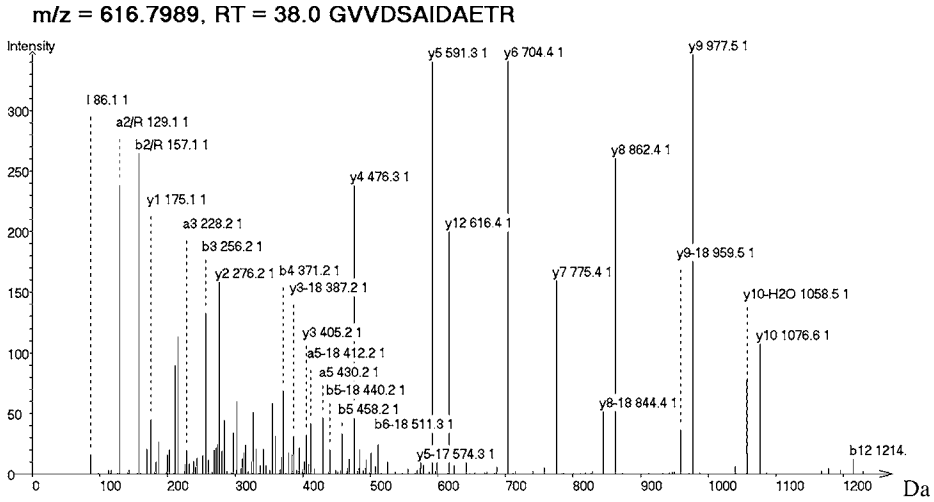


Fig. 8. A tandem mass spectrometry (MS/MS) spectrum of the tryptic peptide GVVDSAIDATER. The MS/MS spectrum was automatically annotated using VEMS v3.0. The annotation is given as ion type, mass (Da), and original charge state of the ion before deisotoping and decharging.

y_{n-1}-ion. The y_{n-1}-ion is often not observed, so the ordering is only completed after finding a database match.

The C-terminal amino acid of a tryptic peptide is either Lys or Arg, except for the last peptide of a protein and for miscleaved peptide bonds. The y_1 can often be observed in Q-TOF data with low intensity at *m/z* 147.11 for Lys and 175.12 for Arg. In addition, the immonium ions of the C-terminal amino acid and fragments of the immonium ions can often be observed. Identification of the C-terminal amino acid decreases the number of possible database matches and increases the search speed.

In the low mass region one also finds the immonium ions, $H_2N^+ = CHR$ (*see* **Appendix D**). The *m/z* of unmodified immonium ions ranges from *m/z* 30 (G) to *m/z* 159 (W). The immonium ions can be used as an indication of the presence of certain amino acids. In practice immonium ions can be observed for peptides containing L/I, V, F, Y, H, P, and W, whereas S, T, R, and K are ineffective sources ***(61)***. If L/I, V, F, Y, H, P, and W are N-terminal residues, the corresponding immonium ion can be abundant ***(62)***. However, immonium ions from internal residues are also quite frequently observed.

Other fragmentation ions apart from a-, b-, and y-ions exist and they complicate MS/MS spectra further. The y_ib_j-ions, also called the internal fragment ions, are one example. The y_ib_j-ions are generated from b- or y-ion fragments ***(62)***.

Low-energy CID of protonated peptides also give rise to fragmentation products, which arise from small neutral losses from b-, y-, and y_ib_j-ion products (**Appendix D**). The a-ions result from the loss of CO from b-ions, and are seen as b minus 28 Da ***(62)***. The loss of NH_3 can occur from b-, y-, and y_ib_j-ions containing the residues N, Q, K, and R. Peptide ions containing the residues S, T, D, and E can lose H_2O ***(52)***. The neutral losses can also be used for assigning MS/MS spectra of peptides containing modified residues. An example is the neutral loss of H_3PO_4 from phosphoserine or phosphothreonine giving anhydro derivatives.

1.14. Fragmentation Methods

Fragmentations observed in MS are divided into three main categories ***(63)***: (1) unstable ion fragments, formed in the ion source, with dissociation rate constants of $k_{diss} > 10^6$/s, (2) metastable ion fragments, formed between the ion source and the detector, with $10^6/s > k_{diss} > 10^5/s$, and (3) stable ions which remain intact during their time in the mass spectrometer and have $k_{diss} < 10^5/s$. The list of techniques for precursor ion fragmentation includes collision-activated dissociation/decomposition (CAD)/CID, PSD, ECD, IRMPD, BIRD, SID, or ETD. IRMPD ***(64)***, BIRD ***(65)***, and SID ***(66)*** will not be discussed further because they are mainly used for top down proteomics. The most frequently used method for precursor ion fragmentation is CAD/CID ***(67)***. However, the other methods have been found useful in specific areas. In the following section, the most common fragmentation methods currently used in bottom up proteomics are described in more detail.

1.14.1. CAD/CID

CID is the most frequently used fragmentation method. It works by inducing collisions between the precursor ion and inert neutral gas molecules. This leads to increased internal energy followed by decomposition of the precursor ion. The activation of precursor ions is separated in time from the dissociation process. The activation time is many times faster than the dissociation, which explains why CID can be modeled as a unimolecular dissociation process and explained by Rice–Ramsperger–Kassel theories ***(68)*** or quasi-equilibrium theory modeling.

$$[M+nH]^{n+} \xrightarrow[\text{increase by rf pulse}]{\text{Kinetic energy}} [M+nH]^{n+}* \xrightarrow[\text{inert gas}]{\text{Collision with}}$$

$$[M+nH]^{n+}** \xrightarrow[\text{fragmentation}]{\text{Unimolecular}} a,\ b,\ \text{and}\ y \quad \begin{cases} * \text{ increased kinetic energy} \\ ** \text{ increased internal energy} \end{cases}$$

Both low- and high-energy collisions are used for peptide fragmentation. Low-energy collisions are used in quadrupole and ion trap instruments and they result mainly in a-, b-, y-ions, immonium and ions from neutral loss of ammonia, and

water from the a-, b-, y-ions. High-energy CID is mainly used in sector and TOF-TOF instruments and yields similar fragment spectra of peptides as low-energy CID. The main difference is that d-, v-, and w-ions corresponding to amino acid side chain cleavage are observed in addition to more intense immonium ions. CAD and CID can be used interchangeably. However, CID is used more frequently in the literature (http://www.msterms.com/wiki/index.php?title = CAD_vs._CID). Many variants of the CID method exist. They can be divided into two main groups, on-resonance and sustained off-resonance irradiation (SORI). SORI-CID is the most widely used technique in FTICR MS ***(69)***. On-resonance excitation works by applying an on-resonance radio frequency (RF) pulse to increase the kinetic energy of the ions. The increased kinetic energy increases the collision energies and the number of collisions with the collision gas. However, the collision energies drop for each collision that the ion experiences. It is therefore an advantage to activate the ions several times, as is the case in multiple excitation collisional activation ***(70)***. SORI also uses multiple excitations by applying RF pulses slightly above and below the resonant frequency which causes the ions' kinetic energy to oscillate. The advantage of SORI compared with on-resonance CID is that the distribution of the number of collisions of the ions is broader for SORI, meaning that a broader range of fragment ions are generated ***(69)***.

1.14.2. ECD

ECD has turned out to have great usability for studying peptides with modified amino acid residues ***(71)***. In ECD, trapped multiple-charged ions are irradiated with a low-energy electron beam (<0.2 eV) generated by a heated filament electron gun ***(72)***. The electron capture by the positively charged peptides leads to intensive backbone fragmentation yielding mainly c and $z^{\bullet}$ fragments.

$$\left[M + nH\right]^{n+} \xrightarrow{e} \left[M + nH\right]^{(n-1)+\bullet}$$
$$\xrightarrow[\text{cleavage}]{\text{Local ECD}} c,\ z^{\bullet}\ \textit{fragments} \qquad \bullet\text{Radical}$$

It is evident from the above reaction scheme that the protonated peptides should have a minimum charge state of +2. The mechanism of ECD cleavage is thought to involve the release of an energetic hydrogen atom upon electron capture. The released hydrogen atom is collisionally de-excited and then later captured at sites having high hydrogen atom affinity, such as carbonyl oxygens and disulfide bonds ***(73)***. The cleavage occurs at the site of hydrogen atom capture. A unique feature in ECD is that the ECD cleavage is faster than the intramolecular energy randomization and is therefore called a nonergodic process.

The advantage of ECD compared with CID is that ECD produces many more peptide backbone cleavages of peptides with unstable modifications, such as glycans and phophorylations ***(74)***.

1.14.3. ETD (75)

In ETD singly charged anthracene anions transfer an electron to multiply protonated peptides. The transferred electron induces fragmentation of the peptide backbone in a similar fashion to what is observed for ECD. The advantage of ETD compared with ECD is that it can be used in combination with mass analyzers that trap ions with RF electrostatic fields such as ion traps. ECD requires that the precursor sample ions are immersed in a dense population of near-thermal electrons making it a technical challenge to use ECD in ion traps. In fact, ECD is thus far only available in combination with FTICR.

1.14.4. PSD (36)

PSD can be performed on a MALDI reflectron TOF mass spectrometer. Ions that have acquired enough energy from photoactivation and molecular collisions during the desorption process can fragment in the mass analyzer to release energy ***(76)***. If the TOF mass analyzer is operated in linear mode (without the reflector) then the precursor and the fragments of the precursor which are generated in the field-free region will have approximately the same flight time. In the linear mode, fragmentation in the field-free region can only be indicated by a peak broadening in the MS spectrum. However, if the reflector is used then the precursor and the precursor fragments will be separated by their kinetic energy. This means that they will be separated by mass because their velocity is the same. The reflector works by applying a potential in opposite direction of the ions' flight path causing the ions to turn direction. The ions with highest kinetic energy will penetrate deeper into the reflector region before being deflected and will therefore have a longer flight path.

1.15. Detectors

It is outside the scope of this book to cover the different detectors used in MS. However, in doing data analysis of MS data one cannot be ignorant to the characteristics of the different detectors. Detector saturation effects can, for example, skew the near Gaussian mass peaks of intense ions downward, so the experimental mass of intense ions will have negative mass error. In **Fig. 9** an overview of the different instrument techniques that can be used is given together with references for further reading on detectors.

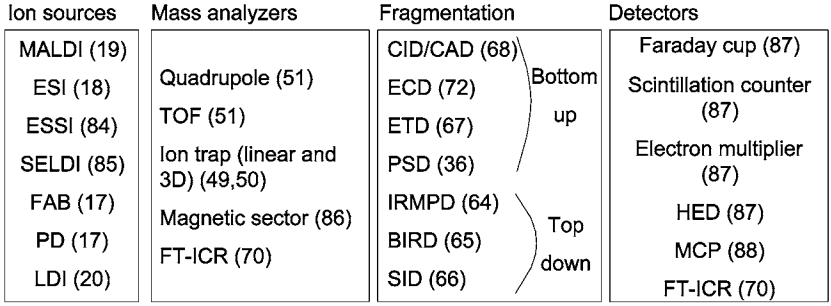


Fig. 9. Overview of different instrument settings. The instrumental methods can be combined in many different ways. Not all methods are discussed in this chapter. However, appropriate references for further reading are provided. Abbreviations are spelled out in **Appendix A**.

1.16. Data Formats Used in Proteomics

The data formats used in MS can be divided in two main groups: (1) formats containing raw data, and (2) formats containing processed data. The raw data formats are often proprietary formats from the MS inventors. Recently, standard raw data formats have been invented to ease data extraction for program developers, such as mzDATA (*see* Chapter 15) and mzXML.

For processed MS/MS spectra a number of file formats exist that are associated with various software applications, such as pkl (MassLynx, Micromass), mgf (Mascot generic file), dat (Data analysis, Bruker), pkx (VEMS), and msm (Xcalibur). In these files the raw and continuous MS/MS spectra have been processed to a peak list. For some of the formats the data can be further processed to contain only singly charged and monoisotopic peaks. The structure of the pkl, pkx, and mgf file formats are shown next:

The mgf file format:

BEGIN IONS
TITLE= Cmpd 2, +MSn(371.23), 31.8 min
PEPMASS= 371.23 762929

101.07 1287
105.12 2277
…
513.22 1081

END IONS

In the Mascot generic file format the fragment ions from a MS/MS spectrum are enclosed between “BEGIN IONS” and “END IONS.” The title line is very

useful for specifying extra information, such as retention times from the chromatogram and spectrum numbers. This is especially useful if searches with Mascot are performed because the title line is associated with the peptide results in the Mascot result display. The *PEPMASS* is the precursor ion mass (*m/z*) and the following lines are fragment ion mass tab delimiter and ion intensity.

The pkl file format

592.5793 617.0220 2
97.0468 1.378e0
112.0647 4.553e0
..
962.1050 3.278e-1

The pkl file format is produced by PLGS v2.05 and Masslynx v4.0 (Waters). In the pkl format the MS/MS spectra are separated by an empty line. The first line in a MS/MS spectrum is precursor ion mass, intensity of precursor ion, and charge state separated by the space character. The following lines are the fragment ion masses and intensity separated by the space character.

The pkx file format

592.5793 617.0220 2 28.6831
97.0468 1.378e0 1
112.0647 4.553e0 1
..
962.1050 3.278e-1 1

In the pkx format that is used in the VEMS program, the MS/MS spectra are separated by an empty line. The first line in a MS/MS spectrum is precursor ion mass, intensity of precursor ion, charge state, and retention time separated by the space character. The following lines are the fragment ion masses, intensity, and original charge state separated by the space character.

1.17. Discussion: Problems to be Solved in MS

In **Subheading 1.1.**, the view that the goal of proteomics should be identification and quantification of proteins in addition to identification and quantification of posttranslational modification in different cells over time was presented. To fulfill these requirements, high-quality data interpretation and storage programs must be available, the fragmentation chemistry should be predictable and fully understood, and good quantification methods should be available.

The data interpretation has lately attracted the attention of many in the proteomics field (*77*). The main problem of today's proteomics is that the introduction of large-scale proteomics has made it almost impossible to analyze all the data

manually. A number of computer programs for automatic interpretation are freely available, such as Proid ***(35)***, Mascot ***(53)***, Lutefisk ***(54)***, and VEMS ***(55)***. However, common to all these programs is that they do not use all the information provided by the MS or MS/MS spectra. The reason is that the MS/MS spectra are complex, containing a large number of mass peaks of known and unknown origin (*see* **Subheading 1.13.**). There are two major problems to be addressed. First, the fragmentation chemistry can be improved. Second, the interpretation algorithms can be improved to recognize and predict abnormal fragmentation chemistry.

Another shortcoming is that most programs use one of the two interpretation approaches. The interpretation algorithms can be divided into two categories: the database-dependent and the database-independent (*de novo*) interpretation algorithms. Both approaches have their strengths and weaknesses. The database-dependent search is better in taking all mass peaks into account, which is especially difficult in *de novo* interpretation if the ion series are incomplete. The advantage of the *de novo* interpretation is that it works even if peptides are either in vivo or in vitro modified. To take advantage of both methods we have developed an algorithm that performs both database-dependent and *de novo* interpretations and automatically compares the result (Chapter 7; VEMS v3.0). Interpretation algorithms can be further improved by including the intensity of the isotopic distribution during the interpretation ***(78)***.

For database-dependent interpretation it is important to have reliable sequence databases available. These can be DNA, mRNA, or protein sequences. In this respect it is important to consider how complete the database is. It is assumed that most databases contain mainly the most abundantly expressed genes and proteins ***(79)***. In addition, the databases are often partly composed of predicted genes and protein sequences. Most of the freely available programs access only some of the available databases, for example Mascot ***(53)*** and Sonar ***(80)***. These programs mainly allow searching against sequence databases from the most studied organisms.

Standardized databases containing the identified, quantified proteins and their modifications need to be improved. These databases need to contain precise information on the organ, tissue or cell type under study, in addition to the result of the identification and quantification. The significance of the result should be evaluated by a statistical test and not only by an algorithm-dependent score. Some examples of the current status of proteome databases are the Expasy server (http://www.expasy.org/) and the Microbial Proteome Database (https://www.abdn.ac.uk/~mmb023/2dhome.htm). Generation of such databases can also be simplified by automatic submission by the program used for interpretation of the mass spectrometric data.

The current methods for quantification are based on 2D-GE or the MudPIT method. Both methods have their limitations. In 2D-GE, quantification is often

based on estimating spot intensities. Currently, fluorescent stains offer the highest dynamic range and facilitate quantification in the 2D-GE method. However, some proteins are known to be excluded from the 2D gels. These proteins have extreme pI, and are small, large, or hydrophobic. In addition, low abundant proteins are difficult to measure. These problems can partly be overcome by prefractionation ***(81)***.

The MudPIT method appears not to underrepresent any particular protein group ***(82)*** and to have a high dynamic range compared with 2D-GE. An additional advantage is that MudPIT can be automated to a larger extent than 2D-GE. The drawback is that quantification is not easily obtained. Quantification with the MudPIT method can be achieved by comparing protein extracts from different sources using different stable isotope labeling for the different sources. One example hereof is the ICAT technique ***(83)***, which is outlined in Chapter 12. The labeled and unlabeled peptides have quite similar physical properties (except from mass), which gives the peptides similar ionization efficiencies and retention times and, thus, the relative intensity of light-labeled and heavy-labeled peptides can be calculated and used as a quantitative measure. If there are overlaps in mass peaks, it becomes necessary to perform peak correction.

2. Notes

1. The rule of thumb is that the enzyme trypsin cleaves after Lys and Arg if not followed by Pro. However, in LC–MS/MS runs, MS/MS spectra are often observed to correspond to a cleavage of Lys or Arg followed by Pro. In these cases, the MS/MS spectra of the peptide corresponding to noncleavage is often also observed and with higher intensity that is if it is not outside the mass range of the mass spectrometer. In general one can also observe missed cleavage sites that often occur when Lys or Arg is neighboring to Asp, Glu, Lys, or Arg.

Acknowledgments

R. M was supported by grants from EU TEMBLOR and by Carlsberg Foundation Fellowships. K. E. M is supported by the BAMSE consortium of Biotechnology industries, Denmark.

References

1. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250,** 4007–4021.
2. Bienvenut, W. V., Sanchez, J. C., Karmime, A., et al. **(**1999) Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot. *Anal. Chem.* **71,** 4800–4807.
3. Willet, W. S., Gillmor, S. A., Perona, J. J., Fletterick, R. J., and Craik C. S., (1995) Engineered metal regulation of trypsin specificity. *Biochemistry* **34,** 2172–2180.

4. Klose, J. and Kobalz, U. (1995) Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. *Electrophoresis* **16,** 1034–1059.
5. Inagaki, N. and Katsuta, K. (2004) Large gel two-dimensional electrophoresis: improving recovery of cellular proteome. *Current Proteomics* **1,** 35–39.
6. Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.* **19,** 1720–1730.
7. Patton, W. F., Schulenberg, B., and Steinberg, T. H. (2002) Two-dimensional gel electrophoresis; better than a poke in the ICAT? *Curr. Opin. Biotechnol.* **13,** 321–328.
8. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.
9. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207.
10. Patterson, S. D. and Aebersold, R. (1995) Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis* **16,** 1791–1814.
11. Matthiesen, R., Bauw, G., and Welinder, K. G. (2004) Use of performic acid oxidation to expand the mass distribution of tryptic peptides. *Anal. Chem.* **76,** 6848–6852.
12. Smith, R. L. and Shaw, E. (1969) Pseudotrypsin. A modified bovine trypsin produced by limited autodigestion. *J. Biol. Chem.* **244,** 4704–4712.
13. Larsen, M. R., Cordwell, S. J., and Roepstorff, P. (2002) Graphite powder as an alternative or supplement to reversed-phase material for desalting and concentration of peptide mixtures prior to matrix-assisted laser desorption/ionization-mass spectrometry. *Proteomics* **2,** 1277–1287.
14. Hagglund, P., Bunkenborg, J., Elortza, F., Jensen, O., and Roepstorf, P. (2004) A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation. *J. Proteome Res.* **3,** 556–566.
15. Pinkse, M. W. H., Uitto, P. M., Hilhorst, M. J., Ooms, B., and Heck, A. J. R. (2004) Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-nanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal. Chem*. **76,** 3935–3943.
16. Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P., and Jørgensen, T. J. D. (2005) Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Proteomics* **4,** 873–886.
17. Burlingame, A. L., Boyd, R. K., and Gaskell, S. J. (1994) Mass spectrometry. *Anal. Chem.* **66,** 634R–683R.
18. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71.
19. Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60,** 2299–2301.

20. James, P. (2001) *Proteome Research: Mass Spectrometry*. Springer-Verlag, New York, pp. 35.
21. Bökelmann, V., Spengler, B., and Kaufmann, R. (1995) Dynamical parameters of ion ejection and ion formation in matrix-assisted laser desorption/ionization. *Eur. Mass Spectrom.* **27,** 156–158.
22. Beavis, R. C. and Chait, B. T. (1990) Rapid, sensitive analysis of protein mixtures by mass spectrometry. *Proc. Natl. Acad. Sci. USA* **87,** 6873–6877.
23. Cohen, S. L. and Chait, B. T. (1996) Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins. *Anal. Chem.* **68,** 31–37.
24. Vorm, O., Roepstorff, P., and Mann, M. (1994) Improved resolution and very high sensitivity in MALDI TOF of matrix surfaces made by fast evaporation. *Anal. Chem.* **66,** 3281–3287.
25. Kussmann, M., Nordhoff, E., Nielsen, R. B., et al. (1997) Matrix-assisted laser desorption/ionization mass spectrometry sample preparation techniques designed for various peptide and protein analytes. *J. Mass Spectrom.* **32,** 593–601.
26. Guilhaus, M. (1995) Principles and instrumentation in time-of-flight mass spectrometry. *J. Mass Spectrom.* **30,** 1519–1552.
27. Vestal, M. L., Juhasz, P., and Martin, S. A. (1995) Delayed extraction matrix-assisted laser desorption time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **9,** 1044–1050.
28. Keller, B. O. and Li, L. (2000) Discerning matrix-cluster peaks in matrix-assisted laser desorption/ionization time-of-flight mass spectra of dilute peptide mixtures. *J. Am. Soc. Mass Spectrom.* **11,** 88–93.
29. Harris, W. A., Janecki, D. J., and Reilly, J. P. (2002) Use of matrix clusters and trypsin autolysis fragments as mass calibrants in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **16,** 1714–1722.
30. Guo, Z., Zhang, Q., Zou, H., Guo, B., and Ni, J. (2002) A method for the analysis of low-mass molecules by MALDI-TOF mass spectrometry. *Anal. Chem.* **74,** 1637–1641.
31. Coursey, J. S., Schwab, D. J., and Dragoset, R. A. (2001) *Atomic Weights and Isotopic Compositions* (v2.3.1). National Institute of Standards and Technology, Gaithersburg, MD; http://physics.nist.gov/Comp. Last accessed 05/27/2006.
32. Snyder, A. P. (2000) *Interpreting protein mass spectra, a comprehensive resource*. Oxford University Press, Oxford, UK.
33. Zheng, P. P., Luider, T. M., Pieters, R., et al. (2003) Identification of tumor-related proteins by proteomic analysis of cerebrospinal fluid from patients with primary brain tumors. *J. Neuropathol. Exp. Neurol.* **62,** 855–862.
34. Jensen, O. N., Podtelejnikov, A. V., and Mann, M. (1997) Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal. Chem.* **69,** 4741–4750.
35. Zhang, N., Aebersold, R., and Schwikowski, B. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2,** 1406–1412.

36. Spengler, B., Kirsch, D., and Kaufmann, R. (1991) Metastable decay of peptides and proteins in matrix assisted laser desorption mass spectrometry. *Rapid Commun. Mass Spectrom.* **5,** 198–202.
37. Chowdhury, S. K. and Chait, B. T. (1991) Method for the electrospray ionization of highly conductive aqueous solutions. *Anal. Chem.* **63,** 1660–1664.
38. Smith, R. D., Loo, J. A., Edmonds, C. G., Barinaga, C. J., and Udseth, H. R. (1990) New developments in biochemical mass spectrometry: electrospray ionization. *Anal. Chem.* **62,** 882–899.
39. Ikonomou, M. G., Blades, A. T., and Kebarle, P. (1991) Electrospray-ion spray: a comparison of mechanisms and performance. *Anal. Chem.* **63,** 1989–1998.
40. Covey, T. R., Bonner, R. F., Shushan, B. I., and Henion, J. (1988) The determination of protein, oligonucleotide and peptide molecular weights by ion-spray mass spectrometry. *Rapid Commun. Mass Spectrom.* **2,** 249–256.
41. Mirza, U. A. and Chait, B. T. (1994) Effects of anions on the positive ion electrospray ionization mass spectra of peptides and proteins. *Anal. Chem.* **66,** 2898–2904.
42. Köcher, T., Allmaier, G., and Wilm, M. (2003) Nanoelectrospray-based detection and sequencing of substoichiometric amounts of phosphopeptides in complex mixtures. *J. Mass Spectrom.* **38,** 131–137.
43. Gatlin, C. L. and Tureçek, F. (1994) Acidity determination in droplets formed by electrospraying methanol-water solutions. *Anal. Chem.* **66,** 712–718.
44. Urvoas, A., Amekraz, B., Moulin, C., Clainche, L. L., Stocklin, R., and Moutiez, M. (2003) Analysis of the metal-binding selectivity of the metallochaperone CopZ from Enterococcus hirae by electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **17,** 1889–1896.
45. Wilm, A. and Mann, M. (1996) Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68,** 1–8.
46. Choudhary, G., Apffel, A., Yin, H., and Hancock, W. (2000) Use of on-line mass spectrometric detection in capillary electrochromatography. *J. Chromatogr. A.* **887,** 85–101.
47. Schneider, B., Schurbert, M., and Ingendoh, A. (1999) *On-Line, Near-Orthogonal Nano-Electrophoresis Coupled With Ion Trap MS for Proteomics Analysis*. Bruker Daltonik GmbH, Bremen, Germany.
48. Wilm, M., Shevchenko, A., Houthaeve, T., et al. (1996b) Femtomole sequencing of proteins from polyacrylamide gels by nano electrospray mass spectrometry. *Nature* **379,** 466–469.
49. Todd, J. F. J. and March, R. E. (1999) A retrospective review of the development and application of the quadrupole ion trap prior to the appearance of commercial instruments. *Int. J. Mass Spectrom.* **190/191,** 9–35.
50. Jonscher, K. R. and Yates, J. R. (1997) The quadrupole ion trap mass spectrometer—a small solution to a big challenge. *Anal. Chem.* **244,** 1–15.
51. Chernushevich, I. V., Loboda, A. V., and Thomson, B. A. (2001) An introduction to quadrupole–time-of-flight mass spectrometry. *J. Mass Spectrom.* **36,** 849–865.
52. Kinter, M. and Sherman, N. E. (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. John Wiley and Sons, New York.

53. Creasy, D. M. and Cottrell, J. S. (2002) Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* **2,** 1426–1434.
54. Taylor, J. A. and Johnson, R. S. (2001) Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* **73,** 2594–2604.
55. Matthiesen, R., Bunkenborg, J., Stensballe, A., Jensen, O. N., Welinder, K. G., and Bauw, G. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* **4,** 2583–2593.
56. Wysocki, V. H., Tsaprailis, G., Smith, L. L., and Breci, L. A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **35,** 1399–1406.
57. Paizs, B. and Suhai, S. (2002) Towards understanding some ion intensity relationships for the tandem mass spectra of protonated peptides. *Rapid Commun. Mass Spectrom.* **16,** 1699–1702.
58. Roepstorff, P. and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **11,** 601.
59. Havilio, M., Haddad, Y., and Smilansky, Z. (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.* **75,** 435–444.
60. Summerfield, S. G., Bolgar, M. S., and Gaskell, S. J. (1997) Promotion and stabilization of ions in peptide b1 phenythiocarbamoyl derivatives: analogies with condensed-phase chemistry. *J. Mass Spectrom.* **32,** 225–231.
61. Schlosser, L. and Lehmann, W. D. (2002) Patchwork peptide sequencing: Extraction of sequence information from accurate mass data of peptide tandem mass spectra recorded at high resolution. *Proteomics* **2,** 524–533.
62. Harrison, A. G., Csizmadia, I. G., Tang, T. H., and Tu, Y. P. (2000) Reaction competition in the fragmentation of protonated dipeptides. *J. Mass Spectrom.* **35,** 683–688.
63. Sleno, L. and Volmer, D. A. (2004) Ion activation methods for tandem mass spectrometry. *J. Mass Spectrom.* **39,** 1091–1112.
64. Ying, G. E., Lawhorn, B.G., Elnaggar, M., Sze, S.K., Begley, T. P., and Mclafferty F. W. (2003) Detection of four oxidation sites in viral prolyl-4-hydroxylase by top-down mass spectrometry. *Protein Sci.* **12,** 2320–2326.
65. Price, W. D., Schnier, P. D., and Williams, E. R. (1996) Tandem mass spectrometry of large biomolecule ions by blackbody infrared radiative dissociation. *Anal. Chem*. **68,** 859.
66. Cooks, R. G., Ast, T., and Beynon, J. H. (1975) Anomalousmetastable peaks. *Int. J. Mass Spectrom. Ion Phys*. **16,** 55.
67. Hadden, W. F., and McLafferty, F. W. (1968) Metastable ion characteristics. VII. Collision-induced metastables. *J. Am. Chem. Soc.* **90,** 4745–4746.
68. Csonka, I. P., Paizs, B., Lendvay, G., and Suhai, S. (2000) Proton mobility in protonated peptides: a joint molecular orbital and RRKM study. *Rapid Commun. Mass Spectrom.* **14,** 417–431.
69. Laskin, J. and Futrell, J. H. (2003) Collisional activation of peptide ions in FT-ICR. *Mass Spectrom. Rev.* **22,** 158–181.

70. Lee, S. A., Jiao, C. Q., Huang, Y., and Freiser, B. S. (1993) Multiple excitation collisional activation in Fourier-transform mass spectrometry. *Rapid Commun. Mass Spectrom.* **7,** 819–821.
71. Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **120,** 3265–3266.
72. Cooper, H. J., Hudgins, R. R., Håkansson, K., and Marshall A. G. (2002) Characterization of amino acid side chain losses in electron capture dissociation. *J. Am. Soc. Mass Spectrom.* **13,** 241–249.
73. Zubarev, R. A., Kruger, N. A., Fridriksson, E. K., et al. (1999) Electron capture dissociation of gaseous multiply-charged proteins is favoured at disulphide bonds and other sites of high hydrogen atom affinity. *J. Am. Chem. Soc.* **121,** 2857–2862.
74. Bakhtiar, R. and Guan, Z. (2005) Electron capture dissociation mass spectrometry in characterization of post-translational modifications. *Biochem. Biophys. Res. Comm.* **334,** 1–8.
75. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *PNAS* **101,** 9528–9533.
76. Chaurand, P., Luetzenkirchen, F., and Spengler, B. (1999) Peptide and protein identification by matrix-assisted laser desorption ionization (MALDI) and MALDI-post-source decay time-of-flight mass spectrometry. *J. Am. Soc. Mass Spectrom.* **10,** 91–103.
77. Patterson, S. D. (2003) Data analysis—the Achilles heel of proteomics. *Nature Biotechnol.* **21,** 221–222.
78. Cannon, W. R. and Jarman, K. D. (2003) Improved peptide sequencing using isotope information inherent in tandem mass spectra. *Rapid Commun. Mass Spectrom.* **17,** 1793–1801.
79. Fey, S. J. and Larsen, P. M. (2001) 2D or not 2D. Two-dimensional gel electrophoresis. *Curr. Opin. Chem. Biol.* **5,** 26–33.
80. Field, H. I., Fenyö, D., and Beavis, R. C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2,** 36–47.
81. Gygi, S. P., Rist, B., and Aebersold, R. (2000) Measuring gene expression by quantitative proteome analysis. *Curr. Opin. Biotechnol.* **11,** 396–401.
82. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.
83. Gygi, S. P., Rist, B., Gerber, S. A., Tureçek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999.
84. Takats, Z., Wiseman, J. M., Gologan, B., and Cooks, R. G. (2004) Electrosonic spray ionization. A gentle technique for generating folded proteins and protein complexes in the gas phase and for studying ion-molecule reactions at atmospheric pressure. *Anal. Chem.* **76,** 4050–4058.

85. Kuwata, H., Yip, T. T., Yip, C. L., Tomita, M., and Hutchens, T. W. (1998) Bactericidal domain of lactoferrin: detection, quantitation, and characterization of lactoferricin in serum by SELDI affinity mass spectrometry. *Biochem. Biophys. Res. Commun.* **245,** 764–773.
86. Barefoot, R. R. (2004) Determination of platinum group elements and gold in geological materials: a review of recent magnetic sector and laser ablation applications. *Analytica Chimica Acta* **509,** 119–125.
87. Siuzak, G. (1996) *Mass Spectrometry in Biotechnology*. Academic Press, San Diego, CA.
88. James, H., Barnes, I. V., and Hieftje, G. M. (2004) Recent advances in detector-array technology for mass spectrometry. *Int. J. Mass Spectrom*. **238,** 33–46.

2

Extracting Monoisotopic Single-Charge Peaks From Liquid Chromatography-Electrospray Ionization–Mass Spectrometry

Rune Matthiesen

Summary

Peak extraction from raw data is the first step in analysis of mass spectrometry (MS) data. The quality of this procedure is very important because it affects the quality of all subsequent analysis, such as database searches and peak quantitation. Many methods have been proposed in the literature, yet the number of practical solutions in terms of available software is rather limited. Virtual Expert Mass Spectrometrist (VEMS) v3.0 includes an algorithm for extracting monoisotopic single-charged peaks and their corresponding retention time from liquid chromatography (LC)–MS data. The extracted peaks can subsequentially be exported to other programs or used internally by VEMS to perform peptide mass fingerprinting searches or peptide quantitation. Additionally, VEMS interfaces the commercial program ProteinLynx Global server v2.0.5 for automatic peak extraction from MS/MS spectra obtained by LC–MS/MS.

Key Words: Noise filtering; peak extraction; deisotoping; decharging.

1. Introduction

Liquid chromatography-electrospray ionization–mass spectrometry (LC-ESI–MS) of tryptic peptides produces a wealth of information in the form of peptide masses and peptide retention time(s) on the LC column. In proteomics, the LC system is typically a single hydrophobic reverse-phase column (one-dimensional separation) or an anionic/cationic column followed by a hydrophobic reverse-phase column (multidimensional separation) ***(1)***. The electrospray ion source is responsible for production of charged peptides in the gas phase resulting in tryptic peptide charge states typically from +1 to +4, where the same peptide can appear with different charge states ***(2)***. The mass

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

spectrometer used for LC–MS in proteomics is most often a tandem mass spectrometer that produces MS or both MS and MS/MS data (*see* Chapter 1). The raw data obtained from these experiments contains, in general, transformed and distorted versions of the ideal physical quantity of interest, which is the masses of the intact peptide, the peptide fragments, and the retention time. The conversion of raw data to a peak list consists of the following three steps in Virtual Expert Mass Spectrometrist (VEMS): (1) the instrument-introduced noise in the spectra should be removed, (2) the monoisotopic single-charged mass should be extracted by decharging and deisotoping, and (3) the retention time(s) for the peptides should be extracted. How this is done in theory and practice with VEMS v3.0 is described in this chapter.

1.1. Noise Filtering

Two types of errors are present in experimental data: systematic and random error. Systematic error is often removed by calibration and will not be discussed further in this section. Random error is also called noise. Filtering out noise from the data ideally gives the true signal. The true difference between noise and signal is that noise is not reproducible, whereas signal is. The quality of signals is often expressed as the true signal divided by the standard deviation of the noise. There are many methods for noise removal, such as linear filters ***(3,4)***, penalized least square ***(5)***, Fourier transform filters ***(6)***, and wavelets ***(7)***. The presentation here concentrates on the linear filters, which are computationally fast and have satisfactory performance for proteomics data. Linear filters convert a time series to a new by a linear operation. Linear filters can in general be expressed as ***(4)***

$$y_t = \sum_{r=-q}^{+s} a_r x_{t+r} \qquad (1)$$

where y_t is the smoothed signal. x_t is the current data point, and r iterates over neighboring data points. The smooth width m is equal to q+s+1. a_r are weights and are dependent on the filter type. For example, for a simple, unweighed sliding, average smooth $a_r = 1/m$ for all r. A frequent filter used in MS is the Savitsky Golay filter ***(3)***, which has weights that result in a smoothed signal that corresponds to fitting a low-order polynomial to all smooth intervals (*see* **Fig. 1**). Savitsky Golay filters have been criticized for having end effect problems because it is a symmetrical filter (*see* **Note 1**). However, this is rarely a problem for MS data and can easily be circumvented by combining symmetrical filters together with asymmetrical filters. The quality of the smoothness can be evaluated by the lack of fit and by either the roughness of the data or by maximum entropy (*see* **Note 2**).

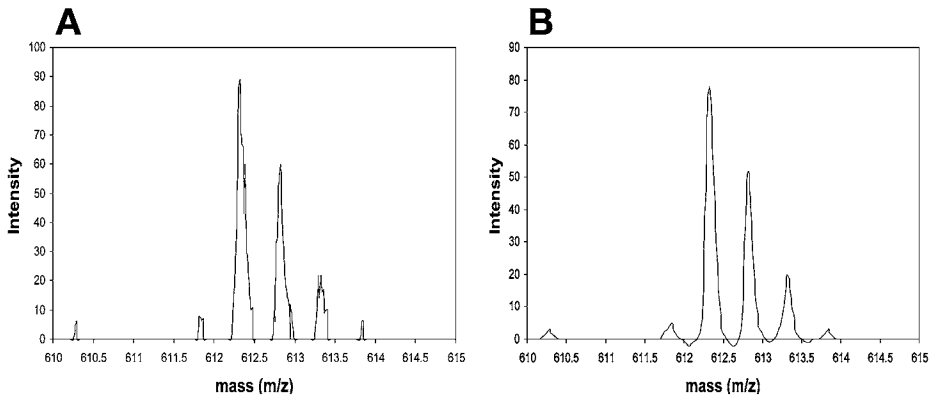


Fig. 1. Savitsky Golay noise filtering. **(A)** Raw mass spectrometry (MS) data. **(B)** MS data from **(A)** after three iterations with a nine-point Savitsky Golay filter.

Alternatively a geometric mean filter with window size 2*m*+1 can be used in combination with a Savitsky Golay filter.

$$y_t = \left(\prod_i x(t + i\Delta t) \right)^{1/(2m+1)} \qquad i = 0, \pm 1, \pm 2, , , \pm m$$

The geometric mean filter can remove spikes because neighboring data points need to be non-zero for a signal to be maintained, and has the additional advantage that the data remains on the same scale ***(8)***.

1.2. Deisotoping and Decharging LC–MS Data

The smoothed raw data is not practical to input into a search engine. Instead, a peak list containing what corresponds to the monoisotopic single-charged ion is often used as input for search engines. The first step is to extract all peaks (*see* **Note 3**) from the smoothed raw data. Peak extraction can be done by extracting peak tops, the centroid method, or by taking the first derivative of the signal (*see* **Fig. 2**). After the peak list is obtained, decharging and deisotoping is done simultaneously by the VEMS program. The algorithm described here for deisotoping and decharging has some similarities to earlier published methods ***(9,10)***. However, the method here is improved by considering information in all MS scan numbers, rather than only considering one scan number at a time. In addition, it considers all combinations of theoretical isotopic distributions of one to two compounds with charge state +1 to +4 to find the best fit to the observed isotopic distribution.

VEMS starts at the first MS scan number from the low mass end, and the program considers high-charge states first.

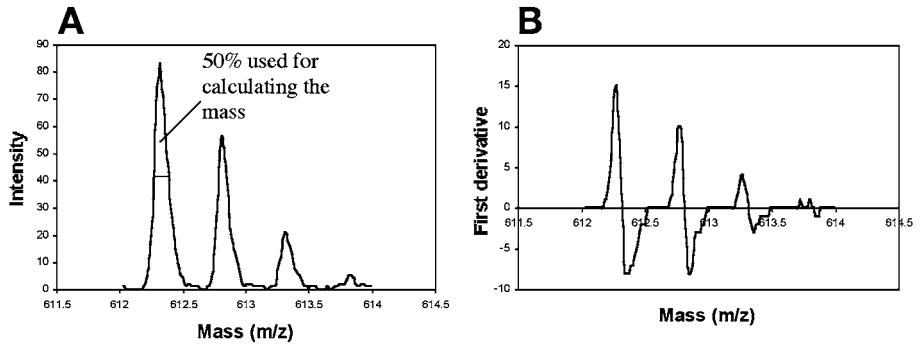


Fig. 2. Converting profile data to peak lists. **(A)** Fifty percent Centroid method. Fifty percent of the resolved part of the peak is used for determining the mass. The mass is calculated by an intensity-weighted average of the masses in the peak. This is equivalent to finding the vertical line passing through the center of gravity of the peak. **(B)** First derivative method. The first derivative of the signal in **(A)** is calculated and the peak masses are determined at the mass points where the first derivative is cutting the *x*-axis.

1. When a peak is encountered by scanning from the low mass end and with low scan numbers, VEMS scans the neighboring MS scans to find the intensity peak maximum in the elution profile.
2. It is likely that the interference from other compounds with similar *m/z* values and retention times is smallest at the scan number obtained in **step 1**. However, there can still be some overlapping peaks. VEMS, therefore, calculates approximate isotopic distributions (*see* **Note 4**) for all possible combinations of two compounds with charge states ranging from +1 to +4, and evaluates which combination fits the observed distribution best by calculating the lack of fit.
3. The best combination obtained in **step 2** is inserted in a new peak list as monoisotopic single-charged mass, intensity, and retention time. After insertion into a new peak list, the theoretical isotopic distribution at the determined charged state is used to remove peaks in the peak list obtained from the raw data corresponding to the observed isotopic distribution over the whole elution profile of the compound. If the best combination was found to contain two compounds, then only the compound corresponding to the peak found in **step 1** is inserted in the peak list and used for removing peaks in the elution profile.
4. **Steps 1–3** are continued until there are no more peaks in the peak list obtained from raw data.

2. Materials

2.1. Required Software

1. VEMS v3.0 (http://yass.sdu.dk). To follow this guide it is also necessary to download the raw data (http://yass.sdu.dk/raw/my00234kr.raw.rar).

2. Microsoft Windows. Currently VEMS is only fully tested on Windows XP and Windows 2000.

2.2. Optional Software

1. PLGS v2.05 and Masslynx v4.0 are commercial programs that can be obtained from Waters (Milford, MA). VEMS interfaces to some of the raw data processing tools of PLGS v2.05 and MassLynx v4.0. It is important that PLGS v2.05 and Masslynx v4.0 are installed in the default directory, otherwise the interfacing from VEMS will not work. If the commercial software is not available, then one can use ExrawNoPKX to convert mzData MS data to the VEMS MS data format. PLGS v2.05 and Masslynx v4.0 are only necessary for the methods described in **Subheading 3.2.**

3. Methods

This section describes how to extract monoisotopic single-charged peaks from raw LC–MS/MS files. In **Subheading 3.1.**, the extraction of the LC–MS data is presented that is accomplished by the VEMS algorithm described in **Subheading 1. Subheading 3.2.** shows how to extract monoisotopic single-charged peaks from all the MS/MS spectra in a number of LC–MS/MS runs.

3.1. Extract Monoisotopic Single-Charged Peaks From MS Scans

Download VEMS from (http://yass.sdu.dk) and uncompress the folder. In the VEMS directory folder there is a folder named “Exraw.” The files in this folder should be moved directly to “c:\data” folder. This folder contains the program “Exraw.exe” which is for format conversion. The program can extract the LC–MS part of LC–MS or LC–MS/MS runs to an indexed format that is used by the VEMS program. The program can, on the time of writing, convert mzXML files and Micromass raw files to the VEMS LC–MS format.

1. Start the “Exraw.exe” program (*see* **Fig. 3**).
2. Use the directory listbox in area 1 (*see* **Fig. 3**) to choose the folder where the raw data folder my00234kr.raw (can be obtain from http://yass.sdu.dk/raw) is located. Press the button “>>” to select the folder. It should now appear in the listbox in area 2.
3. The listbox in area 3 can now be used to specify the output directory.
4. Now press the button “Raw → VEMS” in area 4. The program will now convert all the specified raw data files to the VEMS LC–MS format.
5. In the output folder there should now be a directory named “my00234” containing the LC–MS data in the VEMS format.

The above steps converted a Micromass raw data file to the VEMS LC–MS format. The following steps describe how to convert mzXML files to the VEMS LC–MS format.

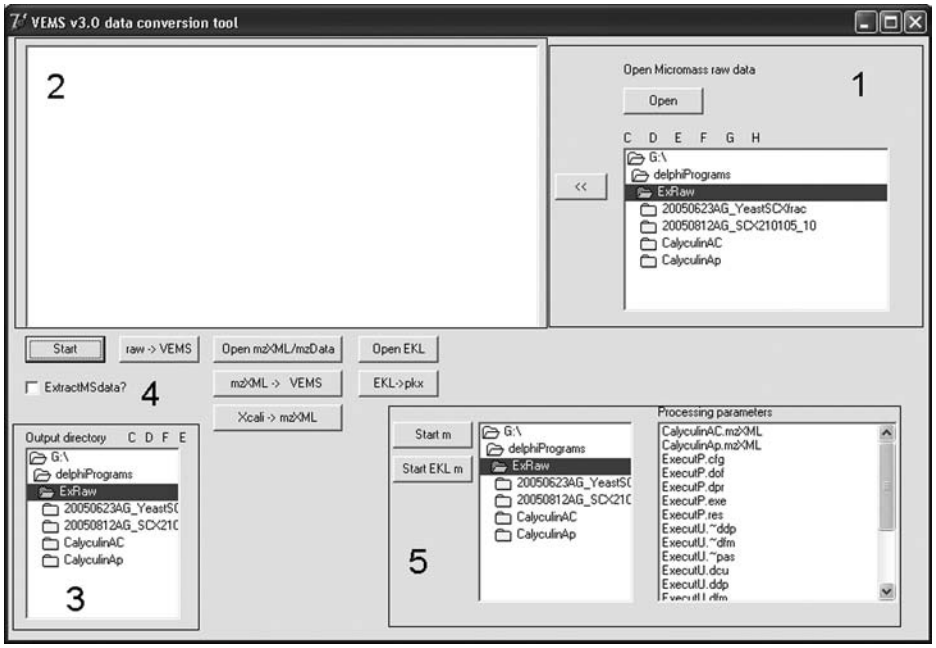


Fig. 3. Screen shot of the VEMS v3.0 data conversion tool. Area 1 is used to specify Micromass raw data files. Area 2 displays the chosen raw data files. Area 3 is used to specify the output directory. Area 4 activates different conversion functions. Area 5 is used to choose files containing different data processing parameters. This is used for optimization of processing parameters.

1. Click the button "Open mzXML" in area 4 to open an mzXML file containing LC–MS or LC–MS/MS data.
2. Choose an output directory in area 3.
3. Click on the button "mzXML → VEMS" in area 4 to create the VEMS LC–MS format in the output directory.

The operations performed so far did not do any data processing, they only extracted the LC–MS data to a more efficient format both in terms of size and data access. The VEMS LC–MS format just created can be used in the VEMS program to extract monoisotopic single-charged peaks from MS scans. VEMS can also use this format for peptide quantitation (*see* Chapter 8). The following describes how to use VEMS to extract peaks from the format. The nomenclature used to describe the user interface is presented in **Appendix E**.

1. Start VEMS_3.exe. Open the data import window from the file menu (File → Open data → Open multiple spectra or press sequentially "Alt"+"F"+"O"+"P").
2. Select the VEMS LC–MS raw data files and close the data import window.

3. Now click on "File → Save → Extract MS peaklist." This will automatically extract peaks from all the specified LC–MS data files and save them in the same folder.

The created peak list(s) can now be specified in the data import window and can be used for peptide mass fingerprinting searches. This is useful when working with a simple protein mixture and higher sequence coverage than achieved by the MS/MS spectra gives is important. Please note that the function activated by **step 1–3** is currently being improved.

3.2. Extract Monoisotopic Single-Charged Peaks From MS/MS Spectra

The VEMS program currently accepts MS/MS peak lists in mgf, pkl, dta, bsc, and pkx. All these formats are ASCII formats containing mass and intensity of parent ion and fragment ions. The pkx format is a VEMS format. The critical reader would probably ask why a new format was made when there are so many already. The reason is that the other formats do not contain all the necessary information for a proper data analysis. For example the pkx format contains the retention time and the original charge state of the peptide fragments before decharging. This section will describe how to make the pkx format from the raw data file "my00234kr.raw."

1. Start the "Exraw.exe" program (*see* **Fig. 3**).
2. Use the directory listbox in area 1 (*see* **Fig. 3**) to choose the folder where the raw data folder my00234kr.raw (can be obtain from http://yass.sdu.dk) is located. Press the button ">>" to select the folder. It should now appear in the listbox in area 2.
3. The listbox in area 3 can now be used to specify the output directory.
4. Now press the button "Start" to create pkx files in the specified output directory.

Alternatively one can check the checkbox "Extract MS data?" then both the VEMS LC–MS format and the pkx formatted files are created in the output window when the "Start" button is pressed.

4. Notes

1. For symmetrical filters $q = s$ in (**Eq. 1**). Symmetrical filters have the drawback that they cannot be evaluated in the start and end of the spectrum that is the q first data points and the s last data points in the spectrum. However, asymmetrical filters where q or s equals zero can be evaluated ***(4)***.
2. The quality-of-function fitting is often evaluated by the lack of fit, which is given by $E_{lof} = \sum_{it}(x_t - y_t)^2$. It is not only the lack of fit that is important for the quality of a fit. For example, the roughness of spectrum, which is given by $R = \sum_{t}(y_t - y_{t-1})^2$, is also important and a best fit can be found by minimizing a weighted sum of E_{lof} and R. Alternatively, the E_{lof} could be evaluated together with maximum entropy of

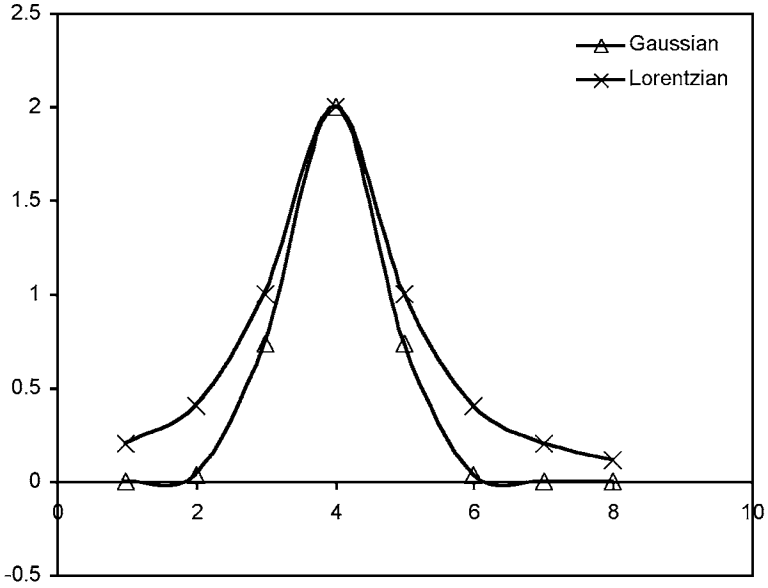


Fig. 4. The peak shape defined by a Gaussian or Lorentzian equation.

residuals, which is given by $S = -\sum_t p_t \log(p_t)$, where p is residuals at different time-points. Maximum entropy is very useful for choosing between different models that give the same $E_{lof.}$

3. Peaks in spectroscopy can have several different shapes that need different mathematical functions for fitting. Peaks can often be approximated by a Gaussian (*see* **Fig. 4**), Lorentzian (*see* **Fig. 4**), or a mixture of the two functions (*see* **Figs. 5–7**). The equation for the Gaussian function is based on the normal distributions and can be formulated as $f(m_i) = A*\exp-(m_i - m_0)^2/s^2$. Where m_0 is at the center, A is the maximum height at x_0, and s defines the peak width. The width at half-height of a Gaussian peak is given by $s(4*\ln 2)^{1/2}$ and the area is $As(\pi)^{1/2}$. The equation for a Lorentzian function is given by $f(x_i) = A/(1+(m_i - m_0)^2/s^2)$, where m_0 is at the midpoint of the peak, and A is the height at the midpoint. The width at half-height of a Lorentzian peak is given by $2s$ and the area is $As\pi$ ***(11)***. In MS the mass of such peaks are often determined by calculating the centroid mass, which is more accurate than just taking the mass at the peak maximum. The centroid mass m_c and the corresponding intensity I_c can be calculated by the following expressions:

$$m_c = \frac{\sum_{y_i > y_{i,\max} x} m_i I_i}{I_c}$$

$$I_c = \sum_{y_i > y_{i,\max} x} I_i$$

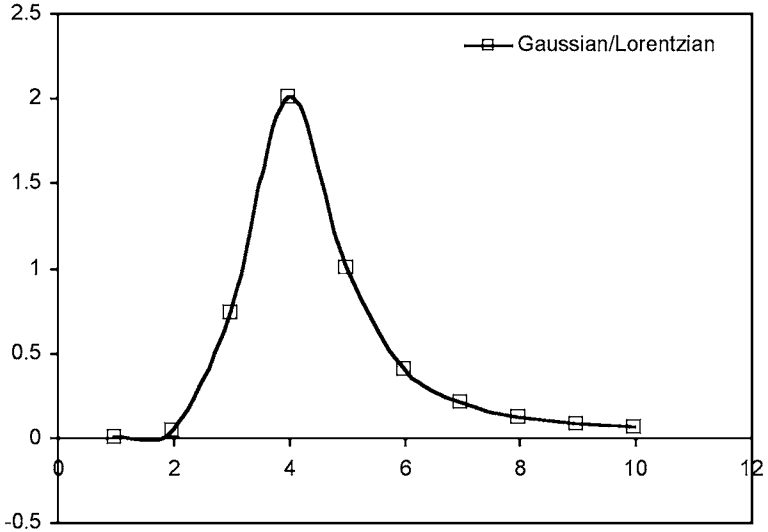


Fig. 5. A mixed model where the peak shape is defined by a Gaussian equation up to the midpoint, and by a Lorentzian function after the midpoint.

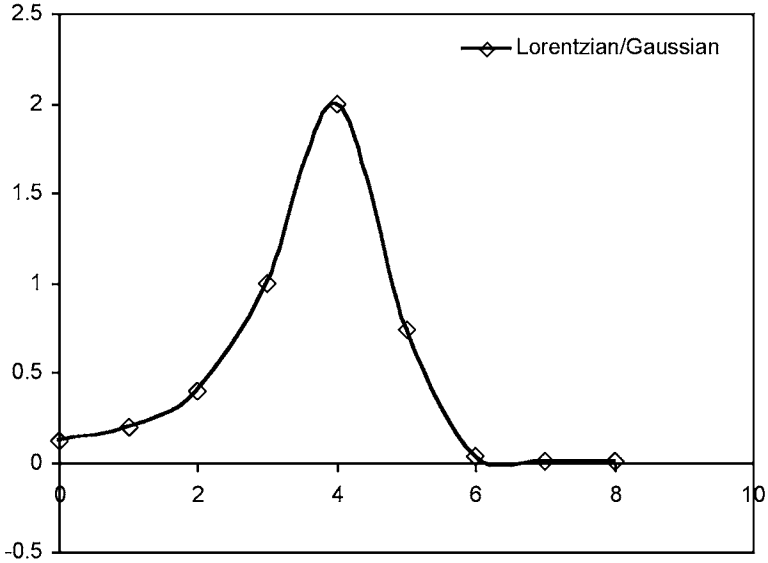


Fig. 6. A mixed model where the peak shape is defined by a Lorentzian function up to the midpoint, and by a Gaussian equation after the midpoint.

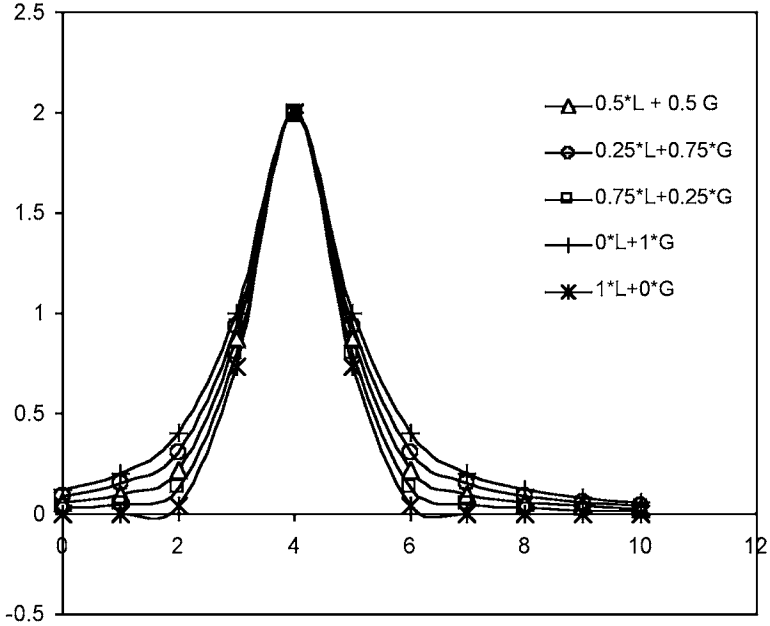


Fig. 7. Examples of Lorentzian and Gaussian mixed models. L is the Lorentzian function and G is the Gaussian.

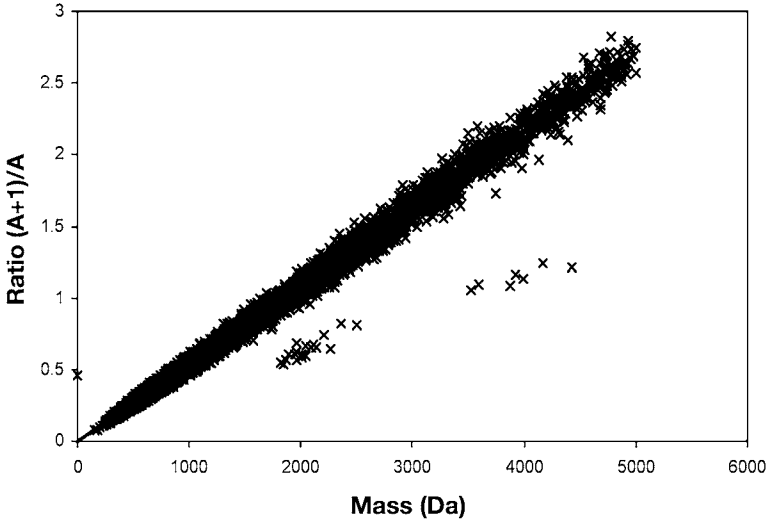


Fig. 8. The ratio between the theoretical abundance of the monoisotopic plus one and the monoisotopic peak plotted as a function of the monoisotopic mass.

where m_i is the mass at a certain mass bin and I_i is the corresponding intensity. x is a specified percentage of the maximum intensity.

4. Approximate isotopic distributions are calculated based on theoretical isotopic distribution of 20,000 standard tryptic peptides. Given the intensity of the monoisotopic peak, the intensity of the following isotopic peaks can be approximated by a linear equation ***(12)***. The ratio R between the intensity of the monoisotopic peak and the monoisotopic plus is approximated by $R = 0.0005412*m–0.01033$, where m is the mass of the monoisotopic peak (*see* **Fig. 8**). Similar approximation can be made for the higher masses in the isotopic distribution. The approximate isotopic distributions are used to generate all possible combinations of two overlapping isotopic distributions. The combination used is the one that gives the best fit on the neighboring peaks. The deisotoping problem can also be solved by linear algebra ***(13)*** instead of checking all reasonable possibilities.

Acknowledgments

R. M was supported by grants from EU TEMBLOR and by Carlsberg Foundation Fellowships.

References

1. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.
2. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71.
3. Savitzky, A. and Golay, J. E. M. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36,** 1627–1639.
4. Chatfield, C. (ed.) (1989) *The Analysis of Time Series: An introduction.* Chapman and Hall, New York, pp. 1–8.
5. Eilers, P. H. (2003) A perfect smoother. *Anal. Chem.* **75,** 3631–3636.
6. Kast, J., Gentzel, M., Wilm, M., and Richardson, K. (2003) Noise filtering techniques for electrospray quadrupole time of flight mass spectra. *J. Am. Soc. Mass Spectrom.* **14,** 766–776.
7. Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., and Kobayashi, R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* **21,** 1764–1775.
8. Bylund, D. (2001) *Chemometrics Tools for Enhanced Performance in Liquid Chromatography-Mass Spectrometry.* Uppsala University, Uppsala, Sweden.
9. Wehofsky, M. and Hoffmann, R. (2002) Automated deconvolution and deisotoping of electrospray mass spectra. *J. Mass Spectrom.* **37,** 223–229.
10. Zhang, Z. and Marshall, A. G. (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass Spectrom.* **9,** 225–233.

11. Brereton, R. G. (2003) *Data Analysis for the Laboratory and Chemical Plant*. John Wiley and Sons, New York, pp. 119–168.
12. Wehofsky, M., Hoffmann, R., Hubert, M., and Spengler, B. (2001) Isotopic deconvolution of matrix-assisted laser desorption/ionization mass spectra for substances-class specific analysis of complex samples. *Eur. J. Mass Spectrom.* **7,** 39–46.
13. Meija, J. and Caruso, J. A. (2004) Deconvolution of isobaric interferences in mass spectra. *J. Am. Soc. Mass Spectrom.* **15,** 654–658.

3

Calibration of Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Peptide Mass Fingerprinting Spectra

Karin Hjernø and Peter Højrup

Summary

This chapter describes a number of aspects important for calibration of matrix-assisted laser desorption/ionization time-of-flight spectra prior to peptide mass fingerprinting searches. Both multipoint internal calibration and mass defect-based calibration is illustrated. The chapter describes how potential internal calibrants, like tryptic autodigest peptides and keratin-related peptides, can be identified and used for high-precision calibration. Furthermore, the construction of project/user-specific lists of potential calibrants is illustrated.

Key Words: Internal calibration; mass defect; contaminants; PeakErazor; MALDI-TOF spectra.

1. Introduction

High mass accuracy is crucial when analyzing biomolecules by mass spectrometry (MS); the more precise the data, the more correctly the spectrum reflects the sample analyzed. For identification of proteins by peptide mass fingerprinting (PMF) based on matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS data (*see* Chapter 1), this also means that the better the mass spectrum is calibrated, the better the chance is of finding only relevant proteins and avoiding false-positive hits resulting from random matching ***(1–5)***. Identification based on MS analysis of a single protein is often a trivial case if sufficient peptides are found with an acceptable mass accuracy (30–70 ppm). High mass accuracy (<30 ppm) is likely to be crucial for identification of the correct protein/proteins if the spectrum contains numerous signals unrelated to the protein in question, if they represent modified peptides, or if only few peptide signals are recorded.

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

Three general approaches are available for spectrum calibration:

1. External calibration.
2. Internal calibration.
3. Calibration based on mass defect.

For the external calibration, a mixture of known peptides (e.g., a tryptic digest of lactoglobulin) or a polymer like propylene glycol ***(6)***, is placed as close as possible to the sample of interest and a spectrum is recorded. The calibration constants calculated for the standard can then be used for calibration of the peaks from the unknown sample. Such external calibration methods will usually only provide a mass accuracy of 50–150 ppm. This is owing to spot-dependent mass accuracy variations found throughout the MALDI plate ***(6,7)***. It has been shown by several groups that this variation in most cases can be sufficiently corrected for by a simple linear transformation using two or more frequently recognized signals from components with known mass values ***(8)***. Traditionally, two or more tryptic autodigest products, if observed in the spectrum, have been used as internal calibrants. Whether or not these signals are observed is determined mainly by the substrate-to-enzyme ratio. However, one or more of these signals may be obscured by other peptide signals with near-identical masses. Alternatively, signals recognized in the spectrum as originating from other compounds, such as keratin-related peptides, matrix peaks, or standard compounds added prior to analysis, can be used in combination as internal calibrants. The advantage is that a multipoint calibration using mass values more or less evenly distributed over the entire range of the spectrum can be performed, and will result in a highly trustable accurate calibration. As the calibrants and the sample have been subjected to exactly the same conditions, the internal calibration approach will offer the most precisely calibrated data; 5–40 ppm depending on the quality of the instrument. The disadvantage is that such internal calibrants often suppress the signals from the sample of interest.

If a mass spectrum does not contain any known peaks on which to calibrate (this is particularly the case if the sample amount is large, e.g., digests of proteins in solution or from Coomassie-stained gel spots) it is possible to calibrate the spectrum using the mass defect/peptide mass distribution ***(9,10)*** of all peptide peaks in the spectrum in order to obtain a mass accuracy better than 50 ppm (dependent on the number of peptides present and the quality of the instrument). But what is the mass defect? All amino acid residues are built from a limited number of atoms: carbon, oxygen, nitrogen, and hydrogen (plus a little sulfur) in approximately the same ratio. The residues all have a mass higher than the corresponding integer values, i.e., alanine weighs 71.037 Da instead of 71.0 Da. The decimal part (i.e., 0.037) is called the mass defect of alanine. The mass defect varies between residues, but is on average 0.0454%. Independent of the composition, there will be

an upper and a lower limit for the mass defect of a peptide as this is calculated as the sum of the mass defect of the individual residues. Analysis of the entire protein database shows that less than 0.5% of all tryptic peptides with a mass greater than 1000 Da deviate more than 125 ppm from the calculated average. This results in peptide masses being distributed into discrete clusters with a width of $2 \times 125 = 250$ ppm. Using this as a constraint, it is possible to calibrate MALDI-TOF spectra to a precision of 30–50 ppm based on the average mass defect of all peptide peaks in the spectrum.

In this chapter, we will illustrate how potential internal calibrants can be identified, evaluated, and used for calibration of peak lists from any MALDI-TOF instrument, and how mass defect-based calibration can be performed if no internal calibrants are present. All steps will be performed using lists of spectrum mass values as query in the software tool PeakErazor ***(11)***. The program works as an interactive tool, which requires human intervention and as such is not high throughput. However, it offers the user the freedom to study the performance of the MS instrument and identify frequently encountered contaminants useful for recalibration of the spectra.

It should be stated that if a MALDI-MS instrument is poorly calibrated or the parameters used during data acquisition are wrong, then high mass accuracy of the MS data can be difficult to obtain even when postacquisition calibration is performed.

2. Materials

The PeakErazor program is freeware and runs under all 32-bit Windows versions. The data format of the input is simply a list of experimentally obtained mass values.

3. Methods

3.1. Installing the Software

1. Download the software from http://www.gpnow.com.
2. Download the available Erazor list (erazorlist.lst) into the same directory.
3. Open the program by double-clicking on the PeakErazor icon.

3.2. Identification of Potential Internal Calibrants

As PeakErazor uses a list of monoisotopic mass values (a peak list) as input, it is required that proper peak extraction has been performed prior to calibration. Peak extraction is described elsewhere in the book (*see* Chapter 2).

The underlying principle in PeakErazor is simple; the experimental peaklist is compared to an Erazor list containing the exact masses of known contaminants like trypsin, keratin, or matrix-related signals. The accepted contaminants

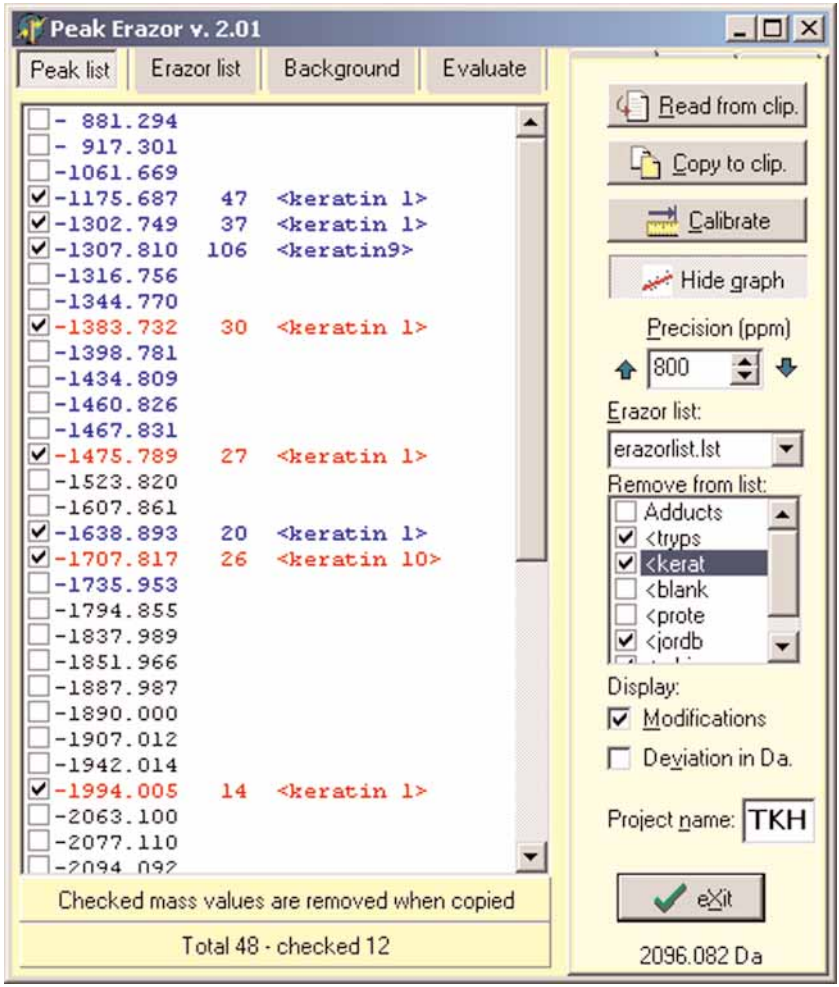


Fig. 1. The main window of PeakErazor. For the data seen here, eight peaks are suggested to be originating from three different keratins: keratin 1, 9, and 10. The mass deviation from the actual mass of the contaminant is in the range of 14 to 106 ppm. If only keratin 1 peptides are taken into account, then the mass range is 14–47 ppm.

are used as internal calibrants through a linear fit, and the contaminants are in the same procedure removed from the list in order to improve the subsequent PMF search (*see* Chapter 4).

1. Start by copying the peak list in question to the clipboard from your preferred spectrum analysis program.
2. Paste the peak list into PeakErazor using the “Read from clip” button in the main window, or press Ctrl + v (*see* **Fig. 1**).

3. Define the "Precision" (i.e., mass tolerance) by which your data should match the theoretical mass values in the Erazor list. Let this mass tolerance be set at a high value (e.g., 500 or 800 ppm) as the accuracy of the data is not known at this stage of analysis.
4. Select an Erazor list (*see* **Note 1**). The construction of user-defined Erazor lists is described elsewhere (*see* Chapter 4).
5. Specify which types of contaminants should be reported, choose between keratin, trypsin, unknown, and so on. The contaminants suggested by the software will be listed to the right of the mass list with deviation and ID (name of component). Be aware that the precision of the calibration is highly dependent on the precision of the mass values of the contaminants. The mass values of tryptic autodigestion products and common keratin peaks are usually known with high precision, whereas frequently observed peaks arising from peptides with unknown sequence (listed as "unknown" in the Erazor list) is only reported with an approximated mass value. As a consequence, these "unknown" contaminants are less useful as calibrants and should be eliminated (by unchecking "unknown" in the list) before the actual calibration is performed. Remember to add the "unknown" (by checking "unknown" in the list) before copying the mass list back to the clipboard in order to remove the unknown from the peak list prior to analysis.
6. Toggle the graph window on by clicking at the "show graph" button. The experimental masses of the suggested contaminants will be displayed as a function of mass deviation (*see* **Fig. 2**). This graph is extremely useful for visualization of possible calibration trends. It is possible to choose which type of labels (deviation, mass, or both) to be connected to each dot in the graph (*see* **Note 2**).
7. Press the "view calib." button below the graph in order to visualize the linear fit of the data.
8. Study the mass accuracy of the data. Remove, by manually unchecking in the main window, mass values for which the mass deviations fall outside the general trend of the data (outliers). In the example given in **Fig. 2**, two mass values fall outside the general trend and should be unchecked (*see* **Fig. 2**). These are presumably not contaminants, but belong instead to the protein in question and should be included in the PMF search (*see* **Note 3**).
9. Use the remaining contaminant peaks as internal calibrants as described in **Subheading 3.3.** Be aware that the best calibration is obtained if calibrants spanning a large mass range are used.

3.3. Calibration Using Internal Calibrants

1. Start the calibration by pressing the "Calibrate" button in the main window (*see* **Fig. 1**).
2. Redefine the mass accuracy by which your data should match the theoretical mass values depending on the precision obtained after calibration. As a result, the window of the graph will be zoomed leaving the user with the opportunity to restudy the calibration (*see* **Fig. 3**).

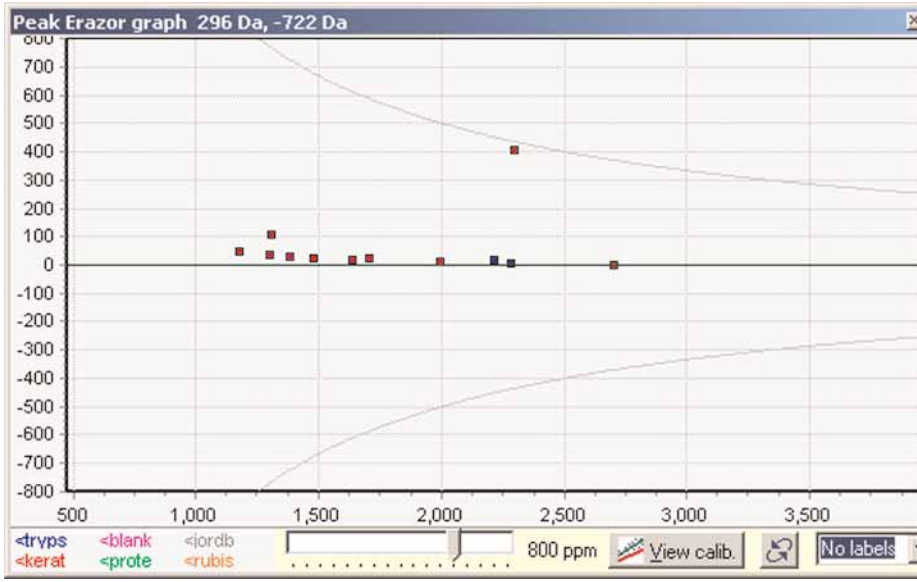


Fig. 2. Graphic representation of the deviation as a function of the mass values. Two values deviate from the other data points. One of them is the peak at mass value 1307 with a deviation of 106 ppm. This peak is the only one suggested to originate from keratin 9. The other one is a peak at mass 2260 with a mass deviation of 400 ppm. As indicated by the gray line, this corresponds to a mass deviation of 1 Da. By inspection of the mass spectrum, it was found that this resulted from wrong peak assignment. The remaining dots (with a mass accuracy within ± 47 ppm) can be used for a linear transformation in order to obtain a higher mass precision of the data (*see* **Fig. 5**).

3. Repeat the calibration step if such a zoom reveals new masses falling outside the general trend of the data, or if new mass values have moved into the displayed precision zone (*see* **Note 4**).
4. Copy the calibrated mass list back onto the clipboard. From here it can be pasted directly into any program of interest, like a PMF search program.

3.4. Evaluation of Contaminants

A wrongly identified contaminant can mess up the calibration, particularly if the calibration is based on only a few mass values. The calibration method used here is based on a list of contaminants frequently observed in our spectra. This list may not reflect the contaminants observed in other laboratories, causing random matching to mass values otherwise never observed and, therefore, to wrongly assigned contaminants. Thus, the frequency of which the contaminant

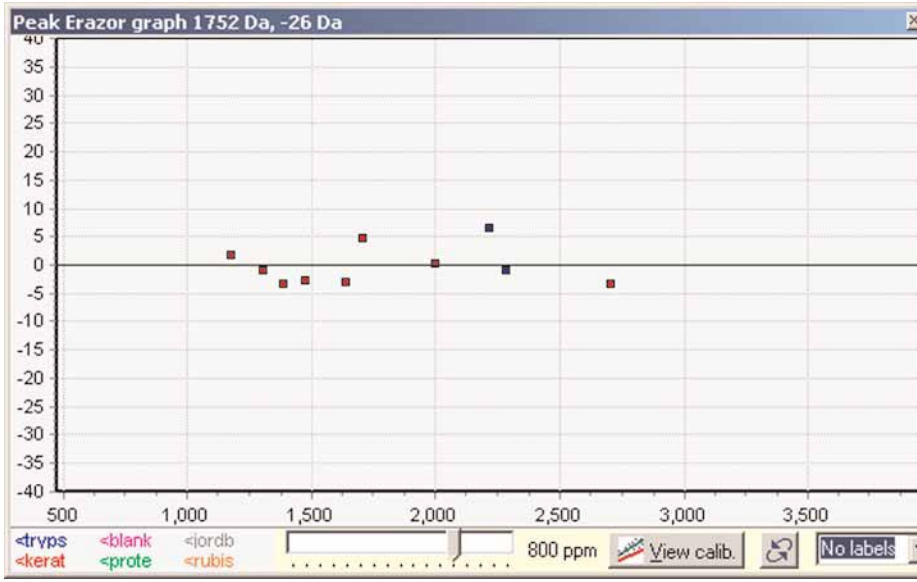


Fig. 3. Calibration using multiple internal calibration. Here is shown the result of an internal calibration of the data shown in **Figs. 1** and **2**. As can be seen, the mass precision went from ±47 to ±8 ppm. Please notice that the mass precision of the graph is ±40 ppm compared with ±800 in **Fig. 2** and in the right-hand insert.

occurs in a specific study can give a hint to whether the contaminant is a true contaminant or not. PeakErazor calculates this frequency based on all mass values ever presented to the program by the user (when copied to the clipboard). The list of contaminants can, therefore, be evaluated from time to time resulting in Erazor lists containing only contaminants observed by the user.

1. Click on the page tab “Evaluation” and open the file “allMass.mss” containing all mass lists treated in PeakErazor by the user (*see* **Fig. 4**). The distribution of all experimentally obtained mass values will be displayed in the graph; the upper part representing the mass values of the sample-specific contaminants and the lower part the contaminants (i.e., values rejected in the program). High-intensity signals from mass values in the upper part (i.e., values commonly observed in spectra) indicate the presence of potential contaminants that were previously unknown to the user. Mass values in the lower part with low intensity represent peptides, which have previously been taken as contaminants but, as they seldom occur, may not be true contaminants.
2. Push the auto-evaluation button.
3. Study the frequency of the mass values in the table (displayed in the second column as the number of times a certain mass value has been observed [#]) (*see* **Fig. 5**). A true contaminant will be represented by a high frequency.

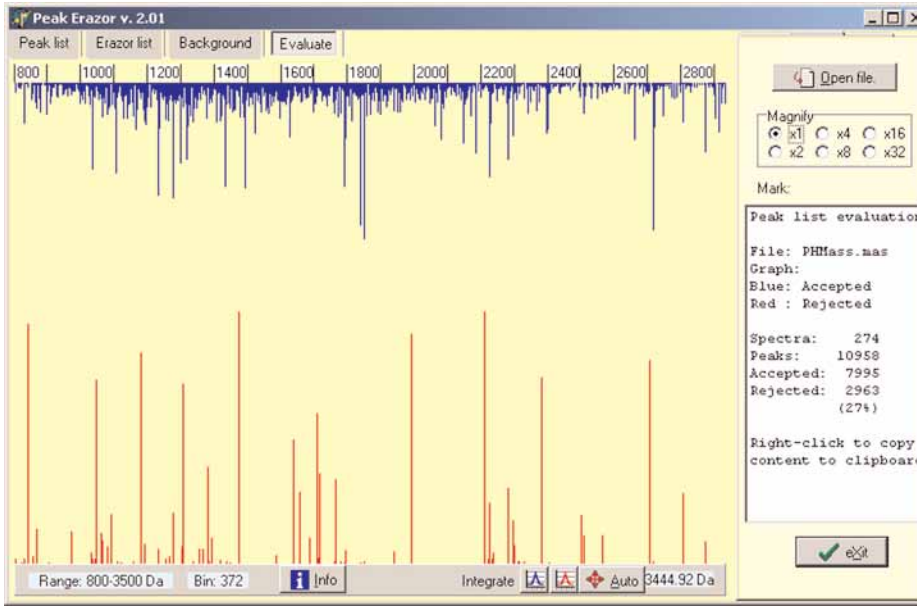


Fig. 4. The evaluation page summarizing the mass distributions of the analyzed data. In the example given here, 274 spectra have been analyzed. In these, 7995 peaks out of a total of 10,958 have been accepted and subjected for further analysis by peptide mass fingerprinting searches (*see* Chapter 4). These are represented by bars in the upper part of the page. The numbers 800–2800 represent the mass values in the analyzed peak lists, and the length of the bars represents the number of times a specific mass value is observed in the 274 spectra analyzed. The bars in the lower part of the page represent the 2963 peaks believed to be contaminants; these have been rejected from the peak list prior to further analysis. In total, 29% of all the peaks have been rejected as contaminants. As can be seen by the length of the bars, some mass values taken as contaminants are only observed a few time (short bars), whereas others are commonly observed (long bars). The long bars are therefore more likely to represent true contaminants, and the mass values resulting in short bars can therefore be removed from the list of contaminants. Conversely, peaks represented in the upper part of the page by long bars are likely to be contaminants and should be included in the list of contaminants.

4. Set the values for construction of a new erazorlist.lst. Select an “Erazor list precision” and a “Base integration width.” Select how many times a certain mass value should be observed in order to be a true contaminant (move the scroll bar in order to change the actual number or click the “+1” button).
5. Push the Remove button.
6. Right-click and choose “save” from the pop-up menu in order to save the new Erazor list (*see also* **Note 1**).

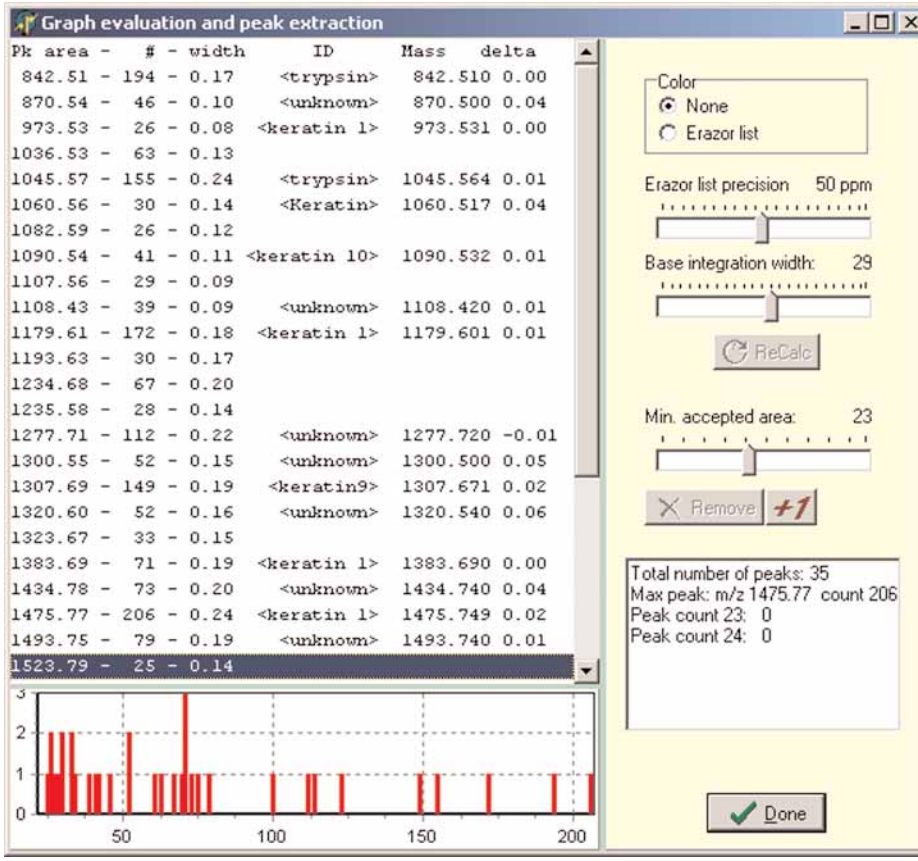


Fig. 5. Graph evaluation and peak extraction. The purpose of this part of PeakErazor is to construct a new Erazor list based on the analyzed data. This will result in a list of contaminants more correctly reflecting the contaminants found in a specific project or specific laboratory. In the example given, both tryptic autodigest peptides <trypsin>, keratin peptides <keratin>, and peptides of unknown origin <unknown> are recognized. These have been observed with different frequencies; from a novel potential contaminant peak at 1523, which has been observed 25 times, to a keratin peak at 1475, which has been observed 206 times. In the dataset shown here, this keratin peptide is observed more times than the commonly known tryptic autodigest peptide at 842.5.

3.5. Calibration Using Mass Defect

A blue-colored mass value in the main window of PeakErazor indicates that the mass defect is larger or smaller than expected for the given mass (default is ± 125 ppm).

1. Click the icon with two arrows (to the right of the “view calib” button) in order to toggle between internal calibration and calibration based on the mass defect. Now

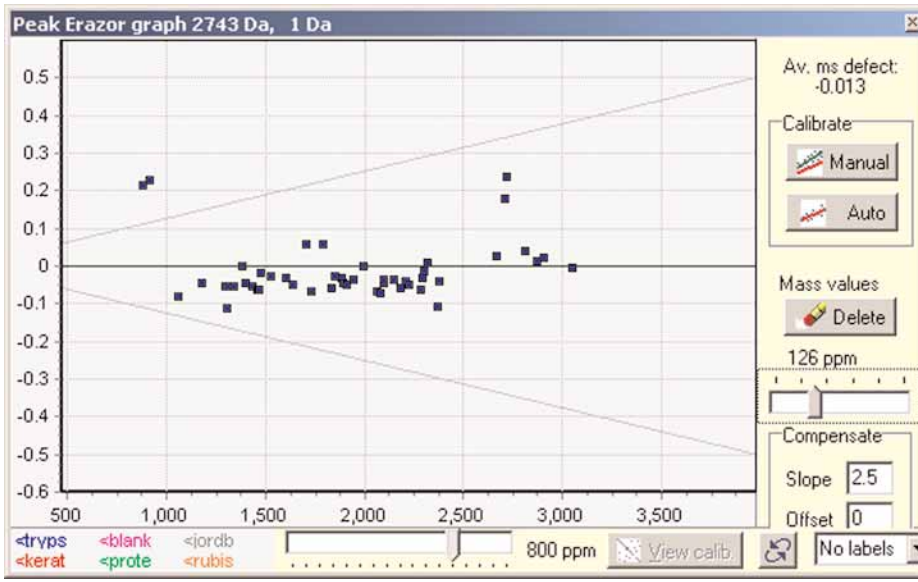


Fig. 6. Mass defect calibration. The mass defect of each mass value, contaminant or not, is represented by a dot. All dots positioned within the gray lines after a multipoint calibration have a mass defect value expected for a peptide. The two dots, which do not fall within these lines, are not peptide peaks, but can instead be assigned as matrix peaks.

every mass value in the peak list will be represented by a dot (showing the deviation from the average mass defect for a given mass), independently of whether it is a contaminant peak or not.

The next step depends on the shape made up of dots; in the case where the mass defect graph is divided into two or more sections (e.g., one group above and one below the zero line) a manual correction has to be performed prior to an automated calibration (*see* example in **Fig 6**).

2. Select “Manual” and drag a line through the low mass group of dots (click and release the mouse at one end of the group, move the cursor to the other end, and click again). This will result in a redistribution of the dots into a single group (*see* **Note 5**).
3. Press the “Auto” button to perform a multipoint calibration.
4. Study the resulting graph. If any points fall outside of the gray lines (default is 125 ppm, set in the right-hand slider) (*see* **Note 6**) they are likely to arise from nonpeptide peaks (typically matrix ions) or from adduct ions (Na, K) and can be deleted through the “Delete” button (*see* **Note 7**). Two peaks arising from matrix clusters are seen in the example in **Fig 6**. The “Slope” values vary a bit between proteins, but a value of 2.5 fits most proteins (*see also* **Note 6**). The “Offset” is currently not used in PeakErazor and should always be zero.

4. Notes

1. The erazorlist.lst from the website is a list of typically observed contaminant (including trypsin autodigest products). This list can be edited to reflect the contaminants observed in a specific laboratory or specific project as described in **Subheading 3.4.** Alternatively, mass values of contaminants can be added manually by selecting the Erazor list page at the top of the main page and clicking the "add" button. Mass values of typically observed contaminants have been reported by several groups ***(12,13)*** and a list of trypsin autodigest products can be found at the Prospector site (http://prospector.ucsf.edu/ucsfhtml4.0/misc/trypsin.htm).
2. It is possible to expand the mass range studied by right-clicking in the graph window and choosing "expanded mass area." This is of advantage if peptides of masses higher than 4000 Da are observed in the spectrum.
3. If a dot in the spectrum after calibration falls on a gray line (±1 Da) it may indicate one of the following situations: (1) either a wrong isotope selection during peak annotation, causing the dot to fall on either of the gray lines (check the spectrum), or (2) it can be the result of a deamidation event (amide to acid), in this case the dot falls on the upper gray line.
4. After protein identification by PMF, the peak list can be recalibrated using the mass values of the identified peptides, thereby improving the mass accuracy even further. This is especially important if the spectrum is suspected to represent more than one protein; improving the mass accuracy will increase the chance of identifying additional proteins.
5. The mass defect-based calibration suffers from the risk of a 1-Da calibration offset (only a problem if the original data has a very low initial mass accuracy). You should therefore always perform the initial manual calibration on the low mass group.
6. In the case of protein identification based on PMF, we are usually looking at tryptic peptides, which always (except for the C-terminal peptide) terminate in a lysine or arginine, resulting in a dataset with skewed composition. Lysine and arginine are two of the residues that have the highest mass defect of all 20 residues, which results in a higher relative mass defect at low *m/z* values, and a lower one at high *m/z* values, which is compensated for by setting the slope value to 2.5. It has been calculated that around 99.5% of all *in silico*-digested tryptic peptides deviate less than 125 ppm from the calculated average (slope = 2.5). It is relatively safe to delete all *m/z* values that are outside of this limit in the calibrated mass list. The slope value should be set to zero if the protein in question has been digested with another enzyme other than trypsin.
7. Mass defect-based calibration of spectra dominated by adduct ions (from, e.g., Na or K) will often result in low mass accuracy as the mass defect of adducts deviate considerably more than 125 ppm from the average peptide.

References

1. Jensen, O. N., Podtelejnikov, A., and Mann, M. (1996) Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps. *Rapid Comm. Mass Spectrom.* **10,** 1371–1378.

2. Jensen, O. N., Mortensen, P., Vorm, O., and Mann, M. (1997) Automation of matrix-assisted laser desorption/ionization mass spectrometry using fuzzy logic feedback control. *Anal. Chem.* **69,** 1706–1714.
3. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (+/– 10 ppm) in protein identification strategies employing MS or MS MS and database searching. *Anal. Chem.* **71,** 2871–2882.
4. Green, M. K., Johnston, M. V., and Larsen, B. S. (1999) Mass accuracy and sequence requirements for protein database searching. *Anal. Biochem.* **275,** 39–46.
5. Chamrad, D. C., Korting, G., Stuhler, K., Meyer, H. E., Klose, J., and Bluggel, M. (2004) Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics* **4,** 619–628.
6. Gobom, J., Mueller, M., Egelhofer, V., Theiss, D., Lehrach, H., and Nordhoff, E. (2002) A calibration method that simplifies and improves accurate determination of peptide molecular masses by MALDI-TOF MS. *Anal. Chem.* **74,** 3915–3923.
7. Egelhofer, V., Bussow, K., Luebbert, C., Lehrach, H., and Nordhoff, E. (2000) Improvements in protein identification by MALDI-TOF-MS peptide mapping. *Anal. Chem.* **72,** 2741–2750.
8. Mortz, E., Vorm, O., Mann, M., and Roepstorff, P. (1994) Identification of proteins in polyacrylamide gels by mass-spectrometric peptide-mapping combined with database search. *Biol. Mass Spectrom.* **23,** 249–261.
9. Gay, S., Binz, P. A., Hochstrasser, D. F., and Appel, R. D. (1999) Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis* **20,** 3527–3534.
10. Wool, A. and Smilansky, Z. (2002) Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting. *Proteomics* **2,** 1365–1373.
11. Hjernoe, K. and Hojrup, P. (2004) Peak Erazor: A Window-based programme for improving peptide mass searches. In: *Methods in Proteome and Protein Analysis*. (Kamp, R. M., Calvete, J. J., and Choli-Papadopoulou, T., eds.), Springer-Verlag, Heidelberg, Germany, pp. 359–369.
12. Mattow, J., Schmidt, F., Hohenwarter, W., Siejak, F., Schaible, U. E., and Kaufmann, S. H. E. (2004) Protein identification and tracking in two-dimensional electrophoretic gels by minimal protein identifiers. *Proteomics* **4,** 2927–2941.
13. Harris, W. A., Janecki, D. J., and Reilly, J. P. (2002) Use of matrix clusters and trypsin autolysis fragments as mass calibrants in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Comm. Mass Spectrom.* **16,** 1714–1722.

4

Protein Identification by Peptide Mass Fingerprinting

Karin Hjernø

Summary

Peptide mass fingerprinting is an effective way of identifying, e.g., gel-separated proteins, by matching experimentally obtained peptide mass data against large databases. However, several factors are known to influence the quality of the resulting matches, such as proteins contaminating the sample in question, modifications altering the mass of the peptides, ionization efficiency of the individual peptides, and the degree of missed cleavage sites. Here, these factors are discussed and methods for elimination of contaminants from the dataset and prediction of various modifications are introduced. Useful tips on how to specify various search parameters and how to manually evaluate the search results are also given.

Key Words: Peptide mass fingerprinting; contaminants; modifications; protein identification; search parameters.

1. Introduction

For gel-based proteomics there are traditionally two strategies for protein identification by mass spectrometry (MS) analysis: (1) peptide mass fingerprinting (PMF) ***(1–5)*** and (2) sequence information obtained by tandem mass spectrometry (MS/MS). This chapter will focus on how to perform PMF searches and how to evaluate the obtained results.

The experimental data used in the PMF-based protein identification strategy are mass lists derived from matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS spectrum of an enzymatic-digested protein. Trypsin is commonly used as the enzyme (*see* Chapter 1; **Note 1**) and the masses of the resulting tryptic peptides are then used as query in PMF search programs and searched against lists of theoretical peptide mass values obtained by *in silico* digest of all the protein sequences in a given database. Each protein in the database is then given a score, dependent on how well the theoretical

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

mass list correlates with the experimental data. Those with the highest score are then most likely to be present in the sample in question. This sounds like a simple task if the protein in question (or a close homolog) is present in the database. However, several phenomena influence the matching of the experimental data to the theoretical ones, making PMF a method that is not always reliable. One problem is that some peptides tend to ionize on the expense of others and as a consequence some peptides will not be observed in the spectrum. The reasons for this are not completely understood, making it difficult to predict which peptides to expect and which not to expect. As a rule of thumb, peaks from arginine-containing peptides are more intense in the spectrum compared with lysine-containing (arginine-deficient) peptides ***(6)***, and arginine-deficient peptides with a mass less than 1000 are rarely observed. Another problem is that only peptides with a mass within the recorded mass range will be observed, e.g., between 700 and 3500. In addition, signals can be observed in the spectrum, which is not predicted by *in silico* digestion of the theoretical sequence. These can be modified peptides, which will not be matched in the PMF search unless the specific modification is accounted for. Alternatively, additional signals can arise from unwanted components in the sample, e.g., keratin from hair and dust. **Figure 1** lists additional reasons why a complete match of all peaks should not be expected.

Various software tools can be used for PMF searches ***(7–10)***. The outcome is most often a ranked list of proteins of which the top hit represents the protein/proteins most likely to be present in the samples analyzed. Many of the algorithms behind these search tools take into account features like the size of the database, the distribution frequency of a particular peptide mass within a given protein size, and the distribution of the mass accuracy. Other parameters, such as the number of missed cleavages sites allowed, the modifications expected, the sequence database against which the data should be searched, and the mass accuracy with which the search should be performed, has to be set by the user. The specificity of the search is determined by these parameters. If, for example, the mass tolerance is five times higher than the true mass accuracy of the data, then the specificity of the search is low, meaning that the risk of having false-positive hits is high. On the other hand, if the mass tolerance is five times lower than the true mass accuracy of the data, then the specificity of the search is too high, and the search program will not be able to report the true-positive hit. Therefore, one has to specify parameter values, which at the same time limit the number of proteins considered in the search, but still allow for the correct protein to fall within the limits. Only few of the search engines take into account other parameters like the ones previously described and in **Fig. 1**.

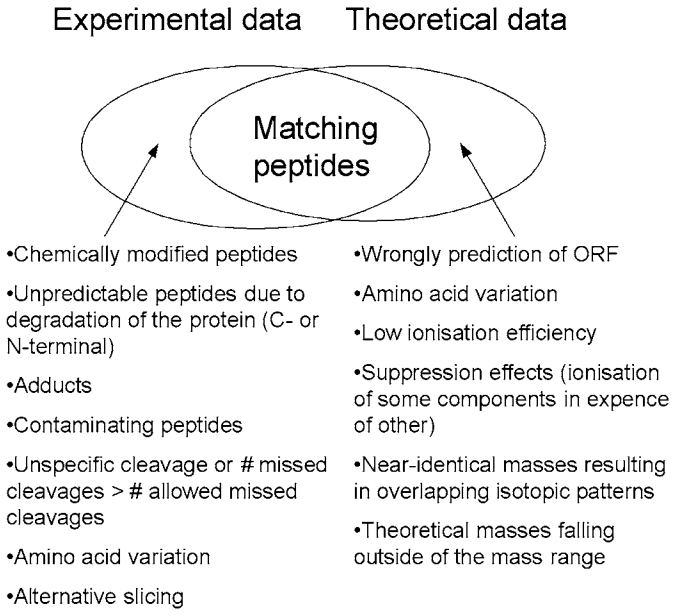


Fig. 1. There are several reasons why a perfect match should not be expected when experimental peptide mass fingerprinting data is compared to the theoretically predicted data. This figure lists some of the reasons.

With the introduction of MALDI tandem mass spectrometers it is now possible to reanalyze the sample in question at a later time-point in order to verify the PMF-obtained result or to study unassigned signals by MS/MS. Here, it is of advantage to initially identify the protein by PMF, confirm the identification by subjecting a few peptides for MS/MS, and then use the rest of the sample amount for analysis of the nonassigned peaks. These are often the ones containing interesting modifications.

This chapter is divided into four parts. The first part is identification and removal of signals originating from contaminants, like keratin-derived peptides. These signals do not belong to the protein in question and can therefore interfere with the specificity of the identification process. The second part deals with modifications. Most proteins are modified at specific residues, either as a result of posttranslational modifications or from artifacts introduced during sample handling. It is tempting to allow for such modifications, as the modified peptides will not be able to match any theoretical peptides otherwise. However, this can introduce a lot of random matches. Here, we introduce a simple strategy, which can be helpful in uncovering the existence of potential variable modifications, and thereby helpful in deciding which modification should be taken

into accounted for a specific search. The latter is illustrated here by the use of the software tool PeakErazor ***(11)***, which was introduced elsewhere (*see* Chapter 3). The third part deals with the problem of specifying parameters, such as mass accuracy, number of missed cleavages, and so on. As these parameters are common between the various search engines they will be described without referring to any specific search engine. The intention is to help the user to get high specificity without losing the real protein hit.

The last part of the chapter will deal with manual evaluation of search results in order to eliminate false-positives. Many proteome studies today are high-throughput studies, meaning that a lot of data is generated, making it impossible to evaluate all of the obtained results manually. However, the current status of search engines makes it advantageous to evaluate the results. This is especially important working with cross-species identifications where the hits are based on databases containing proteins from other species than the one analyzed. It is not possible to present a set of universal rules for such a validation strategy, however, a few useful guidelines can be given based on the experience of the author and other scientists. The validation steps presented here are therefore not quantifiable but we hope that they can help less experienced users in the evaluation process. For more information on protein identification by PMF (*see* **refs.** ***5*** and ***12–14***).

2. Materials

1. The program PeakErazor is used for identification and elimination of contaminants. This program is freely available at the General Protein/Mass Analysis for Windows (GPMAW) homepage; http://www.gpmaw.com and is simple to install and use (*see* Chapter 3).
2. Several different search engines are freely available for PMF, such as Mascot, Profound, MS-Fit, Aldente, and VEMS. Links to most of these can be found at the homepage of Expasy (http://www.expasy.org). VEMS can be found at http://www.yass.sdu.dk. As this chapter deals with parameters common for most of these programs and with problems generally experienced using any programs, this chapter will not focus on specific programs, but instead present a guide, which can be followed using any of the search engines.

3. Methods

3.1. Removal of Mass Values Unrelated to the Protein in Question

Peptides originating from other sources, like trypsin autodigested peptides, keratin-related peptides, and matrix-derived signals will interfere with the PMF search and increase the risk of false-positive hits or failure of the search. PeakErazor (introduced in Chapter 3) (*see* **Note 1**) can be used for extraction of such contaminants before submitting the peak lists to the PMF search program. Several approaches can be taken. Two of these are subsequently described.

3.1.1. Elimination of Peptide Mass Values Matching a List of Trypsin Autodigests Products and Keratin Peptides Commonly Observed by Others (see ***Note 2***)

1. Perform the first five steps in Chapter 3, **Subheading 3.2.** Choose keratin.lst as the Erazor list; a list of mass values against which the experimental mass values should be searched in order to identify potential contaminants.
2. If, for some reasons, one or more of the potential contaminants suggested by PeakErazor is believed not to be a true contaminant, please uncheck these in the list in the main window (*see* **Note 3**).
3. Copy the new list devoid of contaminants to clipboard.
4. Paste the new list into the PMF search engine of choice.

3.1.2. An Alternative to the Standard List of Contaminants (keratin.lst)

It is possible to collectively evaluate spectra from a specific project in order to identify project-specific contaminants. These can then be used as a basis for a new list of contaminants against which the project data can be matched.

1. Chose the page tab "background" in PeakErazor.
2. Paste a number of peak lists into the table by pressing the "Add spec" button for each list of mass values or upload a file containing several peak lists by pressing the "Peaklist set" button.
3. The mass values found to be common in the list are calculated by pressing the "Calculate" button. The criteria used in this calculation are defined by two parameters; the "combine at least" (which is the number of peak lists in which the mass value should be found) and the "prec. (ppm)" (which is the precision within which these mass values should be defined).
4. Save the list of shared mass values as a new Erazor list (e.g., yeastproject.lst).
5. Do as described in **Subheading 3.1.1.1.**, this time choosing the new and project specific Erazor list.

The latter approach is more accurate, as only project-specific contaminants are eliminated and random matching to keratin peptides observed by others is avoided (*see* **Note 4**). *See also* Chapter 3 for optimization of the Erazor lists in order to obtain more precise project-specific lists.

3.2. Identification of Potential Partial Modifications

A large variety of modifications can occur both in nature and during sample handling and the existence of a single peptide in several modified versions within the same spectrum is common. A well-known example is oxidation of methionine, where the modified and the nonmodified peptide have a mass difference of 16 Da. Other examples are modified N-terminal glutamine (pyroGlu, short for pyroglutamic acid, having a mass decrease of 17 Da compared with

the nonmodified peptide) and oxidized tryptophan. The latter will often be observed in several oxidation stages in the same spectrum along with a kynurenine-modified peptide (resulting from a loss of 28 Da from a double-oxidized tryptophan) ***(15)***. As a result, a characteristic pattern can be observed where the signals from tryptophan-containing peptides have a mass difference of +4 (kynurenine), +16 (mono-oxidized), +32 (double-oxidized), and +48 (triple-oxidized) relative to the nonmodified peptide ***(16,17)***. The kynurenine and the triple-oxidized peptide are not observed as frequently as the other three versions of the peptide.

The number of theoretical masses will strongly increase if numerous possible modifications are taken into account in a PMF search. As a consequence the risk of getting random hits (false-positives) will especially increase ***(7)*** (*see* **Note 5**). It is therefore recommended that the spectrum-deduced peak lists are checked for patterns indicating the presence of such partial modifications and that the various modifications are only accounted for in cases where such patterns are observed. PeakErazor can elucidate such patterns as described in the following:

1. Paste your peak list into PeakErazor from clipboard ("Read from clip").
2. Check the box named "Modifications" under display options.
3. Look for characteristic patterns like the +16 for oxidation, +17 for pyroGlu, and +4, +16, and +32 for tryptophan oxidation.

PeakErazor reveals two patterns for potential tryptophan oxidation in the example illustrated in **Fig. 2**.

3.3. Database Search

Before searching your data against a comprehensive database, it is important that the parameters are set correctly. These should reflect the history of the sample. One could ask some simple questions before searching the data: From which organism did my sample come from? Is the sample purified or separated on, for example, two-dimensional (2D)-gels? Did it undergo any special chemical treatments like reduction and alkylation? The answers to such questions can assist in defining the correct parameters for the PMF search in respect to database choice and allowed modifications. In the following, the settings for six different parameters will be discussed.

3.3.1. Choose Which Modifications Should be Allowed

If any of the patterns described in **Subheading 3.2.** is observed, then include the corresponding modification as a variable modification (*see* **Note 6**). In addition, acetylation of the N-terminal residue of the protein is recommended as a variable modification for eukaryotic proteins.

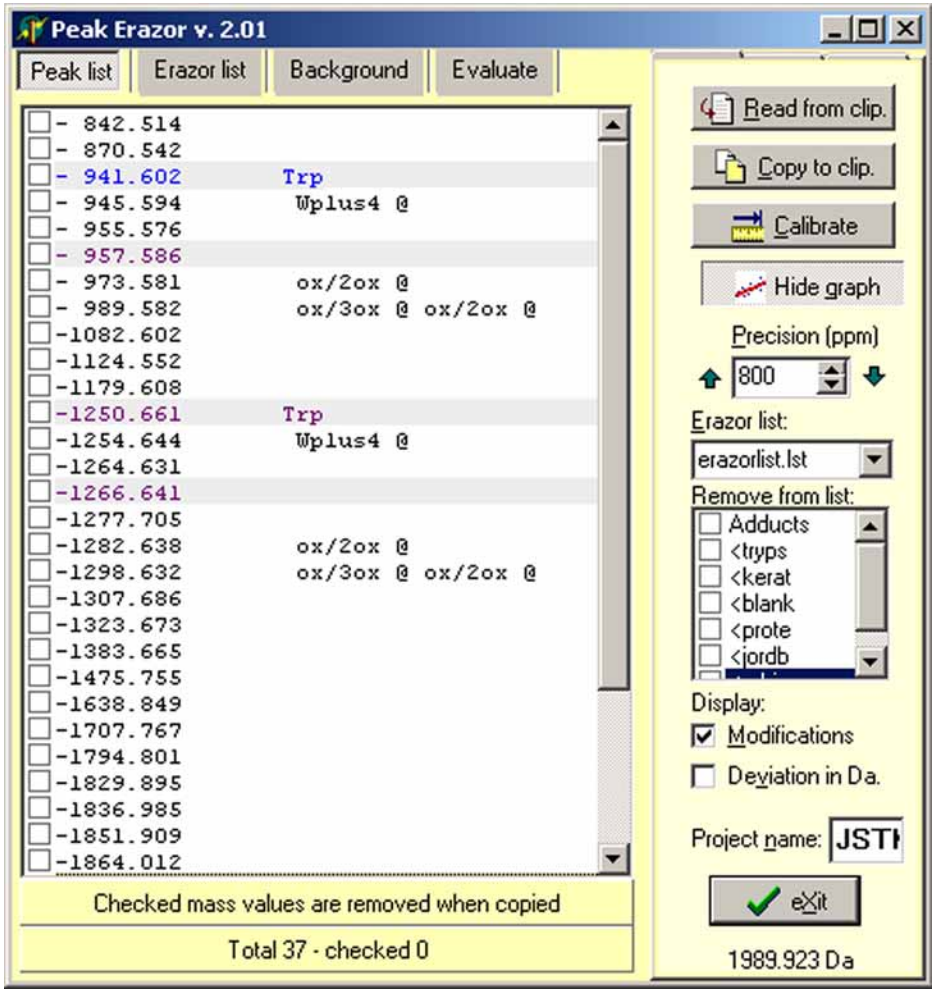


Fig. 2. By checking the checkbox “modification,” it is possible to get information on which partial modifications are likely to be present. The search parameters can be adjusted based on this information, in order to increase the chance of identifying the correct protein. In the example given here, two mass values (941.602 and 1250.661) are likely to represent peptides with tryptophan. This is based on the findings of three additional peptide signals having a mass increase by 4, 16, and 32, respectively. These are likely to represent the same peptide with different tryptophan modifications. It can be recommended to erase all of these peptides except, for example, the double-oxidized peptide, and then specify this modification as a fixed modification in the following database search.

Cysteins are often reduced and alkylated prior to analysis. The modification state of the cysteines depends on the alkylation reagent used; e.g., iodoacetamide treatment will introduce a carbamidomethylation of the cysteine residues (*see* **Note 7**).

3.3.2. Specify the Mass Tolerance

This value describes how well the experimental data has to fit to the theoretical data. The mass tolerance has to reflect the mass accuracy of experimental data. The mass accuracy is often given in either Dalton or in parts per million (ppm).

If a multipoint internal calibration has been performed as described in Chapter 3, then the obtained mass accuracy of the contaminants can make up a norm for the database search; if a mass accuracy of ±20 ppm is obtained for contaminants distributed over the entire mass range, then the mass accuracy of the sample-related mass values can be expected to be within the same range and set to, for example, 25 ppm.

3.3.3. Specify the Enzyme

The most commonly used enzyme for PMF is trypsin (*see* **Note 8**; Chapter 1).

3.3.4. Specify the Number of Missed Cleavage Sites

Some proteases have the tendency to generate peptides containing internal cleavage sites missed by the enzyme. Depending on the efficiency of the enzyme, it is advantageous to allow the search program to take such missed cleavage sites into account. For trypsin it is recommended to allow for one missed cleavage site in each peptide. If the search fails (i.e., no significant and reliable hit is reported by the search engine) it is a good idea to change this parameter to 0 or 2 in order to reflect the actual efficiency of the enzyme. If the parameter is set to 2 or more, please *see* **Subheading 3.4.2.** for a manual evaluation of the search result.

3.3.5. Chose a Comprehensive Database Against Which the Search has to be Matched

The main demand for the database is that the protein, or close homolog, has to be present in the database. A large database like the one from National Center for Biotechnology Information (NCBI) can be chosen. This database covers all the publicly available sequences and should therefore have the largest possibility of containing the protein in question (*see* **Note 9**). The size of the database can be narrowed down by restricting the search to a single taxon or species (*see* **Note 10**). This will not only reduce search time, but will also reduce the chance

of identifying proteins by cross-species identification. Such identification based on homologous proteins from other species can be helpful for organisms for which the entire genome is not sequenced.

3.3.6. Specify the Estimated Isoelectric Point and Molecular Weight Values

As most studies involving PMF searches also involve the use of 2D gels for protein separation, the isoelectric point and molecular weight estimated for the protein in question can be utilized as additional restrictions in the search (*see* **Note 11**). Not all search engines takes these estimated values into account.

3.4. Inspection of the Protein Hits in Order to Evaluate the Search Result

The protein reported as the highest scoring protein with a significant score is most likely the one to be present in the sample of interest according to the PMF search program used. However, how does one evaluate whether the protein reported by the program is actually a reasonable suggestion and not a false-positive? Some rules-of-thumb are presented, which can be used in order to increase the confidence of a search result.

3.4.1. How is the Mass Accuracy Distribution?

The mass distribution has to be either linear or curved, reflecting the calibration obtained on the instrument. If contaminants and/or tryptic autodigest products were eliminated (*see* Chapter 3), then check whether the mass accuracy distribution of the contaminants is comparable to the one obtained in the search hit. If a few mass values are found to be deviating considerably from the rest of the masses, try to remove these masses and redo the search (*see* **Note 12**).

3.4.2. Are Any Overlapping Peptides Identified? Are the Missed Cleavages Observed of the Expected Type?

Overlapping peptides are a consequence of partially missed cleavage sites and adds to the overall confidence of the hit. However, the likeliness of such missed cleavage sites depends on the residues adjacent to the cleavage site. In the case of trypsin, the majority of the missed cleavage sites contain either an acidic residue (glutamic acid or aspartic acid) or a basic residue (lysine or arginine) adjacent to the potential cleavage site ***(18,19)***. Therefore, the likeliness of the missed cleavage sites found and, thereby, the confidence of the search can be evaluated by studying the peptides containing the missed cleavage sites (*see* **Note 13**). This is especially important for search results obtained by allowing two or more missed cleavage sites.

3.4.3. Are any Peptides Carrying Suspected Partial Modifications Identified? What About the Corresponding Nonmodified Peptide?

For modifications caused by sample preparation, such as N-terminal pyroglutamic acid and the oxidation of, for example, methionine, both the modified and the nonmodified peptide are expected to be observed. The finding of both peptides adds to the confidence of the search result. For oxidation of methionine, a metastable signal resulting from loss of the methanesulfonic acid (CH_3-S-OH = 64 Da) group is also expected. The position of the corresponding peak on the mass spectrum depends on the specific parameters of the MALDI-TOF instrument. Observation of this metastable signal further adds to the confidence of the search result.

3.4.4. Do the Identified Peptides Account for the Most Intense Peaks in the Spectrum?

The intensities in the MALDI mass spectrum do not correlate directly to the amount of peptide in the sample because of different factors like different ionization efficiencies, suppression effects, and various degrees of missed cleavages. Therefore, the intensities of the individual signals in the mass spectrum are not taken into account in most PMF search programs today. However, it is relevant to study whether or not the most intense peaks in the spectrum have been accounted for by a given search result. Few intense peaks might have been left unassigned resulting from modifications not taken into account; however, to be a reliable hit, it should be possible to match most of the intense peaks. Additionally, it is known that arginine-containing peptides ionize more efficiently than lysine-containing peptides, often resulting in higher intensity of the arginine-containing peptides compared to the lysine-containing peptides. This general tendency could also be considered in an examination of the reliability of the search result.

3.4.5. What is the Sequence Coverage Obtained Compared With the Expected?

Sequence coverage is the percentage of the protein covered by the matching peptides. Often a new user will ask the question, how many peptides should be matched and how large should the sequence coverage be for a valid hit? However, these numbers depend on the protein in question, so no clear answer can be given. Working with large proteins, we would expect a relatively large amount of peptides to be matched. On the other hand, if the protein in question turns out to constitute only a fragment of the theoretical protein sequence in the database, i.e., because of degradation, then the number of peptides matching can be low and the sequence coverage will appear low even though the coverage in the area constituted by the fragment is high. Working with small proteins,

few peptides might be matched and still result in a relative high coverage. It has recently been suggested that the reliability of the identification can be revised by multiplying the sequence coverage (in percentage) with the number of mass values matching ***(14)***. If the resulting value is greater than 300 the hit is considered valid. It is possible to calculate an expected coverage by excluding all theoretical peptides with a mass outside the mass range in which the spectrum is obtained (typically 700–3000 Da). However, the distribution of possible cleavage sites in the protein of question will also influences such an expected coverage. Infrequently, distributed cleavage sites will result in peptides too large to be observed in the spectrum, whereas frequent cleavage sites will result in many small peptides. The later might be observed as part of peptides containing missed cleavage sites. A way of taking these missed cleavages into account would be to only allow the expected missed cleavages mentioned previously. Peptides originating from the signal peptide of secreted proteins are not expected in the mature protein analyzed, so this part of the theoretical sequence should also be excluded when calculating the expected sequence coverage for the protein in question. A comparison of the experimentally obtained sequence coverage to this calculated expected coverage offers a more realistic view of how well the data correlates.

3.5. Conclusion

In the end it should be noted that large proteins are often overrepresented in the search results because of random matching of peptides to these proteins. Some search engines accounts for this bias in the search algorithms, however, one should always be sceptical toward such large proteins, especially if the part covered by the peptides does not correlate to an expected mass of the protein. Many of these false-positive hits can be avoided by taking the molecular mass estimated from, for example, a 2D-PAGE, into account.

4. Notes

1. Identification of signals arising from contaminating components like keratin is further described in Chapter 3, where it is demonstrated how these signals can be used as internal calibrants. The subject in the present chapter is to eliminate the contaminants prior to the PMF search.
2. An Erazor list, containing mass values corresponding to the masses of various contaminants, is available for download at the same web page as PeakErazor. This list is based on observations done in our laboratory and by other laboratories (*see* Chapter 3).
3. If several mass values are matching the theoretical mass values of contaminants and one or more of these have a mass error, which is deviating from the trend made up of the rest of the contaminants, i.e., it is an outlier, then this mass value is most likely not a true contaminant and should be unchecked.

4. Removing unknown contaminants using this strategy is not without problem, as many projects contain the same protein in several samples. If 10% of the samples in a given project contain the same protein and therefore have near-identical peak lists, the peptides from this protein can accidentally be taken as contaminants, and therefore be removed from the project together with the real contaminants. One suggestion would be to use the obtained list of contaminants (the Erazor list) as query in a PMF search in order to identify such proteins accidentally taken as contaminants. It should be emphasized that peaks should only be rejected if they match the theoretical mass within the accuracy with which the mass spectrum is obtained. A peak rejection feature parallel to the one presented here for PeakErazor is incorporated into several of the PMF search engines ***(14,20)***. In these programs peak rejection will automatically be performed prior to the database search.
5. In many cases, only one or a few peptides will contain a specific modification and omitting these peptide masses from the search may or may not disrupt the search result leaving the protein of interest as a false-negative. As a consequence, the user has to counterbalance the risk of having false-positives or false-negative. Highly modified proteins like collagen are impossible to identify by PMF unless the particular modification (hydroxyproline in this case) are taken into account. However, as long as the protein is not identified, it is impossible to know which specific modification is required. However, studying, for example, bone proteins, it can turn out to be of advantage to include hydroxyproline as a variable modification in searches otherwise found to "fail." The hydroxyproline will have a mass difference of 16 Da compared with the normal proline, and as each peptide arising from collagen has the possibility of containing more than one proline, a characteristic pattern of signals spaced with 16 Da will be observed. Be aware that this pattern corresponds to the pattern observed by multiple oxidations of other residues as well. This example illustrates how knowledge regarding the biology connected to the sample of interest can turn out fruitful in the protein identification process.
6. The double-oxidation modification of tryptophan (formylkynurenine) is not a default option in, for example, the search engine Mascot. It is, however, possible to include additional modifications through the "set search default" page, including formylkynurenine.
7. The –SH group of cysteines are known to be very reactive and will form adduct product with free chemicals like acrylamide, which is commonly used in casting 2D-gels ***(21)***. It has been estimated that polymerization of acrylamide in gels rarely exceed 90% resulting in at least 30 m*M* of free acrylamide ***(22)***. As a consequence, some cysteines may be modified by propionamide. Some may even exist as unmodified cysteins or as disulfide bridges. If full explanation of all the peaks in the spectrum is desired, it can therefore be necessary to take several modifications into account, but be aware that this will increase the risk of false-positive hits.
8. Other enzymes like LysC and ArgC (cleavage at the C-terminal side of lysine and arginine, respectively) generate larger peptides, which can be of advantage for proteins containing many lysines and/or arginines.

9. The databases need to be fairly correct in order to identify the relevant protein. This is especially a problem when searching protein databases derived from more or less automated translation of nucleotide databases (first described by James et al. ***[23]***). The problems are numerous: the quality of the nucleotide-based databases are varying, especially expressed sequence tag sequences derived from cDNA suffer from a high error rate. Gene prediction is a difficult task, especially for small genes for which clear differentiation between true and random stretches of open reading frames is not yet possible. In addition, the rules for alternative splicing are very complex and not yet fully understood. An alternative is to search whole genomic data independently of the predicted open reading frames by scanning the genome with the PMF data. Algorithms for such genome-based PMF searches are under development ***(24)***. Databases like SwissProt and Uniprot are manually evaluated and should be without the previously mentioned problems. However, the number of proteins in the database is lower containing only well-analyzed proteins. One advantage of such well-annotated databases is that you can specify your search not only to species but also to keywords appearing in the database entries, for example, restricting the search to plasma proteins (MultiIdent ***[25]***).
10. Restricting a search toward an organism that is not fully sequenced will increase the risk of failure. This can be partly overcome by allowing the search algorithm to search against proteins from closely related organisms. Again, the search hits have to be evaluated carefully if such cross-species hits are allowed.
11. Experimental values estimated from a gel are very likely to diverge strongly from the theoretical values and the use of these dubious values does not necessarily result in increased identification success. There can be several reasons for this: (1) for example, the protein might be highly modified by glycosylation, (2) the theoretical protein might be wrongly annotated caused by incorrect prediction of the reading frame, and (3) the protein studied might be a fragment of the theoretical protein found in the database (a truncated protein). The search engine Mascot (www.matrixscience.com) has tried to overcome the latter problem by introducing a “sliding window” (first suggested by Yates ***[4]***).
12. It is possible to recalibrate a peak list based on mass values identified as belonging to the protein in question. If the mass accuracy of the original peak list is low, then recalibration and research can increase the score and, thereby, the confidence of the search result.
13. It should be stated here that whether or not the digestion reaches completion is dependent on other issues as well; the cleavage site can be inaccessible to the enzyme, e.g., from steric hindrance, or the enzyme/substrate ratio that can be to low, both cases resulting in unpredictable missed cleavage sites.

References

1. James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1993) Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Comm.* **195,** 58–64.

2. Mann, M., Hojrup, P., and Roepstorff, P. (1993) Use of mass-spectrometric molecular-weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22,** 338–345.
3. Pappin, D. J. C., Hojrup, P., and Bleasby, A. J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Current Biol.* **3,** 327–332.
4. Yates, J. R., Speicher, S., Griffin, P. R., and Hunkapiller, T. (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **214,** 397–408.
5. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from 2-dimensional gels by molecular mass searching of peptide-fragments in protein-sequence databases. *Proc. Natl. Acad. Sci. USA* **90,** 5011–5015.
6. Krause, E., Wenschuh, H., and Jungblut, P. R. (1999) The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Anal. Chem.* **71,** 4160–4165.
7. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567.
8. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (+/– 10 ppm) in protein identification strategies employing MS or MS MS and database searching. *Anal. Chem.* **71,** 2871–2882.
9. Zhang, W. Z. and Chait, B. T. (2000) Profound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* **72,** 2482–2489.
10. Matthiesen, R., Bunkenborg, J., Stensballe, A., Jensen, O. N., Welinder, K. G., and Bauw, G. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* **4,** 2583–2593.
11. Hjernoe, K. and Hojrup, P. (2004) PeakErazor: A Window-based program for improving peptide searches. In: *Methods in Proteome and Protein Analysis*, (Kamp, R. M., Calvete, J. J., and Choli-Papadopoulou, T., eds.), Springer-Verlag, Heidelberg, Germany, pp. 359–369.
12. Jensen, O. N., Podtelejnikov, A. V., and Mann, M. (1997) Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal. Chem.* **69,** 4741–4750.
13. Jensen, O. N., Larsen, M. R., and Roepstorff, P. (1998) Mass spectrometric identification and microcharacterization of proteins from electrophoretic gels: strategies and applications. *Proteins* **(Suppl 2),** 74–89.
14. Thiede, B., Hohenwarter, W., Krah, A., et al. (2005) Peptide mass fingerprinting. *Methods* **35,** 237–247.
15. Finley, E. L., Dillon, J., Crouch, R. K., and Schey, K. L. (1998) Identification of tryptophan oxidation products in bovine alpha-crystallin. *Protein Sci.* **7,** 2391–2397.
16. Suckau, D., Resemann, A., Schuerenberg, M., Hufnagel, P., Franzen, J., and Holle, A. (2003) A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. *Anal. Bioanal. Chem.* **376,** 952–965.

17. Bienvenut, W. V., Deon, C., Pasquarello, C., et al. (2002) Matrix-assisted laser desorption/ionization-tandem mass spectrometry with high resolution and sensitivity for identification and characterization of proteins. *Proteomics* **2,** 868–876.
18. Thiede, B., Lamer, S., Mattow, J., et al. (2000) Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Comm. Mass Spectrom.* **14,** 496–502.
19. Monigatti, F. and Berndt, P. (2005) Algorithm for accurate similarity measurements of peptide mass fingerprints and its application. *J. Amer. Soc. Mass Spectrom.* **16,** 13–21.
20. Chamrad, D. C., Koerting, G., Gobom, J., et al. (2003) Interpretation of mass spectrometry data for high-throughput proteomics. *Anal. Bioanal. Chem.* **376,** 1014–1022.
21. Hamdan, M., Galvani, M., and Righetti, P. G. (2001) Monitoring 2-D gel-induced modifications of proteins by MALDI-TOF mass spectrometry. *Mass Spectrom. Rev.* **20,** 121–141.
22. Patterson, S. D. and Aebersold, R. (1995) Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis* **16,** 1791–1814.
23. James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1994) Protein identification in DNA databases by peptide mass fingerprinting. *Protein Sci.* **3,** 1347–1350.
24. Giddings, M. C., Shah, A. A., Gesteland, R., and Moore, B. (2003) Genome-based peptide fingerprint scanning. *Proc. Natl. Acad. Sci. USA* **100,** 20–25.
25. Wilkins, M. R., Gasteiger, E., Wheeler, C. H., et al. (1998) Multiple parameter cross-species protein identification using MultiIdent-a world-wide web accessible tool. *Electrophoresis* **19,** 3199–3206.

5

Generating Unigene Collections of Expressed Sequence Tag Sequences for Use in Mass Spectrometry Identification

Jeppe Emmersen

Summary

Expressed sequence tag sequences remain the largest resource of DNA sequence for most organisms despite recent advances in genome sequencing. These sequences are short, fragmented versions of the expressed genes. By DNA sequence assembly, the fragments can be assembled into contiguous DNA sequences that are better suited for protein identification by mass spectrometry.

Key Words: Expressed sequence tag; EST; assembly; unigenes; protein prediction; annotation.

1. Introduction

Identification of proteins using mass spectroscopic methods relies on the existence of reliable protein sequence databases. For most organisms, however, the vast majority of sequence information for any given organism remain as cDNA sequences in the form of either singlepass sequenced expressed sequence tags (ESTs) or (more rare) complete full-length cDNA sequences. EST sequences are deposited in a special subdatabase of GenBank, dbEST, and the typical EST sequence is approx 400–500 bases long ***(1)***. It has previously been shown that using the dbEST database for protein identification by peptide mass fingerprinting was feasible as a last resort ***(2)***.

ESTs are either sequenced from the 5′ or the 3′ end of cDNA clones, with the majority of ESTs deposited in the databases being 5′ end ESTs. The reason is that the 5′ end of a cDNA normally contains a larger proportion of coding sequence than 3′ ESTs. It turns out that many EST sequences are derived from cDNA clones truncated at the 5′ end. Although a drawback in other application, the truncation turns out to be an advantage when ESTs are assembled into a

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen

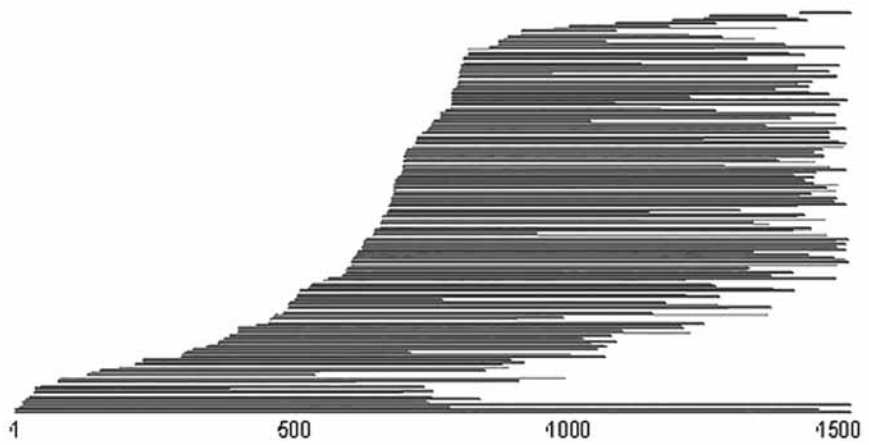


Fig. 1. Assembly of 133 expressed sequence tags (ESTs) from the patatin gene of potato into one contig–Unigene cluster. Each EST is represented by one line. The otherwise undesired 5′ end cDNA truncation is an advantage in this context, as it provides complete coverage of the gene. The ESTs were assembled using three full-length cDNAs as scaffold, shown as the large lines at the bottom of the figure.

unigene collection. This enables genes of larger sizes to be constructed from short EST sequences, which in turn enables better identification of proteins using mass spectrometry (MS) (**Fig. 1**).

EST sequences obtained from public databases should, in principle, be free from contaminating sequences. There is no standard for cleaning EST sequence sets, however, and it is recommended to do a precleaning of EST sequence sets obtained from public sources to remove contaminating sequences from *Escherichia coli*, lambda, and vector sources. If EST sequences are not cleaned and vector inserts masked, the final Unigene collection will harbor the contaminants and the assembly of EST sequences will not be optimal if vector sequences allow nonrelated sequences to be joined.

The benefits of EST sequence assembly can be summarized as follows:

1. Reduced sequence set—reduced search space and search time.
2. More peptides can be identified—peptide sequence located at ends of ESTs are joined.
3. Better accuracy of DNA sequence through consensus sequence—better accuracy of protein sequences. Improves with redundancy of transcripts.
4. Detection of chimera (nonrelated sequences joined as one)—depending on algorithm.
5. EST sequences are free of intron sequences (in most cases).

The only real disadvantage of EST assembly is that sequence variants may be lost during consensus sequence construction. These can be found at a later

stage after protein annotation by examining the multiple sequence alignment of each EST contig. If using TGICL for EST assembly, the alignments are part of the output.

The disadvantage of losing sequence polymorphism information during consensus sequence generation can be circumvented to some extent by using sequences only from the tissue of interest. Thus, if the proteome study is focused on a particular plant cultivar for example, the best Unigene collection would be built by selecting only ESTs from the particular cultivar, if available.

1.1. Prediction of Protein Sequence From Low-Fidelity DNA Sequences

A good DNA sequence set is the first requirement for use in MS identification of proteins. The second requirement is a good translation to protein sequences. The relatively high error rate in EST sequencing is carried over to the protein sequence, though at a lower level because of codon degeneracy at the third position. A wrong DNA base may alter the corresponding amino acid in a tryptic peptide, rendering this peptide lost in peptide mass fingerprinting. For MS/MS sequencing, the loss is not so critical, as the remaining sequence is retained and can be found by nonexact methods, such as BlastP. The worst base-calling error is either a wrong insertion or deletion (indel), which alters the frame of the protein translation and subsequently the remaining sequence will be wrong.

For unattended DNA translation, there are a number of algorithms to perform the translation. The software VEMS uses a traditional algorithm based on simple open reading translation (longest open reading frame [ORF]). Using VEMS to make the translation, it is possible to select the number of longest ORFs chosen for each DNA sequence. Any indels will break the ORF, but this may not be a problem if the real ORF is among those chosen for the translation. For instance, if a deletion occurs at 100 bases in a 1200 base long transcript, the real ORF shifts from frame 3 to frame 2. The result is the original ORF of 400 amino acids being split into two ORFs of 40 and 370 amino acids. If the user selects the three largest ORFs to be stored for this transcript and there are four nonsense ORFs longer than 40 amino acids but shorter than 400 amino acids, the short 40 amino acid sequence is lost. If the deletion occurs in the middle of the transcript, the chance of finding the correct sequence grows.

To compensate for indels in DNA sequences, programs have been written that can detect frameshifts during the translation process and correct the error. One such program is Framefinder, which is part of the Estate package for EST analysis ***(3)***. Framefinder uses local hexamer (stretches of six nucleotides) usage frequencies estimated from known coding sequences of a given organism to predict the location of insertions or deletions. A dynamic programming algorithm, similar to DNA alignment algorithms, such as blast or Smith-Waterman,

Table 1
Some General Precomputed Expressed Sequence Tag Assembly Resources

Source	Organisms	Web address
TIGR	85	www.tigr.org/tdb/tgi/
Sputnik	60	sputnik.btk.fi
UniGene	52	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene

is used to find the most probable translation given the hexamer frequencies estimated for each organism. Both local and global paths can be used to find the correct translation. This way, the algorithm is able to compensate for frameshifts when translating the DNA sequence.

1.2. If You Do Not Want to Do It Yourself

There are a number of publicly available repositories for clustered and assembled EST collections (**Table 1**). The Institute of Genomic Research (TIGR) publishes their EST assemblies for download on their website www.tigr.org <http://www.tigr.org/>. Currently (June 2006), there are assemblies available from 88 species at the TIGR website ***(4)***. The program TGICL for EST assembly has been made publicly available ***(5)***. Another source of precomputed EST assemblies is the UniGene database located at GenBank ***(6)***.

The major difference between the *TIGR* gene indices and the GenBank UniGene collection is the way each EST cluster is represented. A TIGR EST cluster is treated in the same way as genome shotgun sequences, and each cluster is subject to a final assembly process, whereby a consensus sequence is generated, representing all EST sequences in the cluster. Thus, any sequence polymorphism information is lost in this process. A unigene EST cluster is represented by the best EST sequence, thus no consensus sequence is produced. This means that a unigene cluster sequence will not represent the full coding sequence of a transcript, unless the gene is short or there is a full-length mRNA sequence as part of an EST cluster.

As ESTs are selected randomly from the pool of cDNA, the most highly expressed genes will be represented by many EST sequences, leading to a high level of redundancy. If searching with raw EST data in the annotation process, this will lead to longer processing times.

1.3. Annotation of Sequences

EST unigene sequences can be annotated automatically using BlastX against a protein database, such as the nonredundant protein database (nr) from GenBank. In the simplest annotation scheme, only the annotation of the best

Table 2
Software for Expressed Sequence Tag Processing

Name	Description	Web address
TGICL	Assembly	www.tigr.org/tdb/tgi/software/
Mira	Assembly	www.chevreux.org/projects_mira.html
Phrap	Assembly	www.phrap.org/
Cap3	Assembly	genome.cs.mtu.edu/cap/cap3.html
Stack	Assembly	www.sanbi.ac.za/Dbases.html
JESAM	Assembly	corba.ebi.ac.uk/EST/jesam/jesam.html
Lucy	Preprocessing	http://www.tigr.org/software/
SeqClean	Preprocessing	http://www.tigr.org/software/
blast-multi.pl	Annotation	je@bio.aau.dk
Blast	Alignment	ftp://ftp.ncbi.nih.gov/blast/

Blast hit is recorded. The annotation can then be stored in a separate file recording the sequence and the corresponding hit. The annotation can also be stored in the header of the fasta file.

It is also possible to annotate the predicted protein sequences directly using BlastP against the nr protein database. However, if only ORF prediction is performed with extraction of the three longest ORFs, the BlastP annotation may be wrong as BlastP only searches one protein sequence, whereas BlastX searches all six reading frames of the DNA sequence. Thus, it is better to annotate the DNA sequence and then carry the annotation to the protein sequences afterwards if the database match is good (E-value > 1E-20). If a protein sequence predictor like Framefinder is used with the annotated DNA sequences, the header information is also retained.

This chapter will describe how to set up free software to perform EST assembly and sequence annotation of the resulting Unigene collection using a Linux platform. Commands must be executed on the command line (shell) and are highlighted in **bold**. Options for the command are highlighted in *italics*.

2. Materials

You will need one computer with an ×86 Linux distribution. The described programs used in the methods are available free of charge (*see* **Table 2**).

Should it prove impossible to find the software listed because of website reorganizations or other difficulties, the author of this chapter can be contacted for information on how to obtain the software.

3. Methods

The following method will describe how to use the programs TGICL, Seqclean, and Framefinder and blast-multi.pl to assemble, annotate, and create

protein files from a collection of EST sequences. **Table 2** lists the current web addresses where the software can be obtained.

3.1. Setting Up Software

To install the software make a folder for assembly of the EST sequences:

mkdir *est-assembly.*

Then download the software into the *est-assembly* folder:

1. TGICL: <http://www.tigr.org/tdb/tgi/software/>.
2. Framefinder: <http://www.ebi.ac.uk/~guy/estate/> or by request to je@bio.aau.dk.
3. blast-multi.pl: mail to je@bio.aau.dk.
4. Sequence databases for blast: <ftp://ftp.ncbi.nih.gov/blast/db/>.

3.2. Blast Programs and Sequence Databases

The first step is to install the Blast software. It is important to remember that Linux file names are case sensitive. First log-in as root. Set up the path to blast environment:

mkdir */usr/local/bioinfo*

Where the following folders are made using:

mkdir */usr/local/bioinfo/bin/*

mkdir */usr/local/bioinfo/data/*

Move the blast the blast archive to the bioinfo/bin/ folder and extract the program:

tar *–xvf blastsoftware*

Set the environment variables for the particular Blast path in the file ***/etc/.profile*** *by* appending the following lines to /etc/.profile using a text editor:

PATH = $PATH:/usr/local/bioinfo/bin/:usr/local/bioinfo/data/

BLASTMAT = "/usr/local/bioinfo/bin/data"

BLASTFILTER = "/usr/local/bioinfo/bin"

BLASTDB = "/usr/local/bioinfo/data"

export BLASTMAT BLASTFILTER BLASTDB PATH USER LOGNAME MAIL HOSTNAME HISTSIZE INPUTRC

Also *see* **Note 1**.

To obtain the database files for blasting against the nonredundant protein database, download the formatted database files at the GenBank Blast web address: <ftp://ftp.ncbi.nih.gov/blast/db/> to the */usr/local/bioinfo/data/folder*.

To format custom databases such as the *E. coli* genome sequence for Blast searching, the following command is executed:

formatdb *–i E-coli-sequence-file b –p F –o T*

where *E-coli-sequence-file* is the fasta formatted database file. This will generate a DNA database for searching. The generated files must be placed in the ***/usr/local/bioinfo/data/*** folder. The *E. coli* genome sequence database is obtained from <ftp://ftp.ncbi.nih.gov/blast/db/FASTA/> and the Univec database from <ftp://ftp.ncbi.nih.gov/pub/UniVec/UniVec_Core>.

formatdb *–I UniVec_Core b –p F –o T*

3.3. Setting Up TGICL for Clustering and Assembly of EST Sequences

Copy the TGICLI and Seqclean packages into the ests-assembly folder and unpack them:

tar *xfvz* TGICL.tar.gz and **tar** *xvfz seqclean.tar.gz*

3.4. Installing Framefinder Software

The framefinder program is a little tricky to install as only the source files are provided. Transfer the Framefinder software package to the *est-assembly* folder. Then unpack the programs and make the binaries:

tar *–xfvz*, then **cd** *estate* and finally **make**. Move the binary files in the bin folder to your ***homefolder/bin*** or somewhere in the path.

The software is now ready for use. Remember to put the TGICL program in the same folder as the sequence file. *See* **Note 2** on program and user rights.

3.5. Making EST Assemblies

To clean your EST files, use the command:

seqclean *est_file-v/usr/local/bioinfo/data/UniVec_Core.*

The output are two files: the processed fasta file, est_file.clean and a processing report, est_file.cln, with details on the preprocessing.

The fasta file is then used for the assembly (*see* **Note 3**):

.\TGICL est_filedb

After the assembly has been performed, the important files generated are *est_file.singletons* and */asm1/contig* and */asm1/singlets*. The contig file is located in the asm1 folder generated during the assembly process and contains the

assembled sequences from two or more EST sequences. The singlets file in the same folder contains those sequences, which initially aligned with another sequence, but was later rejected during the assembly phase. The singlets in the singleton file are not immediately available in fasta format but can be extracted with the following command:

cdbyank *est_filedb.cidx < est_filedb.singletons > singleton.seqs*

When both the contig file, the singlets file and the singleton file have been gathered in one directory, the following command will concatenate both files into one to generate the final Unigene collection:

cat *contig_file singletons.seq > Unigene.fsa.*

See **Note 4** if you have access to original sequence data.

3.6. Performing Protein Sequence Prediction Using Framefinder

The next step is to generate protein sequence files from the Unigene collection using the protein prediction software framefinder. Before DNA sequences can be translated using framefinder, the statistical framework for protein prediction needs to be set up. The first step is to obtain a set of full-length mRNA sequences from GenBank for the organism of interest. The sequences need to be in GenBank format and not Fasta format (*see* **Note 4**).

The coding sequence is then extracted with command:

flat2coding *-o 'Name of organism' -d 'name of GenBank file' > coding.fasta.*

The entry "Name of organism" should match the organism identifier exactly as written in the GenBank files.

The hexamer frequencies are calculated using the commands:

fasta2usage *-w 6 -j 3 -d coding.fasta > coding.wordcount*

calcwordprob *-w coding.wordcount > coding.wordprob*

The protein sequences can then be constructed using framefinder:

framefinder *– w coding.wordprob -d Unigene.fasta > Unigene.predicted_protein.fasta*

The output file may contain some lines that are not in FASTA format at the beginning of the file. These can edited out using a text editor (*see* **Note 5**).

3.7. Annotate the Unigene Collection to the nr Database

perl blast-multi.pl *Unigene.fsa*

which will annotate all DNA sequences using BlastX against the nr database located in the folder */usr/local/bioinfo/data/*

The protein files can be annotated with the same program using the –p blastp switch

perl blast-multi.pl –p blastp *predicted_protein.fasta.*

Two files are produced: the annotated fasta file, where the annotation information for the best Blast hit is added at the end of the fasta file header. The name of the database used for annotation is prepended to the fasta file name. The second file is a report, producing detailed information for each annotation, such as length of query, length of database match, coverage of alignment for both Unigene sequence, and database sequence. Thus, it is easy to check the "biology" of the annotation and not just rely on E-value numbers.

4. Notes

1. If only one user is going to use the Blast software, the Blast environment can be set up by creating an .ncbirc file with the following lines using a texteditor in the user's home directory:

   [NCBI]

   Data = "/usr/local/bioinfo/data/"

   The Blast executables still need to be in the system path.

2. When setting up the software, it is important that all users can access the programs and the data. This can be done by changing the access rights using the following command (assuming the files were put in the *bioinfo* path and user is root):

   chmod *666 /usr/local/bioinfo/* .

3. If sequence processing is performed on a Windows OS machine but the sequence assembly and protein sequence prediction is performed on a Unix OS, care has to taken when transferring files, as text files are not equal between the two systems when it comes to line breaks. Windows uses a carriage return and a linefeed (\r\n) whereas Unix only uses a linefeed (\n). To convert from Windows to Unix format, each text file can be modified with the following command:

   tr *—d '\15\32' <windows-file.txt > unix-file.txt*

   Otherwise, the Cap3 assembly algorithm of TGICL will not be able to read the sequence and quality files properly.

4. The example given for EST assembly assumes that sequences were obtained from dbEST. Consequently only the fasta-formatted DNA sequence is available and the confidence of each base cannot be estimated. If the original sequence chromatogram (trace) can be obtained, it is possible to estimate the probability of each base being correct if using the base calling program Phred ***(7)***.

5. When framefinder runs, it may insert an X instead of an amino acid letter in the sequence in places where an indel was detected.

References

1. Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993) dbEST—database for "expressed sequence tags". *Nat. Genet.* **4,** 332–333.
2. Mann, M., Hendrickson, R. C., and Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70,** 437–473.
3. Slater, G. C. (1999) *Expressed sequence tag analysis tools.* PhD thesis, March 3, 2006. Human Genome Mapping Project RC, Hinxton, Cambridge, UK.
4. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., and Quackenbush, J. (2000) An optimized protocol for analysis of EST sequences. *Nucl. Acids Res.* **18,** 3657–3665.
5. Pertea, G., Huang, X., Liang, F., et al. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19,** 651–652.
6. Wheeler, D. L., Church, D. M., Federhen, S. L., et al. (2003) Database resources of the national center for biotechnology. *Nucleic Acids Res.* **31,** 28–33.
7. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194.

6

Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching

Alexey I. Nesvizhskii

Summary

The shotgun proteomics strategy, based on digesting proteins into peptides and sequencing them using tandem mass spectrometry (MS/MS), has become widely adopted. The identification of peptides from acquired MS/MS spectra is most often performed using the database search approach. We provide a detailed description of the peptide identification process and review the most commonly used database search programs. The appropriate choice of the search parameters and the sequence database are important for successful application of this method, and we provide general guidelines for carrying out efficient analysis of MS/MS data. We also discuss various reasons why database search tools fail to assign the correct sequence to many MS/MS spectra, and draw attention to the problem of false-positive identifications that can significantly diminish the value of published data. To assist in the evaluation of peptide assignments to MS/MS spectra, we review the scoring schemes implemented in most frequently used database search tools. We also describe statistical approaches and computational tools for validating peptide assignments to MS/MS spectra, including the concept of expectation values, reversed database searching, and the empirical Bayesian analysis of PeptideProphet. Finally, the process of inferring the identities of the sample proteins given the list of peptide identifications is outlined, and the limitations of shotgun proteomics with regard to discrimination between protein isoforms are discussed.

Key Words: Tandem mass spectrometry; proteomics; algorithms; database; protein identification; statistical models; bioinformatics.

1. Introduction

In the last few years, the shotgun proteomics approach ***(1–3)*** has become the method of choice for identifying and quantifying proteins in most large-scale studies (for a recent review, *see* **ref.** ***4***). This strategy is based on digesting proteins into peptides followed by peptide sequencing using tandem mass spectrometry (MS/MS) and automated database searching. Compared with

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

methods of analysis based on extensive protein separation prior to MS-based identification, such as two-dimensional (2D) gels ***(5)***, shotgun proteomics allows higher data throughput and better protein detection sensitivity. It also has an advantage over methods of analysis based on MS/MS sequencing of intact proteins. Intact protein sequencing (reviewed in **ref.** ***6***) is difficult owing to their large molecular weight, and requires more expensive mass spectrometry (MS) instrumentation, which hinders the implementation of this technique in a typical academic laboratory.

The method of shotgun proteomics analysis is schematically illustrated in **Fig. 1**. The first step of this method is digestion of sample proteins into peptides using proteolytic enzymes such as trypsin. Because each protein digested with trypsin produces multiple peptides, the resulting peptide mixtures can be very complex. Peptide samples are then separated by one- or multidimensional liquid chromatography (LC) and subjected to MS/MS analysis to sequence the peptides. In quantitative proteomics, reviewed in **ref.** ***4***, peptides are also encoded with a stable isotope tag, which allows determination of the relative protein abundances with respect to a control sample. Peptides are then ionized, and selected ions are subjected to sequencing to produce signature MS/MS spectra. The MS/MS data acquisition process consists of two stages. The first stage involves reading all peptide ions that are introduced into the instrument at any given time (MS spectrum). At the second stage, selected peptide ions (often referred to as "precursor" or "parent" ions) are fragmented into smaller pieces (fragment ions) in the collision cell of the mass spectrometer in the process termed collision-induced dissociation (CID). The acquired MS/MS spectrum is thus a record of mass-to-charge ratios (*m/z* values) and intensities of all the resulting fragment ions generated from an isolated precursor ion. The fragmentation pattern encoded by the MS/MS spectrum allows identification of the amino acid sequence of the peptide that produced it (*see* **Fig. 2**). After the desired amount of MS data is collected, the effort shifts toward the computational analysis.

The computational analysis typically starts with the identification of the peptides that give rise to the acquired MS/MS spectra. In high-throughput studies, the most efficient peptide identification method is based on searching MS/MS spectra against protein sequence databases. A number of automated database search tools have been described, including widely used commercial programs such as SEQUEST and Mascot. Although these tools can be easily run by someone with little experience in MS, the user should be able to make appropriate choices of the database search parameters. The user should also be aware of various data interpretation challenges, including high rates of false identifications produced by those tools.

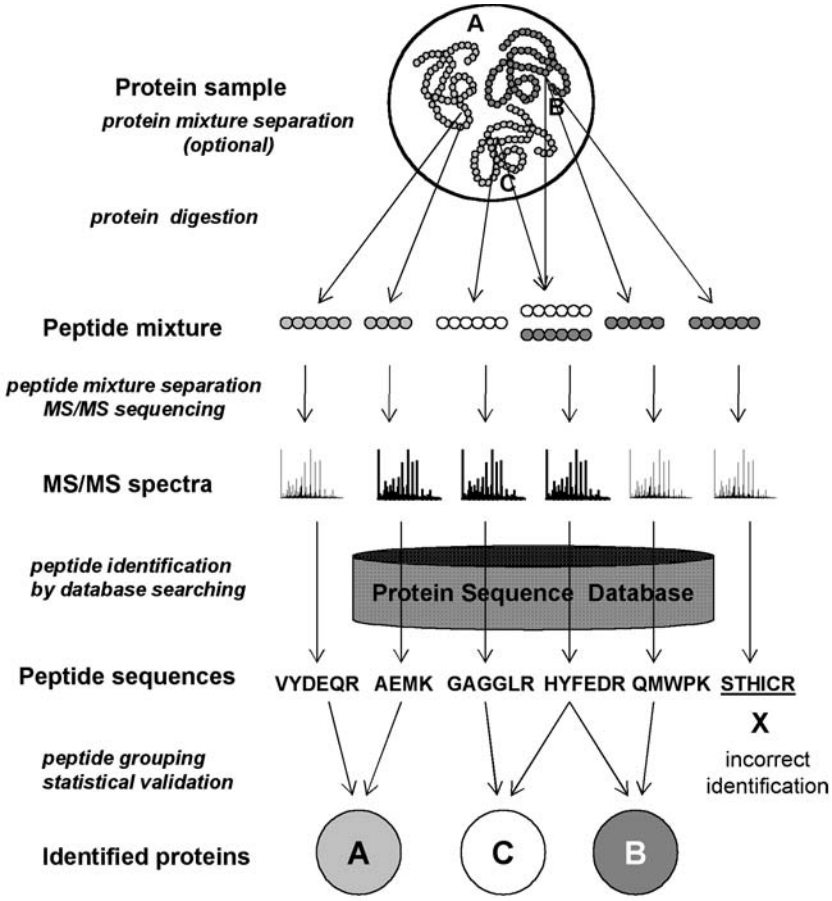


Fig. 1. General view of the experimental steps and flow of the data in shotgun proteomics analysis. Sample proteins are first proteolytically cleaved into peptides. After separation using one- or multidimensional chromatography, peptides are ionized and selected ions are fragmented to produce signature tandem mass spectrometry (MS/MS) spectra. Peptides are identified from MS/MS spectra using automated database search programs. Peptide assignments are then statistically validated and incorrect identifications filtered out (peptide STHICR). Sequences of the identified peptides are used to infer which proteins are present in the original sample. Some peptides are present in more than one protein (peptide HYFEDR), which can complicate the protein inference process.

2. Peptide Identification Methods

In a typical experiment, a single mass spectrometer can generate thousands of MS/MS spectra per hour, and manual spectrum interpretation is not a feasible option. As a result, a number of computational approaches and software tools

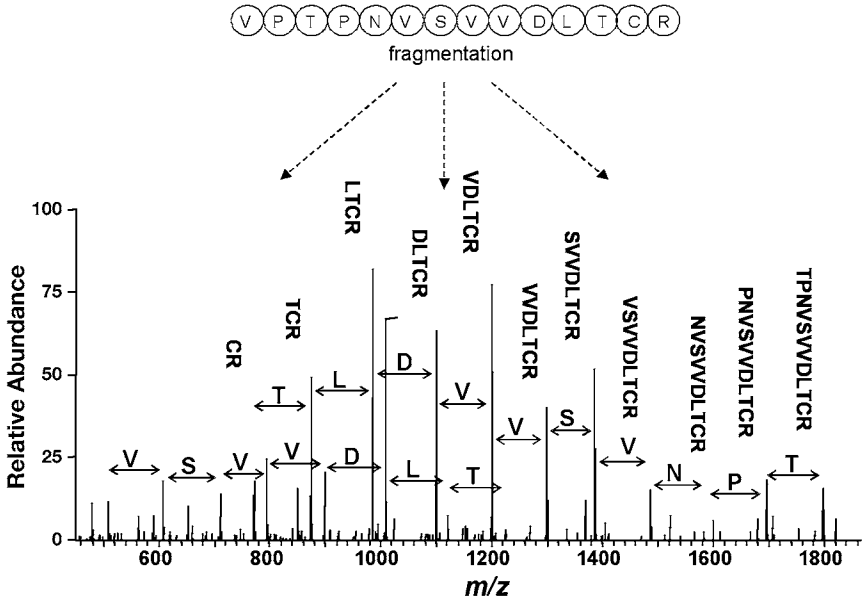


Fig. 2. An example of a tandem mass spectrometry spectrum.

have been developed for automated assignment of peptide sequences to MS/MS spectra ***(7)***. Existing computational approaches can be roughly classified into three categories. In the first approach, peptide sequences are extracted directly from the spectra, i.e., without referring to a sequence database for help (*de novo* sequencing approach) ***(8–12)***. In the second approach, peptide identification is performed by correlating experimental MS/MS spectra with theoretical spectra predicted for each peptide contained in a protein sequence database (database search approach) ***(13–21)***. The third category includes hybrid approaches, such as those based on the extraction of short sequence tags (three to five residues long) followed by "error-tolerant" database searching ***(22–25)***.

Each of the approaches has its own advantages and limitations. The advantage of the *de novo* sequencing approach over the database search method is that it allows identification of peptides whose exact sequence is not present in the searched sequence database. However, *de novo* analysis is computationally intensive and requires high-quality MS/MS spectra. Furthermore, researchers analyzing proteomic data are more interested in knowing what proteins are present in the sample. This means that peptide sequences extracted from MS/MS spectra using *de novo* algorithms need to be matched, e.g., using BLAST, against the sequences of known proteins present in the sequence databases. In the high-throughput proteomics environment, researchers are not always

interested in, or have time to follow up on the peptides identified using *de novo* sequencing tools and for which there is no exact match in the database (assuming the database is fairly complete). As a result, the computational analysis typically starts with database searching, and only then, if desired, *de novo* sequencing tools are applied to the remaining unassigned spectra.

In the case of organisms with unsequenced or only partially sequenced genomes, the database search approach will fail to assign correct peptide sequences to many MS/MS spectra (even if the search is performed against a database containing sequences from multiple related species), and it becomes necessary to use *de novo* sequencing tools ***(26)***. If the analyzed sample contains only a few proteins (e.g., proteins that are first separated on a 2D gel), protein identification can be performed by merging peptide sequences extracted from all spectra using a *de novo* sequencing tool into a single sequence and searching that sequence against a protein sequence database of a homologous organism using a modified version of the BLAST algorithm, MS-BLAST ***(27)***. However, if the sample contains more than several proteins, this approach is likely to fail, and it is not applicable in the case of complex protein samples. Hybrid approaches have also been suggested that combine inference of short sequence tags (partial sequences) from MS/MS spectra with an error-tolerant database search, i.e., search that allows one or more mismatches between the sequence of the peptide that produced the MS/MS spectrum and the database sequence. This approach, first described in **ref.** ***22***, has been recently extended by several groups ***(23,24)***. By limiting the search space to only those database peptides that contain the sequence tag extracted from the spectrum (or one of the several sequence tags if more than one per spectrum is extracted), a significant reduction in the database search time can be achieved. As these methods improve and the software tools that implement them become available, they should make a great addition to the suite of computational peptide identification tools available to proteomics researchers. However, searching MS/MS spectra against protein sequence databases will likely remains the primary method for the identification of peptides from MS/MS spectra in most proteomic studies.

3. Peptide Identification by MS/MS Database Searching

3.1. Basic Principles

Several MS/MS database search tools are currently available, including established and widely used commercial applications such as SEQUEST ***(13)*** and Mascot ***(14)***, integrated programs (that provide other functionalities in addition to database searching) such as SpectrumMill ***(19)*** and Phenyx ***(21)***, and open source database search tools such as X!Tandem ***(17)***, OMSSA ***(18)***, and ProbID ***(19)*** (*see* **Table 1**). All these tools operate in a similar manner. They take an experimental

Table 1
A Partial List of Publicly Available Tools for Assigning Peptides to MS/MS Spectra and for Statistical Validation of Peptide and Protein Identifications

Program	Website database search tools:
SEQUEST	http://www.thermo.com
MASCOT	http://matrixscience.com[a]
Protein Prospector	http://prospector.ucsf.edu[a]
ProbID	http://projects.systemsbiology.net/probid[b]
X!Tandem	http://www.thegpm.org[a,b]
OMSSA	http://pubchem.ncbi.nlm.nih.gov/omssa[a,b]
	Database searching using sequence tags:
GutenTag	http://fields.scripps.edu/GutenTag
InsPect	http://peptide.ucsd.edu/[a,b]
	Integrated systems (include database search tools):
SpectrumMill	http://www.chem.agilent.com
Phenyx	http://www.phenyx-ms.com
	Post-database search processing (no statistical validation):
INTERACT	http://www.proteomecenter.org/software.php[b]
DTASelect	http://fields.scripps.edu/DTASelect
DBParser	http://www.proteomecommons.org
	Post-database search processing (with statistical validation):
PeptideProphet	http://www.proteomecenter.org/software.php[b]
ProteinProphet	http://www.proteomecenter.org/software.php[b]

[a]Free access via the web interface.
[b]Free distribution (open source tools).

MS/MS spectrum as input and compare it against theoretical fragmentation patterns constructed for peptides from the searched database to find a match (*see* **Fig. 3**). The search is typically restricted to a subset of all database peptides based on such user-specified criteria as mass tolerance, proteolytic enzyme constraint, and types of posttranslational modifications allowed. Theoretical fragmentation patterns are calculated for each of the candidate peptides using common peptide fragmentation rules. The output from the database search tools is a list of matches (peptide sequences) ranked according to the scoring scheme implemented in each particular tool; the best scoring peptide match has the highest likelihood of being correct. Several types of search parameters are commonly used with MS/MS database searches, which are described next.

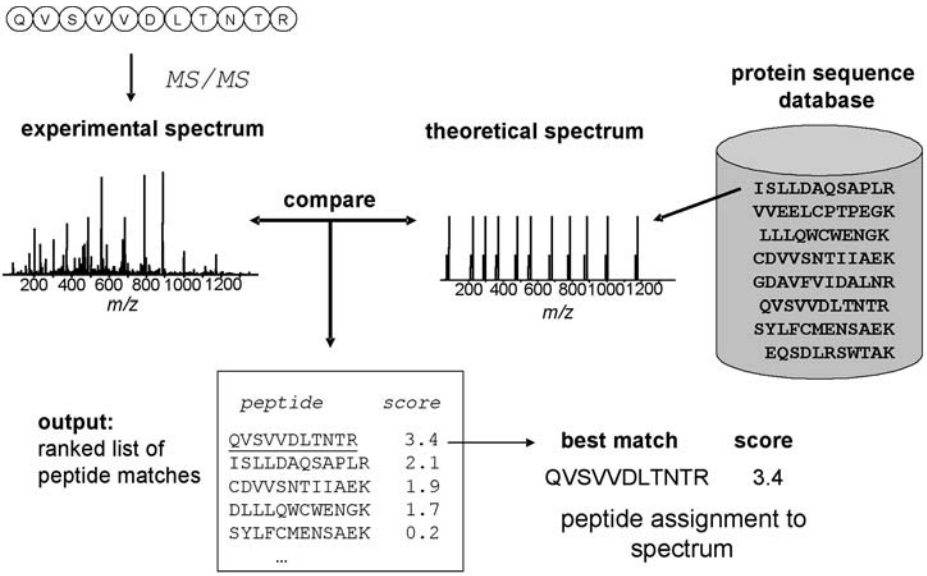


Fig. 3. Tandem mass spectrometry (MS/MS) database searching. Acquired MS/MS spectra are correlated against theoretical spectra constructed for each database peptide that satisfies a certain set of database search parameters specified by the user. A scoring scheme is used to measure the degree of similarity between the spectra. Candidate peptides are ranked according to the computed score, and the highest scoring peptide sequence (best match) is selected for further analysis.

3.2. Database Search Parameters

1. Types of fragment ions. For each candidate peptide, theoretical fragmentation patterns are calculated using common peptide dissociation rules. The fragment ions, composed of both C- and N-terminal ions, are specific for each type of mass spectrometer ***(28)***. Under the low energy (a few keV) CID conditions commonly encountered in mass spectrometers such as ion trap, triple quadrupole, and quadrupole time-of-flight (TOF) instruments, peptide fragmentation primarily results in the so called b- and y-ion. The high-energy CID conditions encountered in other types of instruments, like TOF-TOF instruments, are also capable of generating other ion types such as a-, c-, x-, and z-ions. Fragment ions specific to residue side chain cleavages can also be generated at high collision energy conditions. Most database search tools allow the user to select the types of ions to be included in the generation of the theoretical spectra (default settings for most common types of the mass spectrometer are often provided as well).
2. Monoisotopic vs average mass. Another parameter that needs to be specified by the user is the method of calculation of the peptide mass (monoisotopic mass or average mass). Mass spectrometers do not measure the masses of peptides, but rather the mass-to-charge values (*m/z*) of peptide ions. The *m/z* value of the

peptide ion is measured during the first stage of the MS analysis (MS spectrum), and the mass of the peptide is then computed from it. The mass of the peptide ion determined this way can be closer to the monoisotopic mass (mass of the peptide ion containing monoisotopic ^{12}C atoms only) or the average mass (average over the masses of all peptide ions, including those containing one or more ^{13}C atoms). In the case of high-resolution mass spectrometers, the measured mass is likely to be the monoisotopic mass, whereas it is better to choose the average mass in the case of low-resolution instruments, such as ion traps.

3. Peptide ion charge state. Determination of the peptide mass from MS data requires the knowledge of the charge state of the peptide ion. In the case of high-resolution instruments it can be determined quite reliably from the isotope distribution pattern observed in the MS spectrum. This is not always the case, however, with low-resolution instruments, such as ion traps. In low-resolution instruments, distinguishing between singly and multiply charged ions is often possible, but the exact charge state of a multiply charged ion cannot be easily determined. When the charge state is not known, the same MS/MS spectrum is searched against the database at least twice, once assuming +2 and once assuming +3 charge state (in the case of peptides produced by digesting proteins with trypsin, peptide ions carrying more than three charges are observed less frequently). After the search is done, the best scoring assignments obtained for each of the assumed charge states need to be reconciled, either by selecting the assignment with the highest database search score, or using statistical methods ***(29)***.
4. Parent ion mass tolerance. Only candidate peptides that have the calculated mass within a certain range of the measured peptide mass are selected from the sequence database and scored against the experimental spectrum. The choice of the mass tolerance parameter depends on the type of mass spectrometer used. With low mass accuracy instruments, such as ion traps, one should specify a fairly large mass tolerance of 2–3 Da. With TOF-based mass spectrometers it is possible to achieve a mass accuracy of less than 0.1 Da, and even better mass accuracy of less than 0.05 Da with Fourier transform MS instruments.
5. Enzymatic digestion constrain. Many of the proteolytic enzymes used to digest proteins into peptides cleave specifically after certain residues in the protein sequence. For example, the most commonly used enzyme trypsin cleaves after arginine (R) and lysine (K) residues (but usually not if they are followed by proline). Thus, a peptide resulting from trypsin digestion should contain K or R at its C-terminus (unless it is a C-terminal peptide and the last residue in the protein is not K or R), and in the sequence of its corresponding protein the residue immediately preceding the peptide should also be K or R (or the peptide is located at the N-terminus). The peptide also should not contain any missed cleavages, e.g., internal K or R residues.

   The knowledge of the digestion process can be used to limit the search space to only those peptides that conform to the digestion rules specific to each proteolytic enzyme. All database search tools allow the user to specify the digestion enzyme as a parameter of the database search. If the enzyme is specified, then the

number of candidate peptides that need to be analyzed is significantly reduced (compared with the enzyme unconstrained search allowing any linear stretch of amino acids). This means that the search time is significantly reduced as well, because it depends both on the size of the database and the number of candidate peptides from that database that need to be scored against the experimental spectrum. In the enzyme-constrained search, the user can also limit the number of missed cleavages allowed in the sequence of the candidate peptide.

Performing enzyme-constrained searches, however, has disadvantages. It becomes impossible to identify peptides that exhibit unspecific cleavage, e.g., from posttranslational processing (e.g., removal of the signal peptide) or from contaminating enzymes present in the sample, or because they are products of in-source or in-solution fragmentation of other (tryptic) peptides. Similarly, if the spectrum is produced by a peptide containing more than the allowed number of missed cleavages, it will not be identified. As a result, in order not to lose those identifications some researchers prefer to perform enzyme-unconstrained searches, provided they have significant computational resources to do that. A good alternative, available in some database search tools, is to perform "semi-constrained" searches, i.e., searches allowing for a nonspecific cleavage at one of the peptide termini and for one or two internal missed cleavage sites.

6. Chemical or posttranslational modifications. Database searches can be performed with one or more static (all occurrences of the residue are modified) or variable (the residue may or may not be modified) modifications such as oxidation, methylation, phosphorylation, deamidation, and others. For example, if cysteine residues of all proteins are chemically modified, this should be considered as a static modification. On the other hand, methionine oxidation (Met +16 Da) should be specified as a variable modification because not all but only a small fraction of all methionine residues are oxidized.

3.3. Selection of the Protein Sequence Database

For some organisms (e.g., human), multiple sequence databases are available ***(30)***. The most commonly used databases are the National Center for Biotechnology Information's (NCBI) Entrez Protein database, the NCBI Reference Sequence (RefSeq) database, and UniProt (consisting of Swiss-Prot and its supplement TrEMBL). The International Protein Index (IPI) database, maintained by European Bioinformatics Institute (EBI), is also frequently used, which is available for six organisms, including human and mouse. All these databases can be easily located and downloaded from the World Wide Web. They vary in terms of the completeness, degree of redundancy, and the quality of sequence annotation; which database is the best to use depends on the goals of the experiment. When the identification of sequence polymorphisms is important (e.g., in proteomics studies of certain diseases that are caused by mutations in the genomic sequence), the search should be done against large databases, such as Entrez

Protein. The disadvantage of searching large databases, however, is that in addition to true biologically significant sequence variants they also contain many redundant sequences (partial mRNAs, sequencing errors, and so on). The entries in Entrez Protein and similar large databases are not well annotated. As a result, researchers have to perform manual analysis to eliminate nonbiological redundancies, which can be time consuming. Searching large databases also introduces more false identifications—the larger the database is, the more likely it is to obtain a high scoring but incorrect peptide match by chance. Thus, when the ability to identify sequence variants is not crucial, it is better to search well-curated and annotated databases such as Swiss-Prot or RefSeq. For the six organisms for which it is available, the IPI database represents a good balance between the completeness and the degree of redundancy. It also maintains cross-references to all its source data (Ensembl, UniProt, RefSeq), making biological data interpretation easier. Genomic databases also can be used for MS/MS database searching ***(31,32)***. This is an attractive option when one wants to identify novel peptides not present in any protein sequence database, e.g., novel alternative splice forms, or sequence polymorphisms. However, searching genomic databases should be practiced with great caution. Accurate translation of the DNA sequences into protein sequences is complicated because of frameshifts, incorrectly predicted open reading frames, and other factors. The search against expressed sequence tag (EST) databases is further complicated by the poor quality of the sequence data. Combined with the poor quality of many experimental MS/MS spectra, genomic database searches can lead to many incorrect identifications. This type of analysis is also very computer intensive resulting from the large size of genomic databases. Thus, it is advisable to perform searches against genomic databases only as a last step, i.e., after searches against protein sequence databases failed to assign a peptide to the spectrum with high confidence.

3.4. Sources of Failure to Assign Correct Peptide Sequences

All MS/MS database search tools perform in a similar way. They return the best matching peptide found in the database for each input spectrum, except when there are no candidate peptides in the searched database that satisfy the search parameters specified by the user (e.g., in the case of enzyme-constrained searches with very narrow mass tolerance). However, the best match returned by the database search tool is not necessarily correct ***(7,29,33)***. In some cases, e.g., in the case of ion trap mass spectrometers, the fraction of all searched spectra that get assigned correct peptide sequences is less than 50%. The reasons why the database search tools fail to assign correct peptide sequences to so many experimental MS/MS spectra include:

1. Deficiencies of the scoring scheme. When the correct peptide sequence is in the database, another (incorrect) peptide can score higher than the correct one owing

to a deficiency of the used scoring scheme. The scoring schemes implemented in most commonly used database search tools are based on a simplified representation of the peptide ion fragmentation process. In particular, in creating the theoretical fragmentation spectrum, all fragment ion peaks (from the fragment ion series specified in the search parameters file) are often assumed to be present in the spectrum and with equal intensities. In reality, the intensity of a fragment ion depends on the amino acids located on each side of the corresponding peptide bond. For example, the presence of a proline in the sequence often leads to a very intense fragment ion corresponding to the breakage of the peptide bond on the N-terminal side of the proline. The knowledge of peptide fragmentation chemistry is often used by experienced mass spectrometrists in the process of manual validation of peptide assignments to spectra. Research efforts are currently underway to better understand these phenomena and incorporate them in the scoring schemes ***(34,35)***.

2. Low MS/MS spectrum quality. Correct interpretation of MS/MS spectra is difficult if the spectra are of low quality, i.e., they contain many noise peaks, have low signal-to-noise ratios, and/or have missing fragment ion peaks owing to incomplete peptide fragmentation. Furthermore, some MS/MS spectra are acquired not on peptides, but on various contaminants introduced in the sample during sample preparation. The number of low-quality spectra in a typical shotgun proteomics dataset is high, and most of them cannot be correctly assigned by the database search tools.
3. Fragmentation of multiple peptide ions. Some MS/MS spectra, which can be a significant percentage when complex peptides mixtures are analyzed, result from simultaneous fragmentation of two more different peptide ions having similar *m/z* values. Because most database search tools operate under the assumption that the spectrum is acquired on a single precursor ion, they often fail to assign any of the peptide sequences to the spectrum.
4. Presence of homologous peptides. Another common problem is the presence of several different but homologous peptides in the searched database. This problem is particularly serious in the case of higher eukaryotes. The mass difference between several amino acid combinations (e.g., D/N, E/Q/K) cannot be resolved using low mass accuracy instruments such as ion traps, and two of the amino acids, I and L, have an identical mass. If the database contains several peptides with a similar molecular weight that have a high degree of sequence homology, an incorrect (homologous) peptide can score slightly higher than the correct one ***(36)***. This can lead to incorrect biological interpretation of the data, and the users of the database search tools should apply additional scrutiny when encountering such cases.
5. Incorrectly determined charge state or peptide mass. When the charge state of a multiply charged ion cannot be determined, the spectrum is typically searched against the database twice, once assuming +2 and then +3 charge states ***(29)***. However, if the true charge state is +4 or higher, or if the software incorrectly classified the spectrum as being produced by a multiply charged ion when it is a singly charged one or vice versa, then the correct peptide will not be found. Also, the mass spectrometer can sometimes select the first or the second isotope peak (peptide ion containing one or more ^{13}C atoms) for MS/MS fragmentation. When

this happens, the peptide mass computed from the recorded *m/z* value is incorrect (it differs by 1–2 Da or more from the calculated monoisotopic peptide mass). Thus, even if the correct sequence is in the database, it will not be selected and scored against the experimental spectrum unless the search is performed using a sufficiently large mass tolerance.

6. Restricted database search. Because performing searches allowing for many different types of chemical or posttranslational modifications is time consuming, database searching is often done allowing for no modifications, or only the most common ones, such as methionine oxidation. Similarly, databases are often searched in the enzyme-constrained manner, e.g., searching only for tryptic peptides with less than one missed cleavage. A spectrum will not be identified if it is produced by a peptide containing an unspecified modification, resulting from an unexpected protein cleavage, or containing more than the allowed number of missed cleavages. Several database search tools (X!Tandem, SpectrumMill, Mascot) allow multistep analysis. The first step in such analysis is the enzyme-constrained search not allowing for any modifications, which then can be extended to look for peptides with certain modifications, nonspecific cleavage, or with missed cleavage sites. The additional searches are performed only against the sequences of the proteins that were identified by at least one high scoring peptide in the initial search. Alternatively, similar computational efficiencies can be realized without limiting the protein search space by performing more comprehensive reanalysis of only those spectra that are of high quality and remain unassigned after the initial search ***(37)***.
7. Sequence variants and novel peptides. As discussed previously in **Subheading 2.**, identification of novel peptides, or peptide containing sequence polymorphisms that are not present in the searched protein database is not possible. To identify such peptides, it is necessary to search large genomic databases, or use database-independent peptide identification methods such as *de novo* sequencing.

3.5. Scoring Schemes and Evaluation of the Search Results

Because for many MS/MS spectra the best scoring peptide assignment returned by the database search tool is incorrect, the user has to evaluate the search results and filter out false identifications ***(7,29,38,39)***. Manual validation of peptide assignments to spectra by visual inspection is a very time-consuming process and is simply not feasible in high-throughput analysis of large datasets containing tens of thousands of spectra. It also requires expertise in MS and peptide fragmentation chemistry. Instead of manual analysis, the validation of peptide assignments to MS/MS spectra can be done in an automated or semi-automated fashion (with only partial manual validation) using the database search scores reported by each tool as filtering criteria.

The database search score is a score computed according to some scoring function that measures the degree of similarity between the experimental spectrum and the theoretical fragmentation patterns of the candidate peptides.

Different database search tools use different scoring schemes, and some tools calculate more than one score. A variety of scoring schemes have been described in the literature, including those based on spectral correlation functions, shared fragment counts, spectrum alignment, or based on empirically derived rules. A detailed review of all different scoring schemes goes beyond the scope of this chapter, and the following discussion will focus on those database search tools that are publicly available and presently used in many proteomics laboratories.

SEQUEST ***(13)*** is the first MS/MS database search tool that became commercially available, and it remains one the most commonly used programs. For each experimental spectrum, SEQUEST calculates the cross correlation score (*Xcorr*) for all candidate peptides queried from the database. To compute this score, the intensity of the peaks in the experimental spectrum are normalized, low-intensity peaks are removed, and all *m/z* values in the spectrum are rounded off to the next integer to create a new (processed) experimental spectrum (spectrum X). The theoretical spectrum (spectrum Y) is created for each candidate database peptide using a set of simplified peptide fragmentation rules. The SEQUEST's *Xcorr* score is then computed via the correlation function *Corr(t)* (the product between the vectors X and Y, with Y shifted with respect to X along the *m/z* axis by *t* mass units) as follows:

$$\begin{aligned} Corr(t) &= \sum_i x_i y_{i+t} \\ Xcorr &= Corr(0) - < Corr(t) >_t \end{aligned} \tag{1}$$

where nonzero elements x_i and y_i represent the peaks in the (processed) experimental and theoretical spectra. The *Xcorr* score essentially counts the number of fragment ions that are common between X and Y (allowing for some small differences in *m/z* values resulting from mass measurement errors). The score is also corrected to account for the number of matches occurring at random, which is done by subtracting the average value of the function *Corr*(*t*) within a certain range around $t = 0$. For each experimental spectrum, the best scoring peptide assignment (highest *Xcorr* score) is kept for further analysis. In addition to *Xcorr*, a derivative score, the relative difference between the best and the second best *Xcorr* score, ΔC_n, is computed. Both of these scores are useful for discriminating between correct and incorrect identifications; the higher the scores are, the more likely it is that the best scoring peptide assignment is correct. This is illustrated in **Fig. 4A**, which shows a scatter plot of *Xcorr* and ΔC_n values for a test dataset of peptide assignments to MS/MS spectra generated from a sample of 18 purified proteins, and where for each assignment it is known with high certainty whether it is correct or incorrect ***(29)***. Because *Xcorr* measures the number of matching ions, it is not length-independent, as shown in **Fig. 4B**.

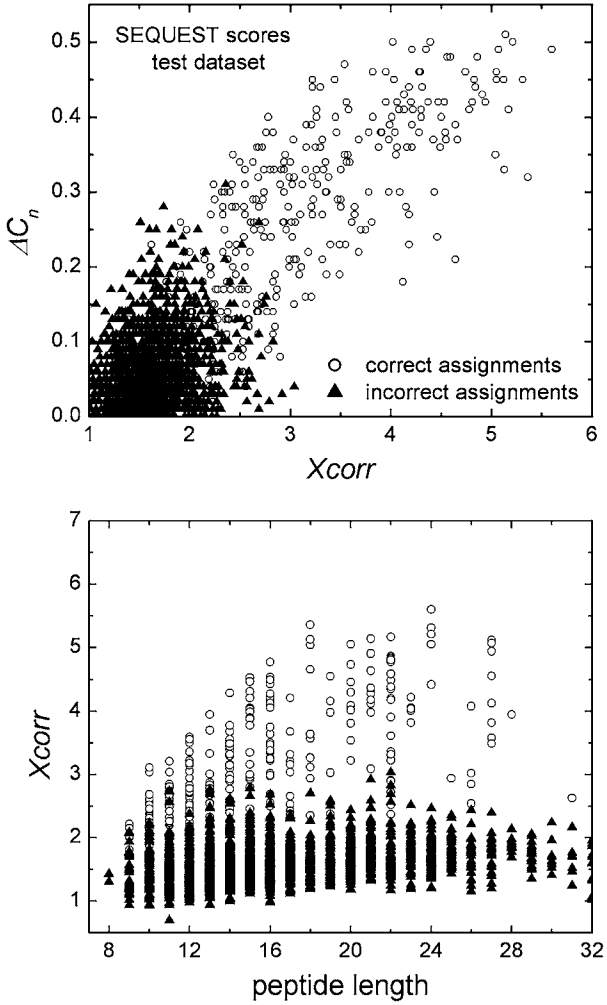


Fig. 4. **(A)** Separation between correct and incorrect results using SEQUEST's *Xcorr* and ΔC_n scores. ΔC_n is plotted vs *Xcorr* for correct (open circles) and incorrect (black triangles) test dataset search results for doubly charged precursor ions (dataset described in **ref. *29***). **(B)** Dependence of *Xcorr* on peptide length.

Furthermore, the distributions of scores are different for peptide assignments to spectra of different precursor ion charge states.

Another widely used database search tool, Mascot ***(14)***, computes a probability-based score called the ion score (often referred to simply as Mascot score). Similar to SEQUEST, the Mascot scoring scheme is based on the shared peak count. The main difference is that the Mascot score is probability based. Instead of reporting the number of matched peaks (shared peak count), Mascot estimates

the probability of that number of matches occurring by chance given the number of peaks in the searched spectrum and the cumulative distribution of *m/z* values of predicted ions for all candidate peptides in the database. At the end, the probability that the match is random is converted into a more convenient scale by taking the logarithm. The ion score is less sensitive to such parameters as the peptide molecular weight and the charge state of the precursor ion than *Xcorr*, but not completely independent. Unfortunately, the details of the Mascot scoring scheme remain unpublished, which prevents a comprehensive and independent assessment of this tool.

A more recently developed database search tool SpectrumMill, as extension of the earlier program Protein Prospector ***(15)***, computes several scores based on a set of empirically derived rules. The main score again counts the number of matched peaks (including the presence of neutral losses, immonium ions, and internal fragments). In addition, a penalty is applied for the presence of unmatched peaks. The score also incorporates the intensity of the fragment ions observed in the experimental spectrum, the length of the peptide, and other factors. The second score measures what fraction of the total peak intensity (i.e., the sum of the intensities of all peaks in the spectrum) is explained by the peaks for which there is a match in the theoretical spectrum. These two scores, and several secondary scores, provide the basis for the evaluation of the significance of each peptide assignment. Several thresholds of varying stringency are suggested, but the decision on how to filter the data to achieve sufficiently low error rate without losing too many correct identifications ultimately lies with the user of the tool.

Another recently developed commercial program, Phenyx ***(21)***, implements a more complicated model of peptide fragmentation that takes into account the likelihood of observing a certain type of fragment ions given the peptide sequence. The score computed by that tool is a likelihood ratio of probabilities computed based on the two alternative hypothesis, that the match is random (H_0) or true (H_1) ***(21,40,41)***:

$$S_L = \log\left(\frac{P(E \mid D, seq, H_1)}{P(E \mid D, seq, H_0)}\right) \tag{2}$$

where $P(E|D,seq,H_1)$ and $P(E|D,seq,H_0)$ describe the probabilities of observing the fragmentation pattern E given the peptide sequence *seq* and various external factors D (such as properties of the sequence database) assuming that the match is correct or occurring by chance, respectively. The calculation of these probabilities requires large training datasets of MS/MS spectra to empirically model the intensities of fragment ions as a function of the sequence composition and other factors. Because peptide fragmentation is dependent on the type of

mass spectrometer and the ionization source, the best performance can be achieved by optimizing the scoring function separately for each instrument type. The practical implementation of this scoring method requires a number of simplifying assumptions (e.g., independence of peptide fragmentation at different positions in the peptide sequence). As a result, the actual score computed using **Eq. 2** does not represent the likelihood ratio of true probabilities. Thus, the score itself does not allow easy evaluation of the correctness of the match.

An important question that always comes up in the analysis of MS/MS database search results is how to judge whether a peptide assignment with a certain database search score (probability-based or not) is statistically significant. This is particularly difficult when the database search tool does not compute any statistical confidence measures. As a result, different groups apply different (often not stringent enough) thresholds, and the resulting datasets may contain large numbers of false identifications. This is certainly the case with SEQUEST, which is not a probability-based tool. Mascot improves upon SEQUEST in this regard, and computes a score threshold, called identity threshold, that can be used as a guideline for filtering the data. It is described as a value of the ion score such that a peptide assignment with a score above that value have less than 5% probability of being a random match ***(14)***. However, the identity threshold is seldom applied in practice. Instead, as in the case of SEQUEST, different research groups use different, often arbitrarily selected thresholds instead of the provided identity threshold. The users of SpectrumMill and Phenyx are facing the same problems.

As a result, many large datasets of peptide and protein identifications published in the literature in recent years contain large numbers of false identifications, which diminishes their value. Furthermore, it is practically impossible to compare or correlate different datasets analyzed using different database search tools, or even using the same tool if different threshold are applied. Thus, when preparing a manuscript for publication, the authors are urged to provide as much detail about the data analysis as possible. A set of guidelines suggesting what information should be provided in manuscripts can be found (*see* **ref.** ***39***). The need to use robust and transparent statistical methods for validation of large datasets of peptide and protein identifications has also been discussed (*see* **refs.** ***7*** and ***42***).

Two recently developed open source database search tools, X!Tandem ***(17)*** and OMSSA ***(18)***, attempt to address the concerns raised above by calculating for each peptide assignments a statistical confidence measure called the expectation value. The scoring functions implemented in these tools are based on the shared peak count approach, although significant differences exist in the way the experimental spectra are processed prior to correlating them with the theoretical spectra. In OMSSA, the score is the number of matches between the experimental and the theoretical spectrum (similar to the *Corr*[0] term in

Eq. 1), whereas in X!Tandem the score (called hypergeometric score) includes extra terms that account for the number of assigned b- and y-ions:

$$s_{hyper} = (n_b ! n_y !) \sum_i x_i y_i \quad (3)$$

The process of validating peptide assignments to spectra in these tools is assisted by the conversion of the search scores into expectation values ***(18,20,43)***. When a peptide assignment to a spectrum with a database search score s is said to have an expectation value E, it means that by searching the database one can expect, on average, to observe E number of peptides getting a score equal or greater than s by random. Thus, the smaller the expectation value is, the less likely it is that the match is random (therefore, it is more likely to be correct). In X!Tandem, the conversion of the database search scores into the expectation value is done empirically. In the empirical approach, illustrated in **Fig. 5**, the observed distribution $f(s)$ of the database search scores (a histogram of the frequency of the occurrence of a particular score s among all performed comparisons between the experimental spectrum and the spectra generated for database peptides) is created for each MS/MS spectrum searched against the database. This distribution is normalized to the total number of candidate database peptides n:

$$f_{norm}(s) = f(s) / n \quad (4)$$

The normalized distribution $f_{norm}(s)$ is then fitted using a model distribution $P(s)$ (e.g., Gaussian distribution). The choice of the model distribution, and the details of the fitting procedure (e.g., the fitting can be performed using the high scoring tail of the distribution only, and after log-transformation) can vary depending on the scoring scheme. The underlying assumption is that $P(s)$ represents the distribution of random matches. The hypothesis that is being tested is that the top scoring peptide assignment with the score s_m is also a random match. The first step is to compute an analog of the p value by integrating the area under the right tail of the model distribution $P(s)$ that extends beyond s_m. Because each MS/MS spectrum is compared with n theoretical spectra generated for all candidate peptides from the sequence database, the expectation value E is computed as

$$E = n \sum_{s \geq s_m} P(s) \quad (5)$$

The empirical approach described can be applied to any database search scoring scheme (the latest version of Mascot also reports the expectation value

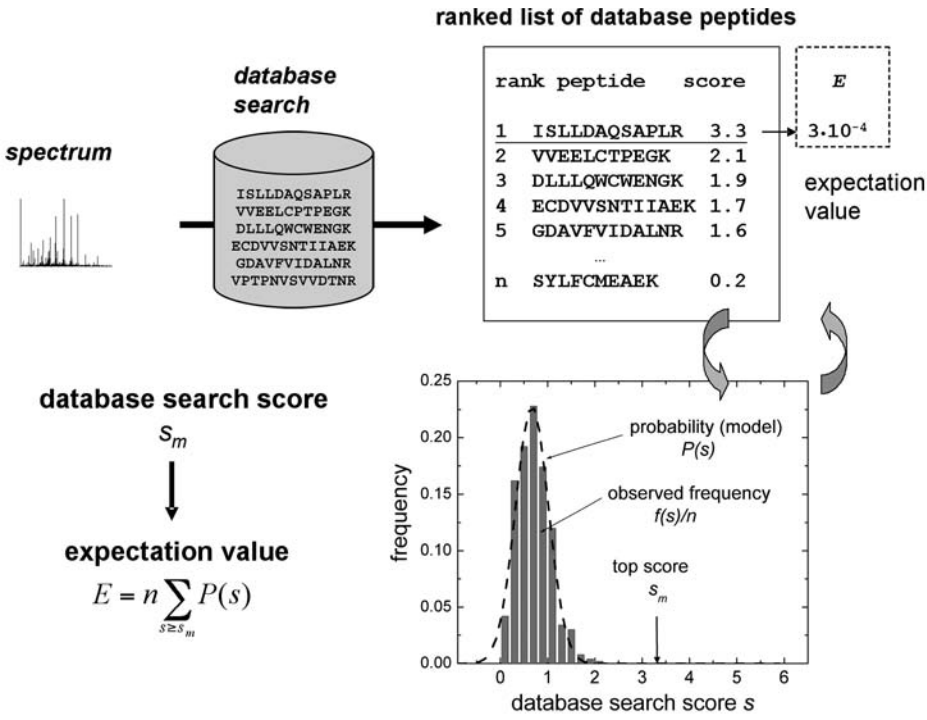


Fig. 5. Evaluation of the significance of the best scoring peptide assignment to a spectrum using expectation values. A histogram of the frequency of the occurrence of a particular score, *s*, among all performed comparisons between the experimental spectrum and the spectra generated for database peptides is constructed (SEQUEST's *Xcorr* score is used in this example), normalized to the total number of candidate database peptides *n*, and fitted using a model distribution *P*(*s*) (Gaussian distribution, dashed line). The area under the right tail of *P*(*s*) that extends beyond the top score s_m is computed, and then converted into the expectation value.

along with the original ion score and the identity threshold). Furthermore, in some cases the expectation value can be computed directly (without empirical distribution fitting) by assuming a certain probability model for the occurrence of a particular number of matches *s* by chance given the properties of the searched protein sequence database ***(18,20)***. This approach is used in OMSSA, where the number of matches *s* between the experimental and the theoretical spectrum is modeled using the Poisson probability function:

$$P(s) = \frac{\mu^s}{s!}\exp(-\mu) \qquad (6)$$

where parameter μ is estimated theoretically as a function of the mass measurement accuracy, the number of peaks in the experimental spectrum, the total number of calculated *m/z* values for all candidate peptides in the database, and the precursor ion mass and charge. For the peptide assignment with the highest number of matches, s_m, the probability $P(s_m)$ computed using **Eq. 6** is converted into the expectation value using **Eq. 5**. The expectation values and *p* values are well-known statistical measures that are used widely in similar applications, e.g., in sequence similarity searches ***(44)***. Conversion of the database search scores into the expectation values reduces the problem of incompatibility between the scoring schemes used in different search tools. Still, it is important to point out an important limitation of these statistical measures when filtering datasets consisting of thousands or tens of thousands of different observations (peptide assignments in this case). Computed independently for each peptide assignment, the expectation values or *p* values on their own are not particularly useful measures for filtering large datasets. It is not always clear what *p* value or the expectation value one should select as a threshold for filtering the entire dataset of peptide assignments, and how to estimate the resulting false identification rates for any particular threshold value. The filtering of large datasets can be best carried out with a help of statistical methods that operate not at the level of a single peptide assignment, but model the distribution of scores observed for all peptide assignments in the entire experiment ***(7,29)***, as discussed in the next section. In that regard, it is interesting to note that a similar limitation of *p* values as statistical measures in the field of gene expression analysis using microarrays has led to the development of a related data analysis approach based on the estimation of false discovery rates ***(45)***.

4. Statistical Validation of Large-Scale Datasets

The general method for filtering large-scale datasets of peptide assignments using cumulative statistical measures, such as false identification rates, has been first described in **ref. *29***. For the sake of clarity, the notion of false identification rates will be first illustrated using a simple approach called reversed database searching. In this method, all MS/MS spectra from the same dataset are searched against a composite database consisting of a normal database and a reversed database in which all protein sequences have been reversed. The number of assignments of peptides that are present in the reversed protein sequences only have a score above a certain threshold s_t counted, $N_{rev}(s_t)$, as well as the total number of peptide assignments above that threshold, $N_{tot}(s_t) = N_{rev}(s_t) + N_{norm}(s_t)$. Because all assignments of peptides present in the reversed sequences only can be assumed incorrect, and assuming that the same number of random matches occurs to the normal sequences [$2N_{rev}(s_t)$ in total], the false

identification rate, $Err(s_t)$, resulting from filtering the data using threshold s_t can be estimated as

$$Err(s_t) = \frac{2N_{rev}(s_t)}{N_{tot}(s_t)} \quad (7)$$

The dataset of peptide assignments can then be filtered using the threshold value s_t that corresponds to the desired false identification rate. For example, if the target error rate is 1%, one would determine the corresponding s_t value using the $Err(s_t)$ curve computed according to **Eq. 7**, and apply that threshold to the dataset.

Although simple and easy to implement ***(46)***, the reversed database approach has several limitations. The false-positive rates determined using reversed database searching for a particular dataset cannot be directly transferred to a different dataset. Thus, to ensure the accuracy of the estimates, the analysis should be repeated for each dataset anew, resulting in longer database search times owing to increase in the database size. Also, it has been observed that searching reversed databases can lead to inaccurate statistical estimates (this problem can be reduced by using randomly generated or scrambled databases instead of reversed databases). Moreover, it becomes increasingly hard to use this approach in the presence of multiple database search scores that are useful for discrimination, or when using additional parameters, such as the number of tryptic termini or missed cleavages, liquid chromatography elution time, and so on. Even more importantly, this approach only allows estimation of the composite false identification rates, but not the probabilities of the individual peptide assignments. The individual probabilities are necessary for the calculation of the confidence measures at the protein level, as discussed in **Subheading 5.** of this chapter. A robust and accurate statistical method for validation of peptide assignments to spectra introduced in **ref.** ***29*** does not have the limitations of the reversed database search approach. This method is implemented in a freely available computational tool PeptideProphet (*see* **Table 1**). PeptideProphet takes as input a dataset of database search results and computes, for each peptide assignment, a probability of being correct. The method is based on the use of the expectation maximization (EM) algorithm to derive a mixture model of correct and incorrect peptide identifications from the data, and can be generally described as an empirical ("learning from the data") Bayesian approach. If the database search tool outputs a single score useful for discriminating between correct and incorrect peptide assignments, the method can be described as an unsupervised leaning method. The basic idea of the method is illustrated in **Fig. 6**. Because every peptide assignment can be classified into one of the two categories, correct or incorrect, the distribution

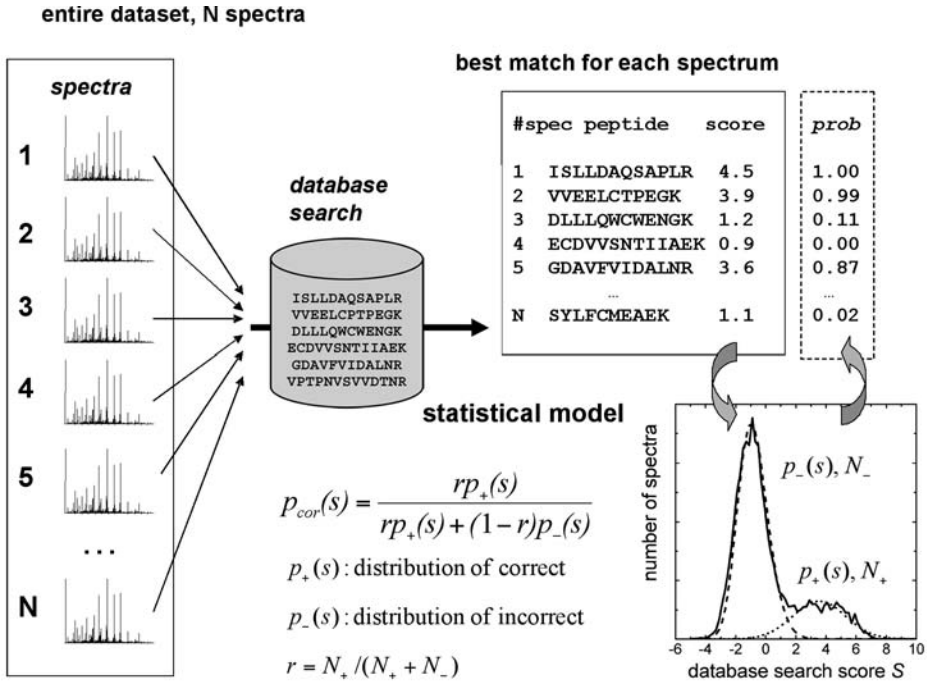


Fig. 6. Statistical model for validation of large datasets of peptide assignments to tandem mass spectrometry (MS/MS) spectra implemented in PeptideProphet. The model takes as input a list of peptide assignments (best database match for each spectrum) and the corresponding database search scores for the entire dataset of MS/MS spectra. It learns the most likely distributions (dashed lines) among correct and incorrect peptide assignments given the observed data (solid line), and computes for each peptide assignment in the dataset a probability of being correct.

of the scores observed for the entire dataset of MS/MS spectra, $f_{tot}(s)$ can be represented as a mixture of two distributions:

$$f_{tot}(s) = f_+(s) + f_-(s) \tag{8}$$

where $f_+(s)$ and $f_-(s)$ are the distributions of the database search score s among correct and incorrect identifications, respectively. The distributions in **Eq. 8**, which describe the results of searching multiple MS/MS spectra against a database, should be distinguished from the distributions discussed previously in the context of a single MS/MS spectrum. The distribution $f(s)$ in **Eq. 4** represented a histogram of the frequency of the occurrence of a particular score s for all comparisons of a single MS/MS spectrum against all n candidate database peptides. In contrast, $f_{tot}(s)$ represents the frequency of the occurrence of a particular score s for all

peptide assignments to N MS/MS spectra in the entire dataset, with only the top-scoring peptide assignment to each MS/MS spectrum considered. The probability $p_{cor}(s)$ that a peptide assignment with a score s is correct can be expressed as

$$p_{cor}(s) = \frac{f_+(s)}{f_{tot}(s)} = \frac{rp_+(s)}{rp_+(s) + (1-r)\,p_-(s)} \tag{9}$$

where $p_+(s)$ and $p_-(s)$ are the probabilities (normalized frequencies) of observing a peptide assignment with a score s among correct and incorrect assignments, respectively, and r is the proportion of correct assignments in the dataset, $r = N_+/N$. The parameters of the distributions $p_+(s)$ and $p_-(s)$ (e.g., the means and the standard deviations if the distributions are modeled as Gaussians), and the overall proportion of correct assignments in the dataset, r, are learned directly from the data in an unsupervised manner using an iterative algorithm (the EM mixture model algorithm). It should be stressed that no training dataset (or reversed database searching) is required for the method to fit the observed distribution $f_{tot}(s)$ and to compute the probability p_{cor} for each assignment in the dataset, except to determine what form of the distribution to use for modeling $p_+(s)$ and $p_-(s)$. In other words, the method learns the underlying distributions $p_+(s)$ and $p_-(s)$ without relying on the knowledge of the distribution parameters determined explicitly by means of reversed database searching (where the distribution $p_-(s)$ is estimated based on the matches to the reversed sequences) or using training datasets generated using control samples (in which case $p_-[s]$ and $p_+[s]$ can be determined using the knowledge of what proteins are present in the sample).

In the case of the database search tool Mascot, the method can be applied as previously described, except that the ion score is renormalized prior to the application of the EM algorithm to reduce its dependence on the molecular weight (which is done by subtracting the identity score from the ion score). Also, an additional penalty is applied by reducing the ion score in those cases where the so-called homology threshold (computed by Mascot along with the identity threshold) reported for that assignment is unusually high (i.e., outside the typical range of homology threshold values observed for other peptides in the same dataset that have similar ion scores). The unusually high homology threshold reflects the presence of other candidate peptides in the database scoring almost as high as the best scoring peptide, which lowers the confidence that the best scoring peptide assignment is correct (in that regard, the difference between the ion score and the homology score is similar to SEQUEST's ΔC_n score). Also, because slightly different distributions of scores are observed for peptide assignments to spectra of different charge state, each charge state is modeled separately. After fitting the distributions

($p_+[s]$ is model as a Gaussian and $p_-[s]$ as an extreme value distribution) and determining the model parameters, the probability of being correct is computed for each assignment in the dataset. The computed probability is then used instead of the original score.

In addition to the database search score, PeptideProphet uses additional peptide properties such as the number of termini consistent with the enzymatic cleavage (*NTT* value), the number of missed cleavages (*NMC* value), the difference between the measured and calculated peptide mass (ΔM), the presence of a certain amino acid or a sequence motif in the peptide sequence (e.g., cysteine residues in ICAT experiments), and other parameters that may be useful for discrimination. Assuming the independence between these parameters and the database search score (which holds true for these variables under most common conditions), the probability that the assignment is correct given the database search score *s* and the values *NTT*, *NMC*, and ΔM can be written

$$p_{cor}(s, NTT, NMC, \Delta M) = \frac{rp_+(s, NTT, NMC, \Delta M)}{rp_+(s, NTT, NMC, \Delta M) + (1-r)p_-(s, NTT, NMC, \Delta M)} \quad (10)$$

The distributions of *NTT*, *NMC*, and ΔM values among correct and incorrect identifications are learned using the EM algorithm along with the distributions of the database search score. Because the *NTT* and *NMC* values are discrete, the distributions of these variables are not modeled using any probability function, but rather are computed as proportions of all peptide assignments having a certain value of *NTT* and *NMC* (details of the calculations can be found in **ref. *29***). The mass difference parameter ΔM, although continuous, is modeled in a way similar to the treatment of *NTT* and *NMC*, except that it is first converted to a discrete variable by binning using a fixed number of bins covering the entire range of ΔM values.

The analysis becomes more complex in those cases where one would like to utilize more than one database search score, but the scores are not independent. This is the case with SEQUEST, where both *Xcorr* and ΔC_n scores are known to be useful for validation of peptide assignments. In addition, SEQUEST calculates several other scores, e.g., the so-called preliminary score S_p that also adds to the discrimination. At the same time, all these scores are not independent, e.g., a strong correlation is observed between ΔC_n and *Xcorr* (*see* **Fig. 4B**). In the statistical model implemented in PeptideProphet, all *m* different database search scores s_i are combined into a single composite score *F* that optimally (under the assumption of multivariate normality and linearity) discriminates between the correct and incorrect peptide assignments:

$$F(s_1, s_2 \ldots s_m) = c_0 + \sum_{i=1}^{m} c_i s_i \quad (11)$$

The discriminant function coefficients (the constant c_0 and the weighting factors c_i that determine the relative contribution of each search score) were optimized for different types of mass spectrometers (ion trap, TOF, and so on) using training datasets generated using mixtures of purified proteins. The statistical model, therefore, is no longer completely unsupervised. However, the role of the training dataset is relatively minor. The distributions of the new composite score F among correct and incorrect assignments, $p_+(F)$ and $p_-(F)$, are not estimated from the training data, but again modeled in the unsupervised manner using the EM mixture model algorithms as described in **Eqs. 8–10**, with the composite score F used in place of the single score s. The distributions of the composite score F vary depending on signal to noise in the MS/MS spectra, the search parameters, and the size of the database used, among other factors. The mixing proportion r varies even to a greater degree, because this parameter is a reflection of the overall quality of the data. However, because $p_+(F)$ and $p_-(F)$, and r are modeled for each dataset anew, the peptide probabilities computed using **Eq. 10** remain accurate even the discriminant function itself is not optimal for that particular dataset. This critical aspect of the statistical model ensures that PeptideProphet is robust toward variations in data quality, proteolytic digest efficiency, database size (to some degree), and other factors.

At present, PeptideProphet can be used to analyze the results of the database search tools SEQUEST and Mascot, and efforts are underway to adopt it to other programs as well. Extensive evaluation of the statistical model implemented in PeptideProphet demonstrated a very good agreement between the actual and computed probabilities in the entire 0 to 1 probability range. Probabilities computed by PeptideProphet are more efficient at separating the correct from the incorrect peptide identifications than the database search scores alone. As a result, PeptideProphet allows researchers to extract more correct identifications from the data with no increase in the number of incorrect identifications. Peptide probabilities can also be used to calculate the false-positive identification rates resulting from filtering the data using a minimum probability threshold p_t:

$$\begin{aligned} Err(p_t) &= \frac{N_{inc}(p_t)}{N_{tot}(p_t)} \\ N_{inc}(p_t) &= \sum_{\{i, p^i_{cor} \geq p_t\}} [1 - p^i_{cor}] \end{aligned} \qquad (12)$$

where p^i_{cor} is the probability that peptide assignments i in the dataset is correct, $N_{tot}(p_t)$ is the number of peptide assignments passing the minimum probability threshold p_t, and $N_{inc}(p_t)$ is an estimate of how many of those peptide assignments are incorrect. The ability to estimate false-positive rates allows consistent and

transparent filtering of large datasets. It also facilitates comparison of different types of mass spectrometers, or the benchmarking of various mass spectrometer settings and experimental procedures to identify those that maximize the number of correct peptide identifications (at the same fixed error rate) per sample or per unit time. More importantly, computed peptide probabilities allow statistical estimation of the likelihood for the presence of proteins corresponding to those peptides in the original sample, as described in **Subheading 5.**

5. Protein Inference

In most studies, researchers are interested in identifying proteins rather than peptides. Thus, peptides need to be grouped according to their corresponding protein, and the statistical confidence measures need to be recomputed at the protein level ***(47)***. Several difficulties have been identified that complicate the process of assembling peptides into proteins.

1. Nonrandom grouping of peptides. This problem can be best explained using an illustration in **Fig. 7**. Peptides that are identified correctly tend to group into a relatively small number of proteins. In contrast, incorrect peptide assignments can be described as random matches to entries in a very large protein sequence database, and almost every high scoring incorrect peptide assignment adds one new incorrect protein identification. As a result, even a small false identification error rate at the peptide level can translate into a high error rate at the protein level ***(47)***. This effect becomes more pronounced as the number of spectra in the dataset increases relative to the number of proteins in the sample. This also makes detection of correct protein identifications based on a single peptide (which is often the case with low abundance proteins) difficult, because most of the incorrect protein identifications also have only one corresponding peptide in the dataset.
2. Shared peptides. Identification of shared peptides, i.e., peptides whose sequence is present in more than a single entry in the protein sequence database, makes it difficult to infer the particular corresponding protein (or proteins) present in the sample. Such cases most often result from the presence of homologous proteins, splicing variants, or redundant entries in the protein sequence database ***(47,48)***. This problem is particularly serious in the case of higher eukaryote organisms. As a result, in shotgun proteomics it is often not possible to differentiate between different protein isoforms, as illustrated in **Fig. 8**. In general, this is less of a problem when proteins are first separated using a multidimensional protein separation technique (e.g., using 2D gels), where additional information, such as the molecular weight of the sample proteins, can assist in the determination of the protein identities. A detailed discussion of the difficulties in interpreting the results of shotgun proteomics experiments at the protein level can be found ***(49)***.

Most database search tools allow the user to view the results in a format that has peptides grouped according to their corresponding proteins. However, in most

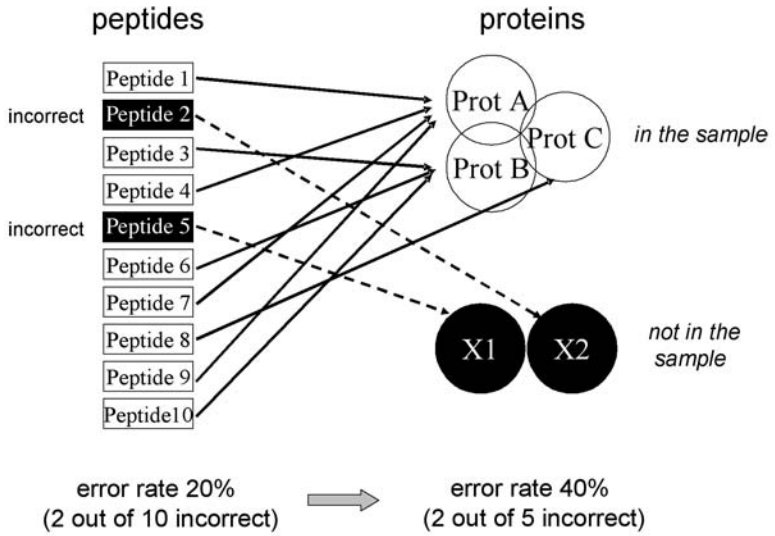
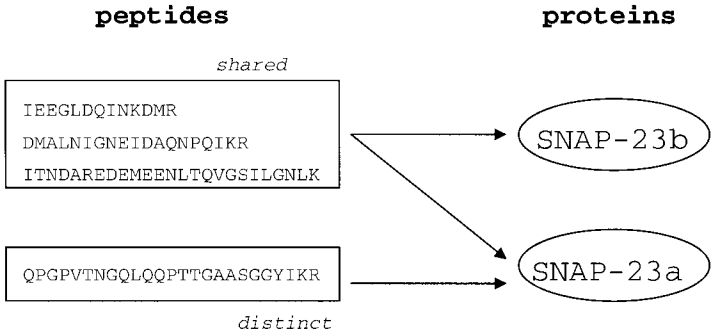


Fig. 7. Illustration of nonrandom grouping of peptides according to their corresponding proteins. 10 tandem mass spectrometry spectra were searched against a protein sequence database and each spectrum was assigned the best matching peptide, with 8 out of 10 assignments being correct (error rate 20%). Incorrect peptide assignments (shown in black) result in two incorrect protein identifications (X1 and X2). Eight correct peptide assignments correspond to only three correct proteins (A, B, C). As a result, in this example a 20% false identification rate at the peptide level (two out of ten peptide assignments are incorrect) translates into a 40% error rate at the protein level (two out of five protein identifications are correct). The figure is for illustration purposes only, and is not representative of actual error rates observed in a typical shotgun proteomics analysis.

large-scale studies one has to deal with multiple datasets of MS/MS spectra that are acquired and processed at different times. Thus, computational tools have been developed for combining peptide assignments from multiple experiments in order to derive a composite list of protein identifications. The earlier tools developed for that purpose such as INTERACT ***(50)*** and DTASelect ***(51)***, which can be used with SEQUEST and Mascot, are helpful for automating this process. However, they do not address any of the difficulties previously described. Another available software tool, DBParser ***(52)*** (compatible with Mascot only), can deal with the cases of shared peptides, but does not at present compute any statistical confidence measures for protein identifications.

All the difficulties discussed are addressed in the statistical model implemented in the computational tool ProteinProphet ***(47)***. This program takes as input a list of peptide identifications and their probabilities (output from PeptideProphet), and computes a probability that a protein is present in the sample by combining together the probabilities of its corresponding peptides



>IPI00216231 IPI:IPI00216231.1|UniProt/Swiss-Prot:O00161-2|REFSEQ_NP:NP_570710|ENSEMBL:ENSP00000207062 Tax_Id=9606 Splice isoform SNAP-23b of O00161 Synaptosomal-associated protein 23

MDNLSSEEIQ QRAHQITDES LESTRRILGL AIESQDAGIK TITMLDEQKE QLNR**IEEGLD QINKDMR**ETE
KTLTELNKCC GLCVCPCNS**I TNDAREDEME ENLTQVGSIL GNLKDMALNI GNEIDAQNPQ IKR**ITDKADT
NRDRIDIANA RAKKLIDS

>IPI00010438 IPI:IPI00010438.1|UniProt/Swiss-Prot:O00161-1|REFSEQ_NP:NP_003816|ENSEMBL:ENSP00000249647 Tax_Id=9606 Splice isoform SNAP-23a of O00161 Synaptosomal-associated protein 23

MDNLSSEEIQ QRAHQITDES LESTRRILGL AIESQDAGIK TITMLDEQKE QLNR**IEEGLD QINKDMR**ETE
KTLTELNKCC GLCVCPCNRT KNFESGKAYK TTWGDGGENS PCNVVSK**QPG PVTNGQLQQP TTGAASGGYI
KRITNDARED EMEENLTQVG SILGNLKDMA LNIGNEIDAQ NPQIKR**ITDK ADTNRDRIDI ANARAKKLID
S

Fig. 8. Three shared peptides are identified that are present in two different splice forms, SNAP-23b and SNAP-23a, of the synaptosomal-associated protein 23. SNAP-23a is also identified by one distinct peptide that corresponds to that isoform only, and thus it can be assumed to be present in the sample. The other isoform cannot be conclusively identified because its presence in the sample is not required to explain the observed peptides.

identified in the experiment. Peptides corresponding to “single hit” proteins (no other peptides identified in the dataset corresponding to the same protein) are penalized, but not excluded, whereas those corresponding to “multihit” proteins are rewarded. The appropriate amount of adjustment depends on the sample complexity, the number of acquired MS/MS spectra, and other factors, and is determined empirically for each dataset. Shared peptides are apportioned among all their corresponding proteins, and a minimal list of proteins is derived that can explain all observed peptides (*see* **Fig. 9**). ProteinProphet also collapses redundant database entries into a single identification and presents proteins that are impossible to differentiate on the basis of identified peptides as a group. The software ProteinProphet provides the user with many convenient interactive options. The output file is a list of proteins, their computed

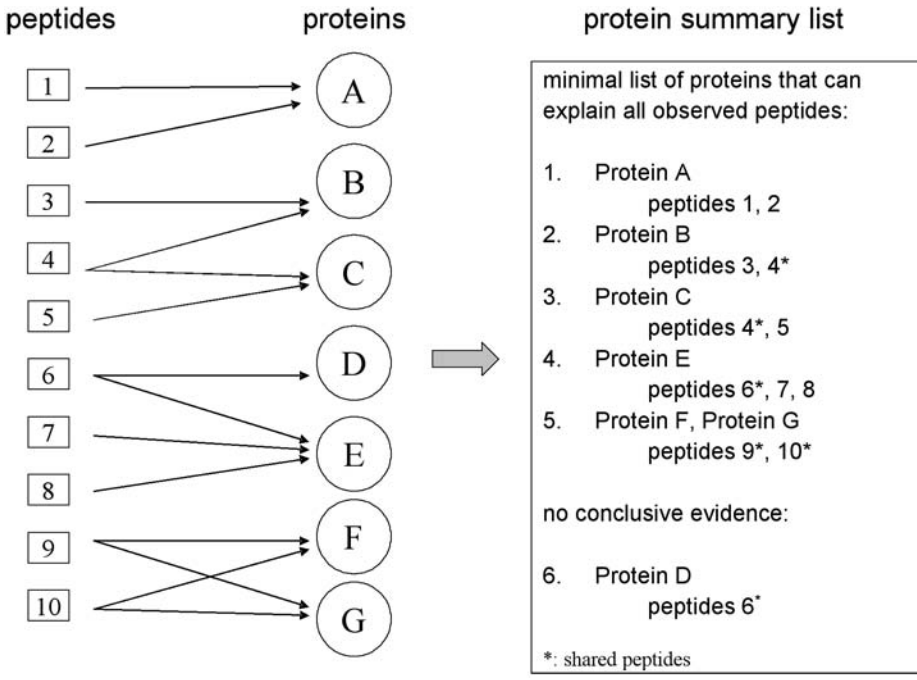


Fig. 9. Peptides are assembled into proteins to infer what proteins are present in the sample. Peptides are apportioned among all their corresponding proteins, and a minimal list of proteins is derived that can explain all observed peptides (A, B, C, E, and at least one of the two proteins, F or G). Proteins that are impossible to differentiate on the basis of identified peptides are presented as a group (F/G). It is not possible to conclude that protein D is present in the sample.

probabilities, and annotations extracted from the searched protein sequence database. For each protein, all supporting peptide identifications are listed, along with their computed probabilities and other information (*see* **Fig. 10**). ProteinProphet is integrated with the quantification tools XPRESS ***(50)*** and ASAPRatio ***(53)***, and can display the relative protein abundance ratios in the case of quantitative proteomics experiments using ICAT, SILAC, or similar approaches ***(4)***. The ProteinProphet output file can be viewed using a web browser, and links are provided that allow easy access to the original search

Fig. 10. *(Opposite page)* A screen shot of a sample ProteinProphet output file. Each protein entry is accompanied by its computed probability of being correct, and when available, quantification information. In addition, annotations extracted from the sequence database and peptide sequences with links to the original tandem mass spectrometry data help the user to interpret the results of the analysis.

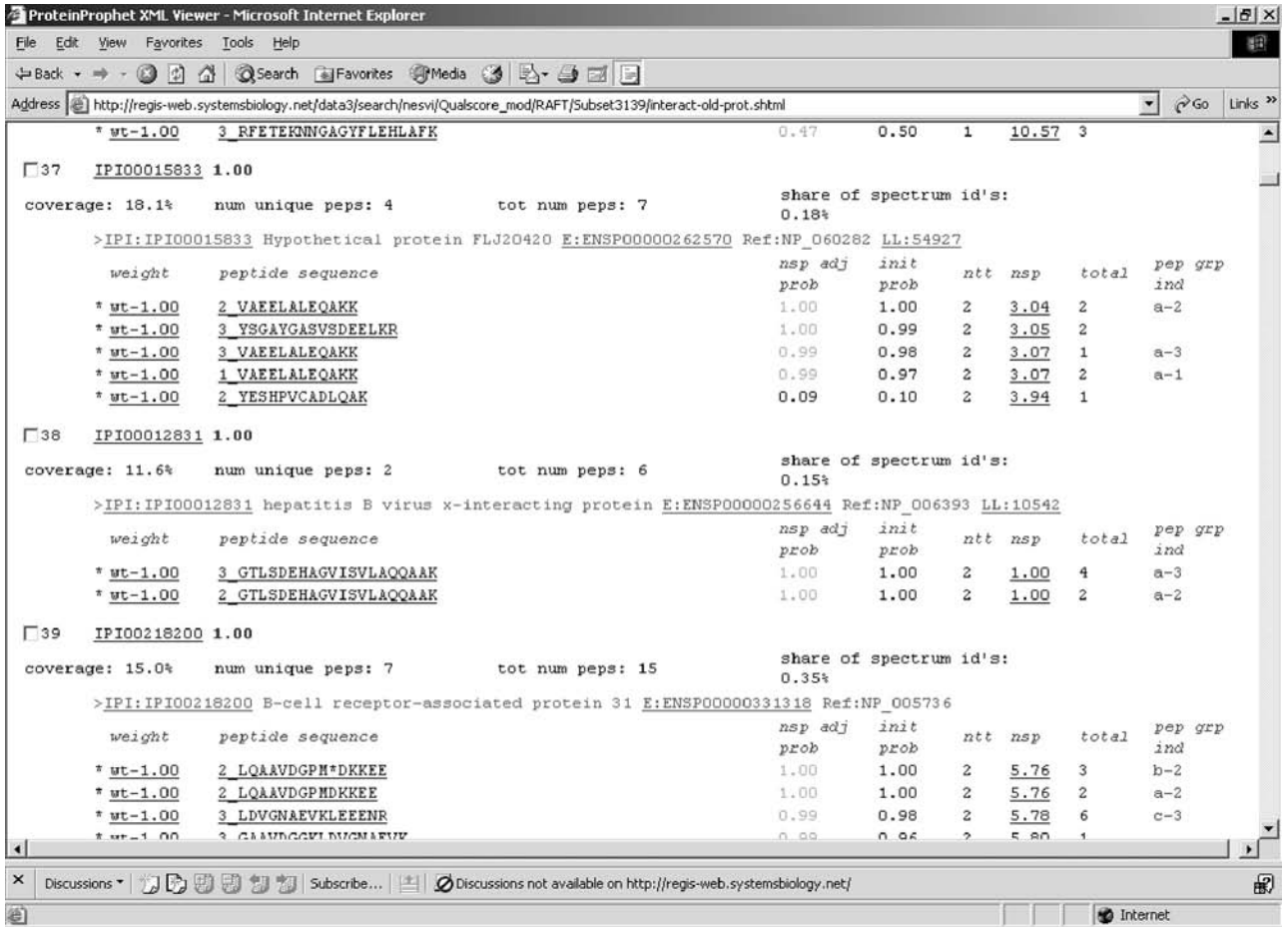

ProteinProphet XML Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://regis-web.systemsbiology.net/data3/search/nesvi/Qualscore_mod/RAFT/Subset3139/interact-old-prot.shtml
Go Links
* wt-1.00 3_RFETEKNNGAGYFLEHLAFK 0.47 0.50 1 10.57 3
37 IPI00015833 1.00
coverage: 18.1% num unique peps: 4 tot num peps: 7
share of spectrum id's: 0.18%
>IPI:IPI00015833 Hypothetical protein FLJ20420 E:ENSP00000262570 Ref:NP_060282 LL:54927
weight peptide sequence nsp adj prob init prob ntt nsp total pep ind grp
* wt-1.00 2_VAEELALEQAKK 1.00 1.00 2 3.04 2 a-2
* wt-1.00 3_YSGAYGASVSDEELKR 1.00 0.99 2 3.05 2
* wt-1.00 3_VAEELALEQAKK 0.99 0.98 2 3.07 1 a-3
* wt-1.00 1_VAEELALEQAKK 0.99 0.97 2 3.07 2 a-1
* wt-1.00 2_YESHPVCADLQAK 0.09 0.10 2 3.94 1
38 IPI00012831 1.00
coverage: 11.6% num unique peps: 2 tot num peps: 6
share of spectrum id's: 0.15%
>IPI:IPI00012831 hepatitis B virus x-interacting protein E:ENSP00000256644 Ref:NP_006393 LL:10542
weight peptide sequence nsp adj prob init prob ntt nsp total pep ind grp
* wt-1.00 3_GTLSDEHAGVISVLAQQAAK 1.00 1.00 2 1.00 4 a-3
* wt-1.00 2_GTLSDEHAGVISVLAQQAAK 1.00 1.00 2 1.00 2 a-2
39 IPI00218200 1.00
coverage: 15.0% num unique peps: 7 tot num peps: 15
share of spectrum id's: 0.35%
>IPI:IPI00218200 B-cell receptor-associated protein 31 E:ENSP00000331318 Ref:NP_005736
weight peptide sequence nsp adj prob init prob ntt nsp total pep ind grp
* wt-1.00 2_LQAAVDGPM*DKKEE 1.00 1.00 2 5.76 3 b-2
* wt-1.00 2_LQAAVDGPMDKKEE 1.00 1.00 2 5.76 2 a-2
* wt-1.00 3_LDVGNAEVKLEEENR 0.99 0.98 2 5.78 6 c-3
Discussions Subscribe... Discussions not available on http://regis-web.systemsbiology.net/
Internet

Fig. 10.

results for each spectrum, and other information. The user can also locate all proteins that share a particular peptide, and follow up on some of the proteins that cannot be conclusively identified as present in the sample based on the sequences of the identified peptides alone. ProteinProphet output files can also be exported to Excel format for further analysis or journal submission.

ProteinProphet has been extensively tested using both control samples and complex biological mixtures. It has been shown to produce accurate protein probabilities having high power to discriminate between correct and incorrect protein identifications. Importantly, the protein probabilities computed by ProteinProphet allow estimation of the false identification rates resulting from filtering the data at the protein level. In combination, PeptideProphet and ProteinProphet allow fast, consistent, and transparent analysis of shotgun proteomic data, and provide a consistent way for publishing large-scale datasets of peptide and protein identifications in the literature.

PeptideProphet and ProteinProphet are distributed as a part of the open source proteomics pipeline (www.proteomecenter.org/software.php), and are available in both Windows and Linux versions. More detailed information regarding the installation and use of these tools can be found on the website previously cited.

Acknowledgments

We would like to acknowledge Jimmy Eng for numerous discussions. This work was funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract number N01-HV-28179.

References

1. Link, A. J., Eng, J., Schieltz, D. M., et al. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol*. **17,** 676–682.
2. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol*. **17,** 994–999.
3. Washburn, M. P., Wolters, D., and Yates, J. R. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.
4. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207.
5. Gorg, A., Weiss, W., and Dunn, M. J. (2004) Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **4,** 3665–3685.
6. Reid, G. E. and McLuckey, S. A. (2002) 'Top down' protein characterization via tandem mass spectrometry. *J. Mass Spectrom*. **37,** 663–675.
7. Nesvizhskii, A. I. and Aebersold, R. (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Disc. Today* **9,** 173–181.

8. Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., and Pevzner, P. A. (1999) *De novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol*. **6,** 327–342.
9. Taylor, J. A. and Johnson, R. C. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal. Chem*. **73,** 2594–2604.
10. Chen, T., Kao, M. Y., Tepel, M., Rush, J., and Church, G. M. (2001) A dynamic programming approach to *de novo* sequencing via tandem mass spectrometry. *J. Comput. Biol*. **8,** 325–337.
11. Ma, B., Zhang, K., Hendrie, C., et al. (2003) PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17,** 2337–2342.
12. Frank, A. and Pevzner, P. (2005) PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77,** 964–973.
13. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*. **5,** 976–989.
14. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. C. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567.
15. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem*. **71,** 2871–2882.
16. Field, H. I., Fenyo, D., and Beavis, R. C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. *Proteomics* **2,** 36–47.
17. Craig, R. and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467.
18. Geer, L. Y., Markey, S. P., Kowalak, J. A., et al. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* **3,** 958–964.
19. Zhang, N., Aebersold, R., and Schwikowski, B. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2,** 406–1412.
20. Sadygov, R. G. and Yates, J. R. (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **75,** 3792–3798.
21. Allet, N., Barrillat, N., Baussant, T., et al. (2004) In vitro and in silico processes to identify differentially expressed proteins. *Proteomics* **4,** 2333–2351.
22. Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399.
23. Tabb, D. L., Saraf, A., and Yates, J. R. (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **75,** 6415–6421.
24. Frank, A., Tanner, S., Bafna, V., and Pevzner, P. (2005) Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res.* **4,** 1287–1295.
25. Salek, M., Di Bartolo, V., Cittaro, D., et al. (2005) Sequence tag scanning: a new explorative strategy for recognition of unexpected protein alterations by nano-electrospray ionization-tandem mass spectrometry. *Proteomics* **5,** 667–674.

26. Liska, A. J. and Shevchenko, A. (2003) Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics* **3,** 19–28.
27. Shevchenko, A., Sunyaev, S., Loboda, A., et al. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73,** 1917–1926.
28. Aebersold, R. and Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chem. Rev*. **101,** 269–295.
29. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392.
30. Apweiler, R., Bairoch, A., and Wu, C. H. (2004) Protein sequence databases. *Curr. Opin. Chem. Biol.* **8,** 76–80.
31. Kuster, B., Mortensen, P., Andersen, J. S., and Mann, M. (2001) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1,** 641–650.
32. Choudhary, J. S., Blackstock, W. P., Creasy, D. M., and Cottrell, J. S. (2001) Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1,** 651–667.
33. Patterson, S. D. (2003) Data analysis: the Achilles heel of proteomics. *Nat. Biotechnol*. **21,** 221–222.
34. Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H., Lin, D., and Yates, J. R. (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* **75,** 1155–1163.
35. Kapp, E. A., Schütz, F., Reid, G. E., et al. (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem*. **75,** 6251–6254.
36. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., et al. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem*. **76,** 3556–3568.
37. Nesvizhskii, A. I., Roos, R. F., Grossmann, J., et al. (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics* **5,** 652–670.
38. Von Haller, P. D., Yi, E., Donohoe, S., et al. (2003) The application of new software tools to quantitative protein profiling via ICAT and tandem mass spectrometry: II. Evaluation of tandem mass spectrometry methodologies for large-scale protein analysis and the application of statistical tools for data analysis and interpretation. *Mol. Cell. Proteomics* **2,** 428–442.
39. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data. *Mol. Cell. Proteomics* **3,** 531–533.
40. Bafna, V. and Edwards, N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **17 (Suppl.),** S13–S21.

41. Havilio, M., Haddad, Y., and Smilansky, Z. (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem*. **75,** 435–444.
42. Baldwin, M. A. (2004) Protein identification by mass spectrometry: issues to be considered. *Mol. Cell. Proteomics* **3,** 1–9.
43. Fenyo, D. and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75,** 768–774.
44. Karlin, S. and Altschul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
45. Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100,** 9440–9445.
46. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large scale protein analysis: the yeast proteome. *J. Proteome Res*. **2,** 43–50.
47. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658.
48. Rappsilber, J. and Mann, M. (2002) What does it mean to identify a protein in proteomics? *Trends in Biochem. Sci.* **27,** 74–78.
49. Nesvizhskii, A. I. and Aebersold, R. (2005) Interpretation of shotgun proteomic data: The protein inference problem. *Mol. Cell. Proteomics* **4,** 1419–1440.
50. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry *Nat. Biotechnol.* **19,** 946–951.
51. Tabb, D. L., McDonald, W. H., and Yates, J. R. (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res*. **1,** 21–26.
52. Yang, X., Dondeti, V., Dezube, R., et al. (2004) DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res*. **3,** 1002–1008.
53. Li, X. J., Zhang, H., Ranish, J. A., and Aebersold, R. (2003) Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.* **75,** 6648–6657.

7

Virtual Expert Mass Spectrometrist v3.0

An Integrated Tool for Proteome Analysis

Rune Matthiesen

Summary

The number of tools described in the literature for analysis of proteome data is growing fast. However, most tools are not able to communicate or exchange data with other tools. In Virtual Expert Mass Spectrometrist (VEMS) v3.0 an effort has been made to interface and export to already existing tools. In this chapter, an outline of how to use the VEMS program to search tandem mass spectrometry data against databases is described. Additionally, examples on how to extend the analysis with other external tools are given.

Key Words: Database searching; integrated analysis; validation; Virtual Expert Mass Spectrometrist; VEMS.

1. Introduction

Working with mass spectrometry (MS) data analysis in proteomics requires a broad set of analytical tools. The flow of data is generally divided into three levels ***(1)***: first, continuous raw data from the mass spectrometer should be noise, isotopic, and charge deconvolued to obtain a peak lists (*see* Chapter 2). Second, the MS data should be searched against a database in order to identify peptides and proteins. Depending on the experiment, the peptides and proteins could also be quantified at this level. Third, the data should be classified and stored in databases so that it can be systematically compared to predictions, broadening existing knowledge and knowledge to be obtained in the future. The goal of data analysis is, in general, to identify proteins, quantify the identified proteins, classify the data, and compare the results with already existing knowledge or predictions to extract biological relevance. To obtain a versatile tool that

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

can perform all this would be extensive work, therefore, the choice made for the Virtual Expert Mass Spectrometrist (VEMS) v3.0 ***(2–4)*** was to implement already exciting tools whenever possible. VEMS v3.0 is interfaced to the retention time predictor program rt (*see* Chapter 13), basic local alignment search tool (Blast) ***(5)***, TandemX ***(6)***, and Lutefisk ***(7)***. In addition, VEMS v3.0 has data export/import functions for exporting to BioEdit ***(8)***, SATv1.0 (*see* Chapter 10), Proteios ***(9)***, Excel, significance analysis for microarray data (SAM) ***(10)***, in-house developed PostGreSQL database (*see* Chapter 16). VEMS v3.0 accepts a range of data formats that will be outlined in the following sections.

This chapter explains, the basis of searching tandem mass spectrometry (MS/MS) data against sequence databases and illustrates how the obtained search result can be analyzed further by the tools in the VEMS program and the external tools interfaced from VEMS.

2. Software

2.1. Required Software

1. VEMS v3.0 (http://yass.sdu.dk).
2. Microsoft Windows. Currently only fully tested on Windows XP and Windows 2000.

2.2. Optional Software

1. Blast (*see* **Note 1**). Blast can be obtain from (http://www.ncbi.nlm.nih.gov/BLAST/) for Windows and Linux. For this purpose the Windows versions should be downloaded (blast-20041205-ia32-win32.exe). This file is a self-extracting archive and it should be executed from the Extern directory folder that is located in the VEMS directory folder. Now VEMS is able to export and import data to Blast.
2. TandemX (*see* **Note 2**) can be obtained from http://www.proteome.ca/open-source.html. TandemX should be located in the Extern/TandemX directory folder that is located in the VEMS directory folder.
3. Lutefisk (*see* **Note 3**) can be obtained from http://www.hairyfatguy.com/Lutefisk/. Lutefisk should be located in the Extern/lutefisk directory folder that is in the VEMS directory folder.
4. BioEdit (*see* **Note 4**) can be obtained from http://www.mbio.ncsu.edu/BioEdit/bioedit.html. Bioedit can be located any where on the host machine. VEMS interacts through the BioEdit through the clipboard.
5. SATv1.0 (*see* Chapter 10) can be obtained from (http:\\yass.sdu.dk). SATv1.0 can be located any where on the host machine. VEMS interacts through the SATv1.0 through the clipboard.
6. Proteios (*see* Chapter 17) can be obtained from http://www.proteios.org/. VEMS export XML that can be imported into the Proteois database.
7. SAM (*see* **Note 5**) can be obtained from http://www-stat.stanford.edu/~tibs/SAM/. VEMS inserts the relevant quantitative data directly into Excel so that it can be used in SAM.

8. ProteinLynx global server (PLGS) v2.0.5 (*see* **Note 6**) is a commercial program that can be obtained from Waters (Milford, MA). VEMS interfaces to some of the raw data processing tools of PLGS v2.05. However, VEMS also has built-in functions for raw data handling.

3. Methods

3.1. Searching the Liquid Chromatography–MS/MS Data

Having obtained a large set of liquid chromatography (LC)–MS/MS data for the given sample the next problem is to identify peptides, peptide modifications, and proteins. In VEMS one can analyze several LC–MS/MS runs at the same time even if the different LC–MS/MS runs corresponds to different samples. This is possible in the VEMS program because VEMS keeps track of which LC–MS/MS run the MS/MS spectra comes from, and because one can group the LC–MS/MS runs that belong to the same sample. This feature is important for recalibration and for quantitative time studies (*see* Chapter 9).

In the following, search strategies with the VEMS program are discussed and based on a test set of data that are freely available on http:\\yass.sdu.dk homepage. The raw data used for this chapter can be obtained by downloading the 357 Mb zip file (http://yass.sdu.dk/my00234kr/my00234kr.zip). Download of the VEMS program and the archive file PatatoPrx.pkx is available in the directory TestData. This file contains processed MS/MS spectra from the raw data file (*see* Chapter 2).

1. Start the VEMS application by executing the file "VEMS.exe."
2. Open the data import window from the file menu (File → Open data → Open multiple spectra or press sequentially "Alt"+"F"+"O"+"P").
3. The window in **Fig. 1** should now be visible. Area 1 is for choosing files containing multiple processed spectra. Area 2 is for choosing multiple raw data files and area 3 is for choosing multiple MS peak lists for PMF searches. Areas 4–6 show the selected files. It is important that the files containing the processed spectra are in the same order as the raw data files. This normally is not a problem because the program sorts the spectra alphabetically. However, right-clicking in the area of selection gives possibilities for manual editing in the list. Such activity can lead to wrong file associations. The drive letters for the directory listboxes can be changed by clicking on the small letters. The button ">>" is for choosing a single file and ">>>" is for choosing multiple files.
4. Choose the processed spectra file "PotatoPrx.pkx" and raw data file "MY00234 kr.raw."
5. Click "Transfer" and close the window.
6. When clicking on the page-tab named "Input," a list of all parent ions of the loaded MS/MS spectra in the listbox on the right should be visible. There should be 148 MS/MS spectra in total. On the top of the page one can choose how many missed cleavages to look for (*see* **Note 7**). For this data setting the default value to one is fine.

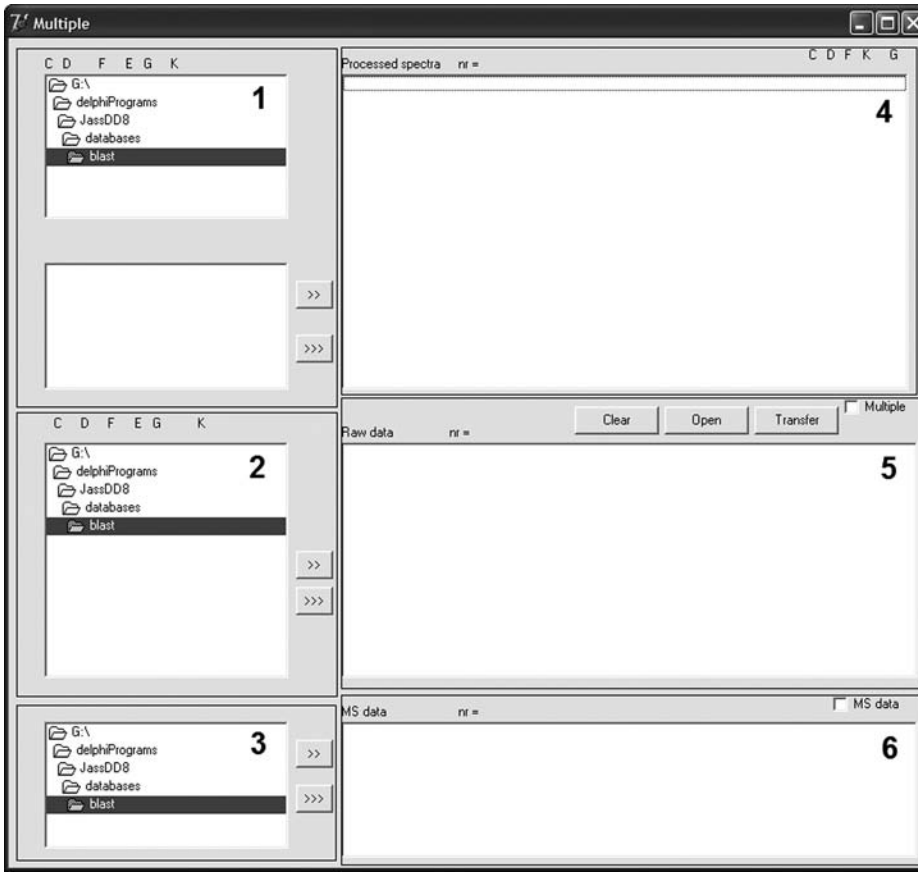


Fig. 1. Screen-shot of the data import window in VEMS. Areas 1–3 are for specifying the processed liquid chromatography–tandem mass spectrometry (LC–MS/MS) runs (peak lists), the raw data files, and MS peak lists. Areas 4–6 are the chosen LC–MS/MS peak lists, raw data, and MS peak lists.

7. Now the expected modifications should be chosen. First, choose which fixed modification to use (*see* **Note 8**). In most cases, one would choose carbamidomethylation of cystein (*see* Chapter 1) and would therefore press the radio button "CAM." However, the test sample under went performic acid oxidation, which double oxidizes methionine, triple-oxidize-cysteine, and double-oxidizes trypthophan (also giving other oxidation products of trypthophan). Assuming that completeness of the reaction is unknown, it is safer to assume that these modifications are variable or potential and therefore the radiobutton "std" for a standard amino acid table should be chosen. The tables with masses are stored as text files in the VEMS directory and

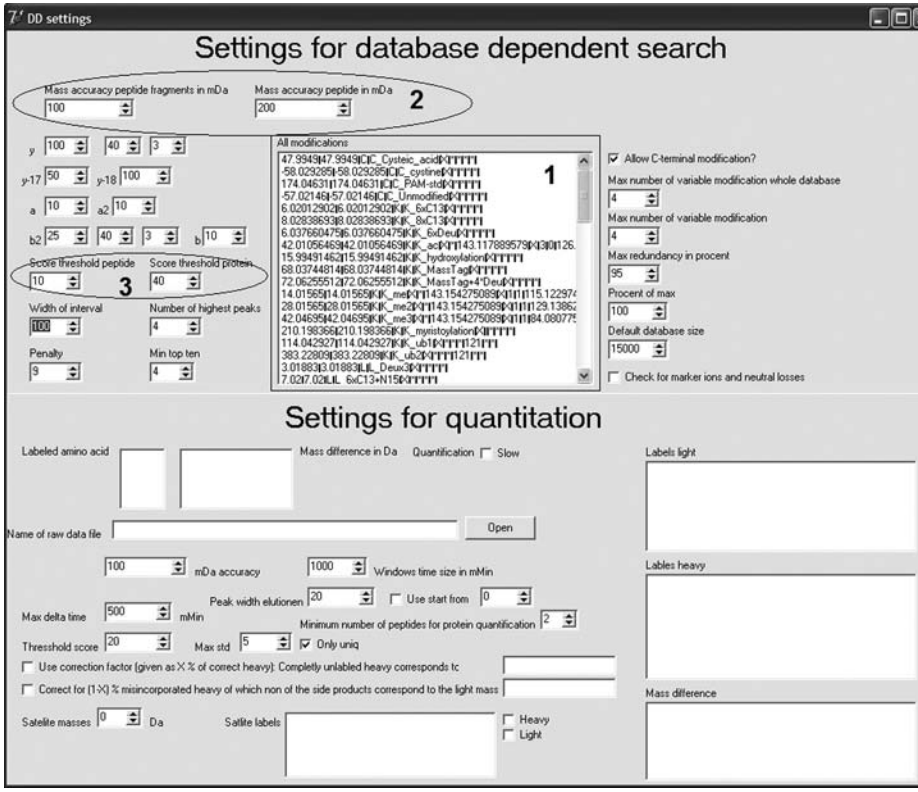


Fig. 2. The settings page. The most important areas are here enclosed either by a box or an ellipse. Area 1 is where one chooses the variable modifications. This is done by right-clicking in area 1 and choosing “Add to variable modification-whole database.” Area 2 is for setting the mass accuracy. Area 3 is for setting the threshold score for peptide and protein.

can easily be modified. After pressing the “Settings” button the variable modification can be chosen by right-clicking on the modification in the listbox (**Fig. 2**). By pressing different keys on the keyboard one gets a reduced list with only the modifications on the amino acids that correspond to the pressed key. Pressing key “1/2” will bring up the full list again. Now choose “oxidize methionine,” “triple oxidize cysteine,” and “double oxidize trypthophan.” On the setting page there are a number of settings for the search. In general, there is no need to change the default values except for the mass accuracy, which by default is set to 0.5 Da for parent ions and 0.5 Da for fragment ions. If the sample has been enriched for phosphopeptides then it is not expected to contain many peptides from proteins, and therefore it is worthwhile to lower the default protein score threshold in such cases.

8. To search the FASTA database choose from the listbox "FASTA Databases" (*see* **Note 9**). In the listbox, FASTA sequence files that are stored in the directory "databases" in the VEMS folder are shown. There should already be a small database of potato peroxidases named "PotatoSmall.txt." Alternatively download a more complete database and locate the FASTA file in the "databases" folder. Click on "PotatoSmall.txt" and press the button "<<." The "PotatoSmall.txt" should now be in the listbox at the left, meaning that it has been selected for the search.
9. Start the search by pressing the tab-page "Output" and pressing the button "Start" or sequentially press "Alt"+"S."
10. Go to "File → Open annotated data." A window pops up where one can annotate the searched data and give suitable comments.
11. The search result together with all the data and settings can now be saved in a single file by pressing "File → Save → Export to database." Save the file in the "Projects" folder in the VEMS directory.

Try to close the program and restart it. Go to "File → Open annotated data." The saved file should appear in the listbox in the bottom. Clicking on the file will display the annotated data. Pressing the "Load" button will reload the MS/MS data, the search setting, the annotation, and the search result. The following describes how one can continue analyzing the data. Close the window and press the page-tab "Output" and click on the radiobutton "View search" or "View quantification."

3.2. Recalibrating the LC–MS/MS Spectra

As a result of temperature fluctuations of the flight tube, one can often observe a linear systematic error in mass accuracy. The individual LC–MS/MS runs can be recalibrated in VEMS by fitting a linear equation to the theoretical and observed masses of the most confidently matched peptides. The test data previously listed (*see* **Subheading 3.1.**, **step 4**) is already recalibrated so the method in this section will not improve the mass accuracy. However, the sequence of steps below can still be tested with the given test set.

1. Right-click in the output window and chose "Re-calibrate."
2. Select a minimum threshold score. Only peptides with a score above the threshold will be used for the recalibration.
3. Select which LC–MS/MS run to recalibrate.
4. Click on the "Extract" button to extract peptides from the chosen LC–MS/MS run with a score above the threshold.
5. Press the button "linear fit" to make a linear fit.
6. If outliers are observed then click on the "Remove outlier" button.
7. Continue to click 5 and 6 until all outliers are removed. For each time button 5 and 6 is pressed the program removes the worst outlier, recalculates the linear fit, and plots the new set of data points and the linear fitted line. When all the data points are in the neighborhood of the linear fitted line then all the outliers are removed.

8. By clicking on the "Mark" button all the peptides that were found to be outliers get a comment.
9. Set the checkbox "Recalibrate all ions" if fragment ions also should be recalibrated according to the linear fit.
10. Click the "Recalibrate" button and close the window.

To get an overview of the mass accuracy before the recalibration, click on "Analysis → Validation → Overall performance" and go to the page-tab "Misc." One can now search the data again by sequentially pressing "Alt"+"S" or by going back to the "Output" tab-page and clicking on "Start." The new search result should have a higher mass accuracy. Try running the function "Overall performance" again and compare with the previous result.

3.3. Repeated Search of the LC–MS/MS Spectra

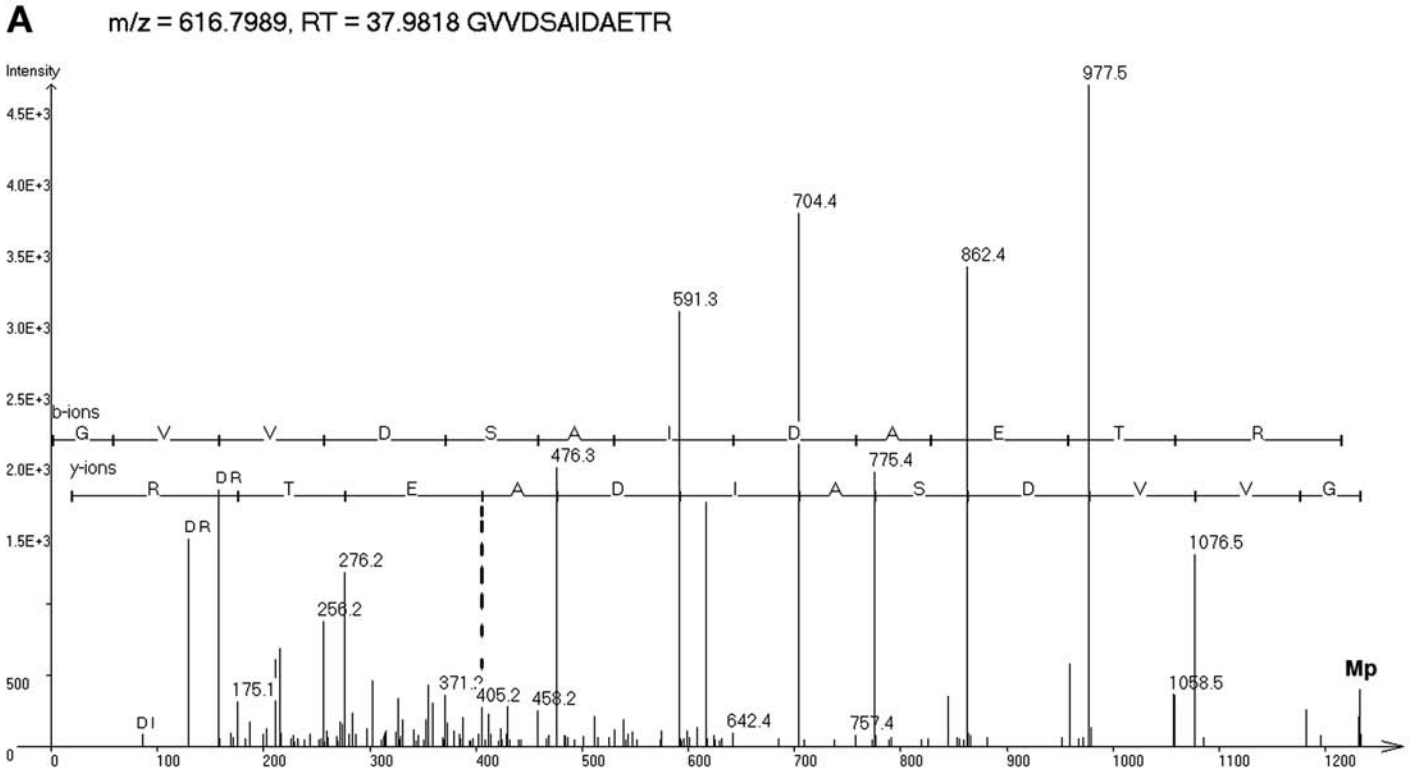
In the VEMS program one can repeated search MS/MS spectra that are not assigned in the first search. It is possible to specify new databases (*see* **Note 10**), mass accuracies, enzymatic cleavage patterns (*see* **Note 11**), and variable modifications for the repeated search. After the first search it is clear that the test data has much higher mass accuracy than what was specified for the first search, and because the search space now will be dramatically expanded it is advantageous to specify the mass accuracy as precisely as possible to minimize random matches. The following steps will show how to search the identified protein sequences for all possible miscleavages and all the variable modifications in the listbox at the setting window. The list of variable modification is stored as a text file in the VEMS directory. One can load a new list by right-clicking on the listbox and choosing "load."

1. Go to "Tables" page-tab and click on the "Settings" button. Set the mass accuracy to 0.1 Da for both the fragment and the parent ion in the top of the window.
2. Go to "Analysis → Re-search data → Un-specific and variable."

Now one should obtain more matched peptides. The number of extra matched peptides will depend on the size of the search space (*see* **Note 12**). It is highly recommended to manually validate the new assigned peptides.

3.4. Validating the Search Results

Validation of the database-dependent search result is an important part of analysis. All database-dependent search algorithms, to date, make errors. In general, the errors are only found in the low-scoring assignments. However, one has to remember that the database-dependent search is in principle an optimization process. This means that if a spectrum of a peptide is searched against a database which does not contain the peptide, then the spectrum can be matched


A
m/z = 616.7989, RT = 37.9818 GVVDSAIDAETR
Intensity
4.5E+3
4.0E+3
3.5E+3
3.0E+3
2.5E+3
2.0E+3
1.5E+3
500
0
b-ions
G V V D S A I D A E T R
y-ions
R D R T E A D I A S D V V G
977.5
704.4
862.4
591.3
476.3
775.4
1076.5
276.2
256.2
175.1
371.2
405.2
458.2
642.4
757.4
1058.5
Mp
0
100
200
300
400
500
600
700
800
900
1000
1100
1200

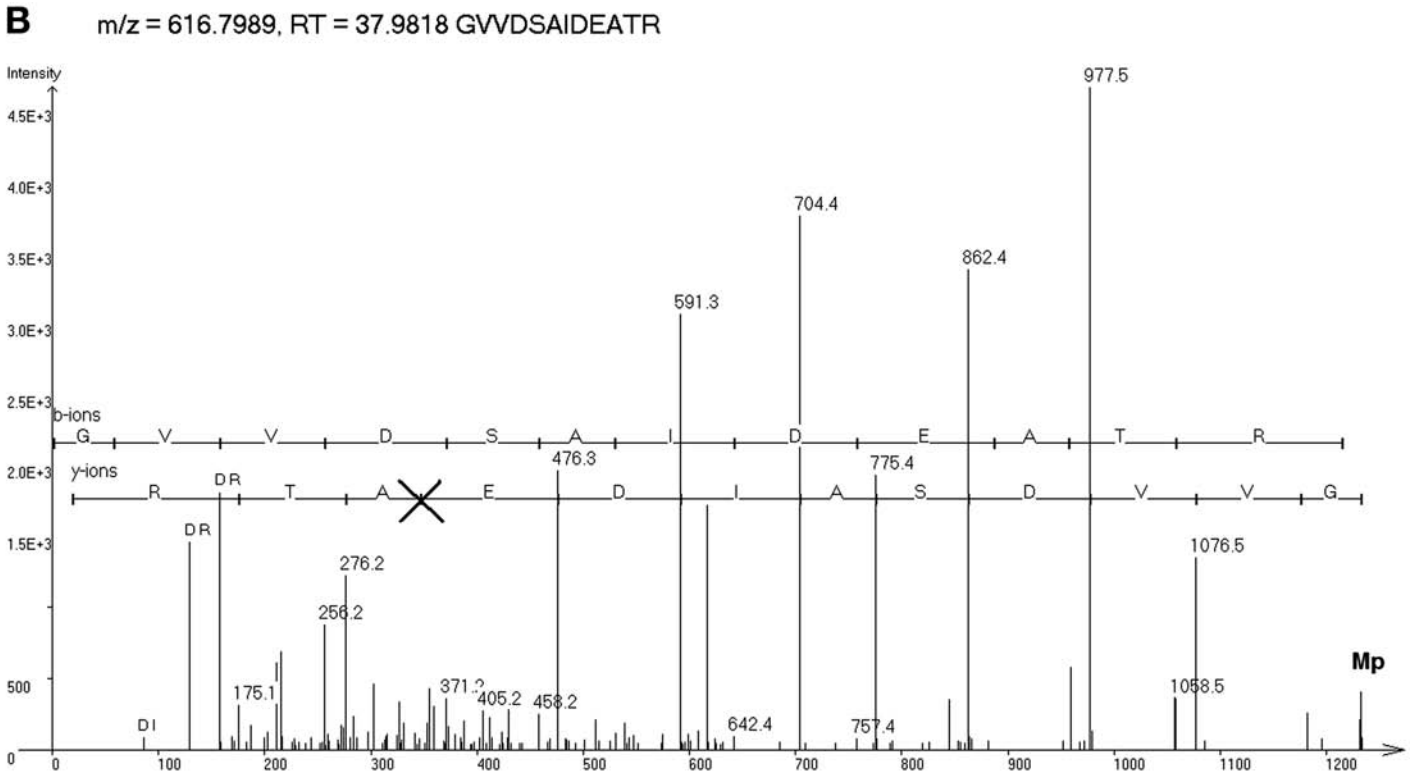


Fig. 3. Different peptides with similar MS/MS spectra. The correct solution **(A)** explains the peak 405.2 Da, whereas the solution in **(B)** does not.

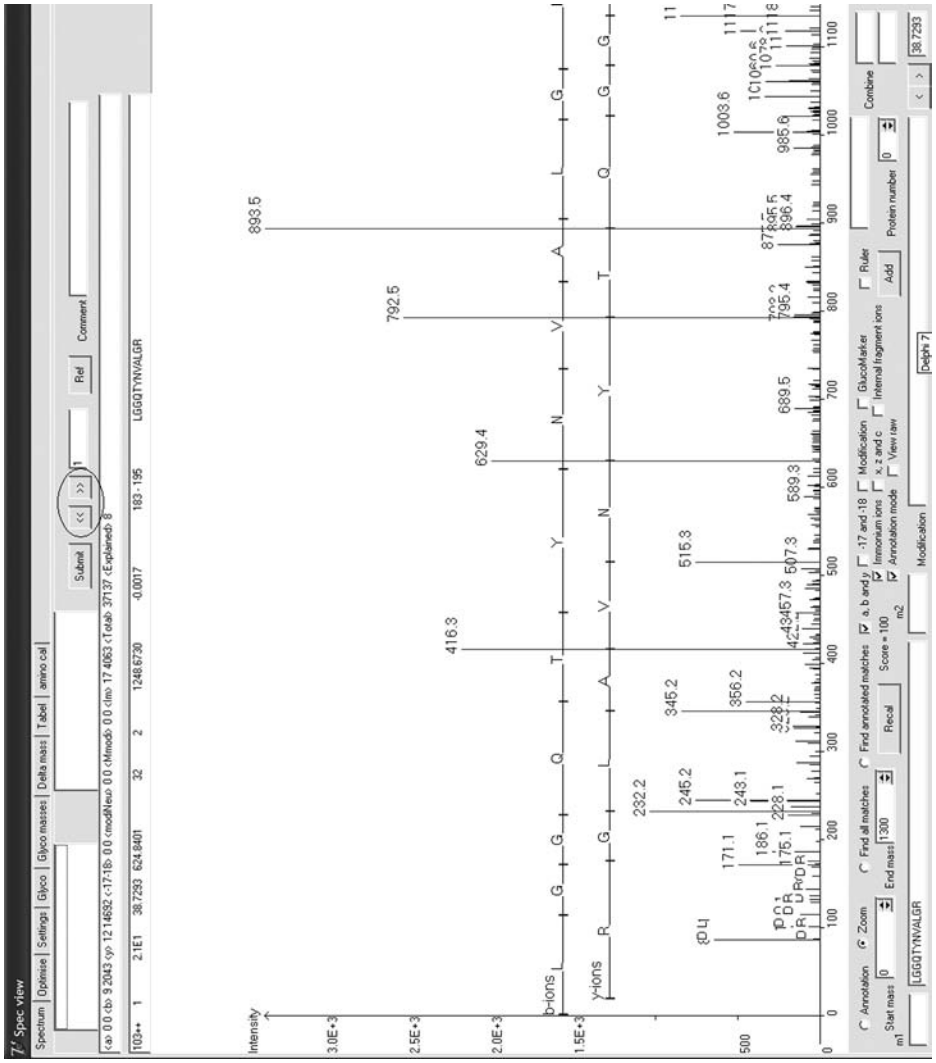


Fig. 4. The spectrum viewer window. In this window the spectrum is display and the peaks automatically annotated by the VEMS program. The button for browsing through all the peptide assignments found by the database-dependent search is highlighted by the ellipse.

to a peptide that has a MS/MS spectrum that is quite similar to the correct peptide (*see* **Fig. 3**). It is also not possible to distinguish isoleucine and leucine in low energy collision-induced dissociation, so the peptides ISTVGEK and LSTVGEK would have the same spectrum. In addition, most peptides are redundant, meaning that they occur in different proteins. Owing to these shortcomings

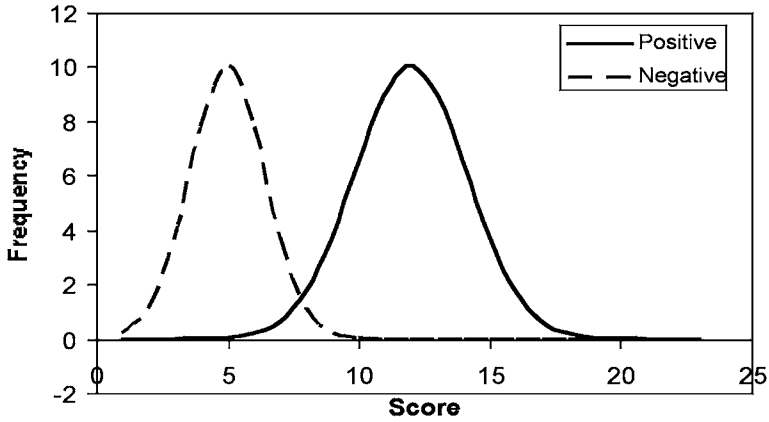


Fig. 5. Score distribution for the negative and positive dataset. The further the two distributions are separated, the better the scoring function is.

it is recommended to have at least two peptides for protein identification. In VEMS there are a number of filters so one can hide or remove search results that do not fulfill different validation criteria. These can be found on the tab-page "Validation." In the output window one can right-click on a peptide assignment and choose "View spectrum." This will show the spectrum viewer. The spectrum viewer displays the spectrum with automatic peak assignments (*see* **Fig. 4**). From the spectrum viewer one can browse through all peptide assignments by clicking the buttons "<<" and ">>."

3.5. Quality Test of Scoring Functions

In VEMS there is the possibility to make custom-made scoring functions as dynamic link libraries (dlls). If appropriate test set of MS/MS spectra is available then VEMS can evaluate the different scoring functions by plotting ROC curves (receiver operate characteristics). The test set should consist of two populations of MS/MS spectra. One population is composed of MS/MS spectra for which the corresponding peptide is present in the searched database (positive set). The other population consists of MS/MS spectra of which the corresponding peptide is known not to be present in the searched database (negative set).

The ROC curves were invented for radio communications in the 1960s and have recently been used in several publications for comparing different scoring functions ***(11)***. It is of great importance to have a good positive and negative set. The positive set used to test VEMS is composed of spectra from known purified proteins and the negative set was made of spectra corresponding to in vitro acetylated peptides. In the test search acetylation is then not specified as a variable modification, so whatever the search and scoring algorithm finds on

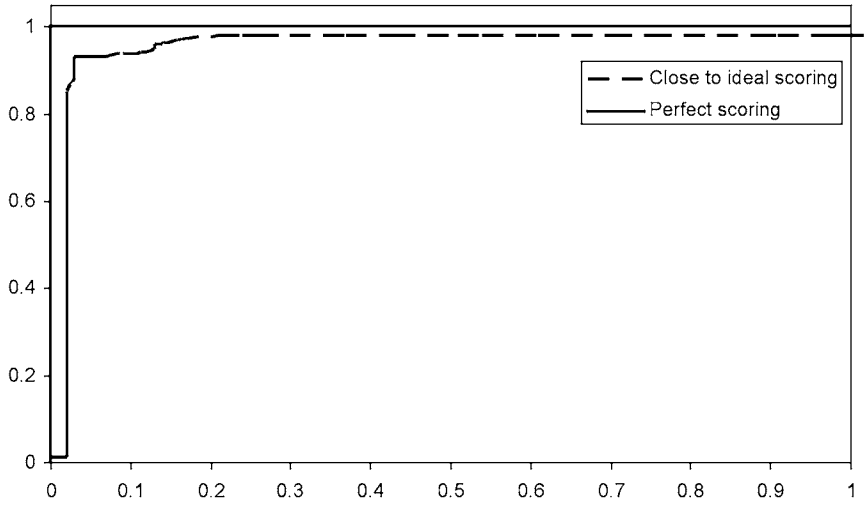


Fig. 6. ROC curves made by testing the same test set with two different scoring functions.

the spectra from the negative set will be wrong. After searching the complete test set (both the positive and negative), the resulting plot of scores may have the appearance shown in **Fig. 5**. The further the two distributions are apart the better the scoring function is. However, the ROC curve is a better way to visualize the performance of a scoring function. In the ROC curve the true-positive rate is plotted vs false-positive rate at different score thresholds (**Fig. 6**). The closer to the upper-left corner the curve is, the better the scoring functions.

The ROC curves can be constructed from the two score distributions in **Fig. 5** by starting with a high threshold where the rate of true and false-positives are zero. This could correspond to a threshold score of 20 in **Fig. 7A** and would correspond to the point (0,0) in the ROC curve in **Fig. 7B**. Lowering the threshold to say 15 would give approximately a true-positive rate of 12% and still no false-negatives. Lowering the threshold further would result in false-positives. For example, a threshold of 10 would give 14% of the false-positives and 88% true-positives point (0.14, 0.88) in the ROC curve.

3.6. Significance of Peptide Assignments

There exist four basic ways of testing the significance of a peptide assignment. First, a probability model can be made that determines the probability that the observed spectrum can be explained from a proposed peptide solution. Such a model is only dependent on how well the theoretical spectrum correlates with the observed spectrum and how independent it is from the size of the searched database, the amount of data searched, and the number of variable

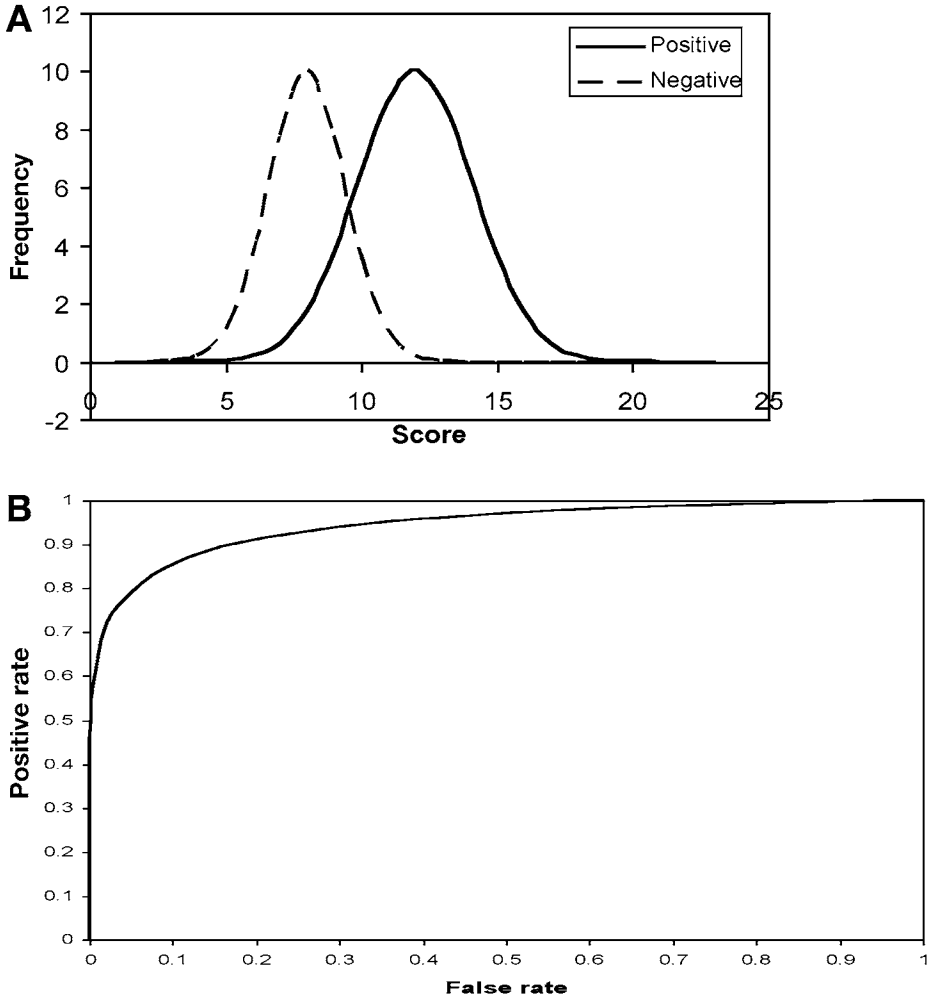


Fig. 7. **(A)** Hypothetical score distribution for the negatives and positives. **(B)** The corresponding ROC curve to the distributions in **A**.

modifications searched. Such a model is best if one can explain all peaks in the observed MS/MS spectra. However, it is often the case that several intense peaks are left unexplained making such a probability model risky. An alternative probability model could be to determine the false-positive rates by searching, for example, the same reversed protein sequence databases with the same data and search settings. Such models have been claimed to be dependent on how well the theoretical spectrum correlates with the observed spectrum, the size of the searched database, the amount of data searched, the number of variable modifications searched, and the search settings ***(12)***. It has also been claimed

that such models can model the complexity of isoforms because all protein isoforms are still present ***(12)***. However, the isoform information is only maintained in the reverse sequence direction and has no or little relevance to the MS data that correspond to sequences in the correct forward direction. A third possibility is to make survival analyses where the low-scoring matches during the search are recorded and used to estimate the probability function for random matches ***(2,3,13)***. The low-scoring matches are assumed to be random. Such a model would truly maintain isoform information in the correct sequence direction. However, such a probability model has the drawback that it requires searches of large amounts of data against large sequence databases in order to have a high enough frequency to make a good estimate of the probability for random hits. Recently, Wany et al. ***(14)*** proposed an elegant algorithm to generate random isobaric peptides of which the score is used to estimate the probability function for random matches rather than using sequences in a protein sequence database. The description given here is slightly modified compared with the one given by Wany et al. The algorithm consists of two steps. First, a mass array using the integer data type of size 400,000 is set up by multiplying all mass with 100. This gives a mass array that covers from 0 to 4000 Da and has a mass accuracy of 0.01 Da. Each entry in the mass array can be true or false. If a theoretical peptide mass corresponding to the mass of the entry exists, then it is set to true, if not then it is set to false. The array can now be computed in linear time by the following recursion:

$$A[i] = \begin{matrix} true\ if\ \ A[i - residue\_mass(aa)x100] = true \\ false\ if\ \ A[i - residue\_mass(aa)x100] = false\ and\ \ A[i] = false \end{matrix}$$

where A[1901] is initialized to true and *residue_mass* returns the residue mass of an amino acid. Iterating through the 400,000 entries in *A* and using the equation completes the array. In the next step, the mass array *A* is traced backward from a given theoretical peptide mass. The back-tracking technique can generate many random sequences of near-isobaric mass very fast. One can further elaborate on the algorithm by using the amino acid frequency for the organism in study for the back tracking.

The last possibility is to make a "theoretical" probability function to describe the probability of random hits. However, all theoretical models so far do not completely agree with reality and it has been claimed that fully theoretical models are not practical because of the large amount of parameters that affects them ***(15)***.

3.7. Grouping LC–MS/MS Data

The LC–MS/MS datasets can be grouped according to which sample they belong. It often happens that one cuts out spots or bands from a gel, digests with trypsin, and performs LC–MS/MS runs of each spot or band. In VEMS there is

no need to search runs corresponding to one spot independently because the LC–MS/MS runs can be grouped. This feature is especially important for quantitative time studies (*see* Chapter 8). Grouping is made on the page-tab "LC–MS grouping." It can be done manually or automatically by giving the number of groups in the spinedit "Group" and clicking on "Automatic" (automatic grouping is only possible if the number of LC–MS/MS runs can be grouped in equal size).

3.8. Interrogating the LC–MS Profile

In all database-dependent search programs the full information in a LC–MS profile is not used. In VEMS the use of information in the LC–MS profile in conjunction with the processed MS/MS spectra have been initiated. For the current version, only manual interrogation of the LC–MS profile is available. There are many examples of useful information hidden in the LC–MS profile. It can reveal overlapping elution time of isobaric peptides, reveal buffer contaminations, locate isotope-tagged molecules, and locate peptides with multiple phosphorylations. Once isotope-tagged molecules or potential phosphorylated peptides have been located they can be exported into an inclusion list for future LC–MS/MS runs.

3.9. Higher Level Data Analysis

VEMS can export all protein identifiers with associated GO annotations and a matrix with quantitations values so that it can be easily imported into the bioinformatic and statistical R package ***(15)***. The Bioconductor ***(16)*** R package has many tools for knowledge extraction from various biological databases, in addition to various graph packages that can be used for visualization. The package GOCluster can be used to statistically analyze whether specific GO categories are significantly up- or downregulated ***(17)***.

4. Notes

1. For grouping the proteins VEMS blasts all identified proteins against all indentified proteins. The resulting expectation values or number of similar amino acids obtained from this blast result is used to group the indentified proteins.
2. TandemX is a free search engine that can do database-dependent searches of MS/MS data. TandemX is used to validate and compare search results obtained by the VEMS program.
3. Luetfisk is a *de novo* sequencing program. Whereas VEMS and TandemX use sequence databases for interpreting the MS/MS data, Lutefisk only uses information in spectra for interpretation.
4. BioEdit is a general purpose sequence handling program. VEMS can, via the clipboard, export protein sequences to BioEdit. BioEdit has many bioinformatics tools such as the ClustalX sequence alignment tool, grouping sequences

into families, protein hydrophobicity/hydrophilicity plots, and graphical annotation of sequences.

5. SAM is an Excel add-in for finding statistically significant regulated genes based on multiple comparisons of microarray quantitative data. In VEMS, the same methods are used for finding significantly regulated proteins or posttranslational modification based on multiple comparisons of MS intensity data.
6. PLGS is a commercial program for data processing, data searching, quantitation, and data storage. VEMS interfaces to some of the data processing tools in PLGS.
7. Missed cleavages occur when trypsin does not cleave a site fully. Missed cleavages are often observed if there are basic or acidic residues in the proximity of the Lys or Arg. It is often assumed that trypsin does not cleave Lys and Arg if the next amino acid is proline. However, one can often observe a peptide with low intensity corresponding to the cleaved peptide. In VEMS v3.0, the algorithm assumes that it is cleaved; however, both versions can be found if one specifies a minimum of one missed cleavage in the search parameters.
8. Fixed modification is defined as a modification that occurs in all cases on the specified amino acid. Variable modification, on the other hand, may occur but will not occur in all cases of the specified amino acid.
9. VEMS has two search algorithms. The one demonstrated in this chapter iterates through all sequences in the specified FASTA database and compares the peptide sequence with an index set of MS/MS spectra. This algorithm is best when the searched database is large and many variable modifications are searched. The other algorithm iterates through all the spectra and compares the spectra with an indexed sequence database. See the quick guide documentation (http://yass.sdu.dk) for details on using indexed databases.
10. In VEMS one can chose a new database to search against unmatched spectra. This may be relevant if contamination from other species is suspected. A well-known example is contamination with keratin from human hair or from autolysis products from trypsin. Once the contaminations are known, the sequences can be combined with the species-specific databases searched and the search can be redone (*see* **Note 12**).
11. Miscleavage is here defined as nonstandard trypsin cleavages. This can be caused by pseudotrypsin (*see* **Subheading 1.2.** in Chapter 1;) or by other proteases from the sample. VEMS finds miscleavage by searching all possible peptide fragments without any cleavage rules or as semi-tryptic, meaning that the C-terminal part of the peptide has a standard trypsin cleavage site but the N-terminal is unspecificly cleaved.
12. The number of possible modifications, *m*, of a peptide with length *n* is given by:

$$m = \prod_{i=1}^{n} (V_i + 1) \tag{1}$$

n is the number of residues in the peptide, V_i is the number of possible variable modifications at residue *i*, and *i* iterates over the sequence.

Acknowledgments

R. M was supported by grants from EU TEMBLOR and by Carlsberg Foundation Fellowships.

References

1. Listgarten, J. and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell Proteomics* **4,** 419–434.
2. Matthiesen, R., Lundsgaard, M., Welinder, K. G., and Bauw, G. (2003) Interpreting peptide mass spectra by VEMS. *Bioinformatics* **19,** 792–793.
3. Matthiesen, R., Bunkenborg, J., Stensballe, A., Jensen, O. N., Welinder, K. G., and Bauw, G. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* **4,** 2583–2593.
4. Matthiesen, R., Trelle, M. B., Højrup, P., Bunkenborg, J., and Jensen, O. N. (2005) VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J. Proteome Res.* **4,** 2338–2347.
5. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
6. Robertson, C. and Ronald, C. B. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467.
7. Taylor, J. A. and Johnson, R. S. (2001) Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* **73,** 2594–2604.
8. Tippmann, H. F. (2004) Analysis for free: comparing programs for sequence analysis. *Brief Bioinform.* **5,** 82–87.
9. Garden, P., Alm, R., and Hakkinen, J. (2005) PROTEIOS: an open source proteomics initiative. *Bioinformatics* **21,** 2085–2087.
10. Tusher, V. G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* **98,** 5116–5121.
11. Colinge, J., Masselot, A., Cusin, I., et al. (2004) High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. *Proteomics* **4,** 1977–1984.
12. López-Ferrer, D., Martínez-Bartolomé, S., Villar, M., Campillos, M., Martín Maroto, F., and Vázquez, J. (2004) Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST. *Anal. Chem.* **76,** 6853–6860.
13. Eriksson, J. and Fenyo, D. (2004) The statistical significance of protein identification results as a function of the number of protein sequences searched. *J. Proteome Res.* **3,** 979–982.
14. Wany, Y., Yangz, A., and Chen, T. (2006) PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. *Anal. Chem.* **78,** 432–437.

15. http://www.r-project.org/. Last accessed 05/27/2006.
16. Gentleman, R. C., Carey, V. J., Bates, D. M., et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80.1–R80.16.
17. Wrobel, G., Chalmel, F., and Primig, M. (2005) goCluster integrates statistical analysis and functional interpretation of microarray expression data. *Bioinformatics* **21,** 3575–3577.

8

Quantitation With Virtual Expert Mass Spectrometrist

Albrecht Gruhler and Rune Matthiesen

Summary

Quantitative analysis of proteins and peptides by mass spectrometry has been greatly advanced by the development of proteomic technologies within recent years. Particularly, labeling of peptides and proteins with stable isotopes such as ^{2}H, ^{15}N, and ^{13}C facilitated the unbiased comparison of protein amounts in distinct samples in a single mass spectrometric experiment. These methods can be applied to detect quantitative changes in protein amounts and posttranslational modifications such as phosphorylation. Quantitation of mass spectra requires accurate and efficient bioinformatics tools, which can match corresponding peptides, determine peak intensities, and calculate relative amounts. In this chapter, we describe the use of virtual expert mass spectrometrist for the quantitation of mass spectra from samples with peptides and proteins encoded with stable isotopes.

Key Words: Quantitation; SILAC; chemical labeling; stable isotopes; virtual expert mass spectrometrist; VEMS.

1. Introduction

Mass spectrometry (MS)-based quantitative analysis of proteins and peptides has become an important proteomic tool within recent years ***(1–3)***. Quantitative proteomics can be employed to study many complex biological processes because it enables the unbiased measurement of changes in protein abundance in response to a certain stimulus. Applications are manifold and include, for example, the comparison of wild-type to mutant genes, the examination of different developmental stages, or the analysis of cell differentiation. Of special interest is the quantitative analysis of posttranslational modifications such as phosphorylation because most signaling processes are regulated by protein phosphorylation and dephosphorylation. Recent examples of the successful application of quantitative phosphoproteomics include the study of the activation of a yeast mitogen-activated protein kinase pathway by α-factor and the investigation

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

of the kinetics of extracellular growth factor receptor signaling ***(4,5)***. Many clinical applications of proteomics also require quantitative protein analysis. One such example is the search for biomarkers for specific diseases in body fluids and analysis and characterization of tumor tissues by MS techniques ***(6,7)***. Yet another example for the application of quantitative proteomics is the study of the kinetics of protein transport between different intracellular compartments ***(8)***.

Comparative quantitation of peptides and proteins by MS can be achieved by marking proteins or peptides with stable isotopes and thereby introducing a mass difference without changing the chemical properties of the peptides, which would influence their analysis by MS. Labeled and unlabeled samples can be mixed and analyzed together in the same mass spectrometric experiment. Intensities of the peptide pairs are then used to calculate relative peptide amounts and to monitor changes in abundance. Encoding of proteins with stable isotopes can be performed in cell culture by growing the cells in the presence of stable isotopes, either by adding a nitrogen source containing ^{15}N ***(9,10)***, or by substituting amino acids with labeled counterparts such as $^{13}C_6$-arginine or $^{2}H_3$-leu (a method termed SILAC for stable isotope labeling by amino acids in cell culture) ***(11,12)***. Alternatively, proteins and peptides can be labeled by chemical derivatization in vitro. A number of reagents exist with distinct properties and affinities for functional groups, such as thiols or free amines ***(13)***. Some of the more commonly used ones are the isotope-coded affinity tag (TCAT) reagent ***(14)*** that reacts with cysteines and the amine-reactive isobaric tagging (iTRAQ) reagent ***(15)***, which is specific for free amines. All of these reagents exist in a light and at least one heavy form that contains multiple ^{2}H, ^{15}N, or ^{13}C, allowing the quantitative comparison of two or more samples by MS The advantage of chemical derivatization is that it can be applied to all samples, even to primary cells and tissues. This is not possible for the in vivo labeling, which in most cases is limited to cell cultures. There are some drawbacks, e.g., the need for additional sample handling, often entailing sample loss and the possibility of incomplete labeling or side reactions.

However, isotopic labeling is not practical in all cases, so some approaches have been developed to quantitate the MS spectra of unlabeled samples. A rough estimate of protein abundance can be achieved by comparing the number of identified peptides for an individual protein in different samples: this number generally decreases at lower concentrations, because the sensitivity of the MS instrument is not sufficient to identify every peptide. In another approach, peak intensities originating from the same peptides are compared with each other in different samples analyzed under identical conditions by matrix-assisted laser desorption and ionization MS or liquid chromatography (LC)-electrospray ionization–MS ***(16)***.

All those approaches have in common the need for dedicated software for the processing of mass spectra and the calculation of quantitative data. The

comparison of several LC–MS experiments requires the matching of identical peaks in spite of slightly variable retention times of the LC analysis, the normalization of the individual LC–MS runs, and the comparison of identical peak intensities. Likewise, quantitation by stable isotope labeling requires the matching of peak pairs and the estimation of peak intensities of the heavy and light peptides. Virtual Expert Mass Spectrometrist (VEMS) is a program developed for searching mass spectra (*see* Chapter 7) and the quantitation of peptides and proteins. In this chapter, we describe the use of VEMS for the quantitation of stable isotope-labeled peptides and proteins from mass spectra.

2. Materials

2.1. Required Software

1. VEMS v3.0 (http://yass.sdu.dk).
2. Microsoft Windows. Currently only fully tested on Windows XP and Windows 2000.
3. ProteinLynx global server (PLGS) v2.0.5 (*see* **Note 1**) is a commercial program that can be obtained from Waters (Milford, MA). VEMS interfaces to some of the raw data processing tools of PLGS v2.05. However, VEMS also has built-in functions for raw data handling. Alternatively, MS data in mzXML (http://tools.proteomecenter.org/) can be converted to the VEMS raw data format and used for quantitation. If PLGS v2.2 is used instead of PLGS v2.05 then the directory "C:\PLGS2.2" should be renamed to "C:\PLGS2."
4. A test dataset containing a result file including all the relevant settings (ResSILAC.txt), raw data in VEMS format (MS_SILAC_VEMS), and the processed peak list file (SILAC.pkx) can be found at http://yass.sdu.dk/SILACtest/SILAC.zip.

3. Quantitation of Isotope-Marked Peptides With VEMS

3.1. Requirements for Quantitation

Quantitation of peptides with VEMS requires that the MS or tandem mass spectrometry (MS/MS) data have been searched against a sequence database. Installation of VEMS and its use for searching MS/MS spectra against databases is described elsewhere in this volume (*see* Chapter 7). LC–MS/MS runs of complex samples contain many isobaric peptides and the correct identification of isotope-labeled and unlabeled-peptide pairs in an LC–MS profile relies both on parent ion masses and the retention time of the peptide on the reverse phase column. The peak list file used for searching peptides against databases should therefore include the retention times. This can be achieved by using the PKX format (*see* Chapter 2) or by using the mgf format for MASCOT searches. In addition to the search result and PKX files, the raw LC–MS or LC–MS/MS data are required, either in the Micromass raw data format or the VEMS raw data format (*see* Chapter 2 to convert mzXML to VEMS raw). The following

steps explain how to quantify peptides and proteins. The search result file ResSILAC.txt is used as an example.

1. Load the result file ResSILAC.txt. Go to "File → Open data → Import from database" and choose the file. Alternatively go to "File → Open annotated data" and choose the file by left-clicking on it in the bottom listbox and press the button "Load." The loaded file already contains all the necessary settings. The explanation of the various settings is given next.
   The VEMS program needs to know where the raw data is located. Therefore, the file path to the raw data should be updated. Go to "File → Open data → Open multiple spectra" (*see also* Chapter 7 for a detailed explanation). Right-click in the listboxes on the right of the window and choose "clear." Now use the listboxes on the left to specify the new file paths for the files SILAC.pkx (file name appears in the window "Multiple processed MSMS spectra") and MS_SILAC_VEMS (in the middle window "raw data") and close the window. The order of PKX and corresponding raw data files has to be identical, as otherwise unrelated files will be associated with each other. Note also that the VEMS program counts the number of files with processed spectra and raw data. This is useful for validating that the files are correctly specified when many files are processed.
2. Click on the radiobutton "View quantification" (**Fig. 1**, no. 1) in the output window to see the search result loaded. Values in the column named H/(L+H)% (*see* **Note 2**), which displays quantitation results, are 0.00 because the data is still not quantified.
3. Quantitation of peptides with VEMS requires that peptide sequences have been assigned to the fragmentation mass spectra from an LC–MS analysis, because peptide composition is necessary to calculate the masses of the isotope-labeled and unlabeled-peptide pair and for extraction of their ion currents from the corresponding mass spectra. However, because peptides elute from the reverse phase column over a time period longer than a MS/MS cycle of the mass spectrometer, the same peptide pair is usually observed in several consecutive mass spectra. Ratios between heavy and light peaks can slightly differ between the spectra because ions statistics are sometimes of lower quality in mass spectra measured at the start or end of the elution profile, where peak intensities are low. Therefore, quantitation should be performed over the whole elution range of a certain peptide pair in order to ensure the best possible accuracy. VEMS has a function that allows you to find the spectra with the highest peak intensities of the identified peptides: "Edit → Find elution maximum." Having performed this search, VEMS will start the quantitation of a peptide from the mass spectra with the highest intensity. This has the advantage that the probability of a wrong assignment of the corresponding peaks will be lowest. In spectra with low peak intensity and with high differences between the intensities of the peptide pair, background peaks can sometimes be erroneously assigned to peptide peaks, which will distort the quantitation.
4. In the settings window of VEMS, which is opened by "Input page-tab → Settings" (**Fig. 2**, no. 1), the parameters for the type of labeling and the stringency of peptide and peak selection, and peptide and protein quantitation can be defined. The listbox in the center of the settings window (**Fig. 3**, no. 1) contains a list with all the

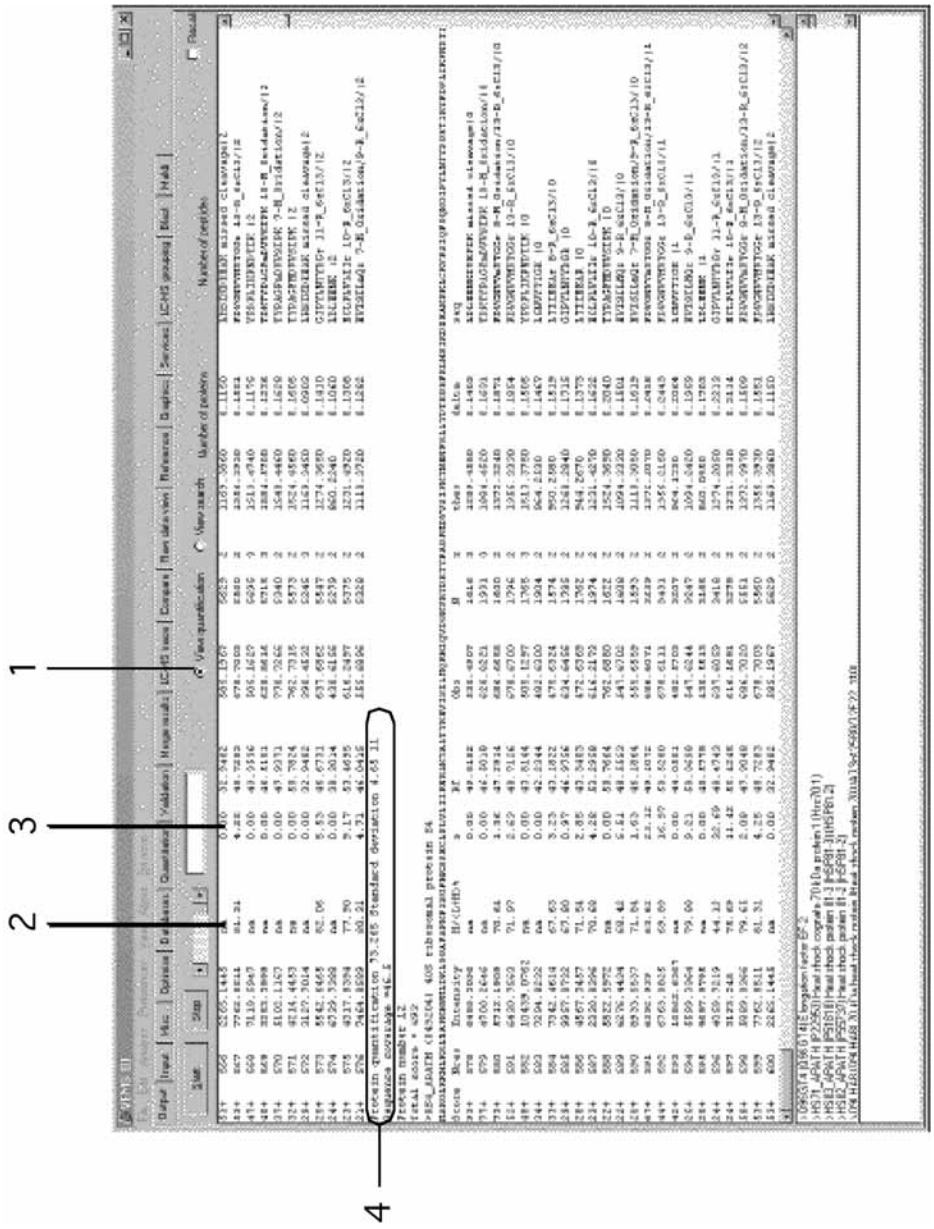


Fig. 1. The output window where the search and quantitation report is displayed.

available modifications, which are displayed in abbreviated form; for example, leu_deux3 stands for [$^{2}H_{3}$]-leucine, labeled with three deuterium instead of hydrogen atoms. Two other commonly used isotopic-labeled amino acids are [$^{13}C_{6}$]-arginine and [$^{13}C_{6}$]-lysine, which are represented by R_13C6 and K_13C6.

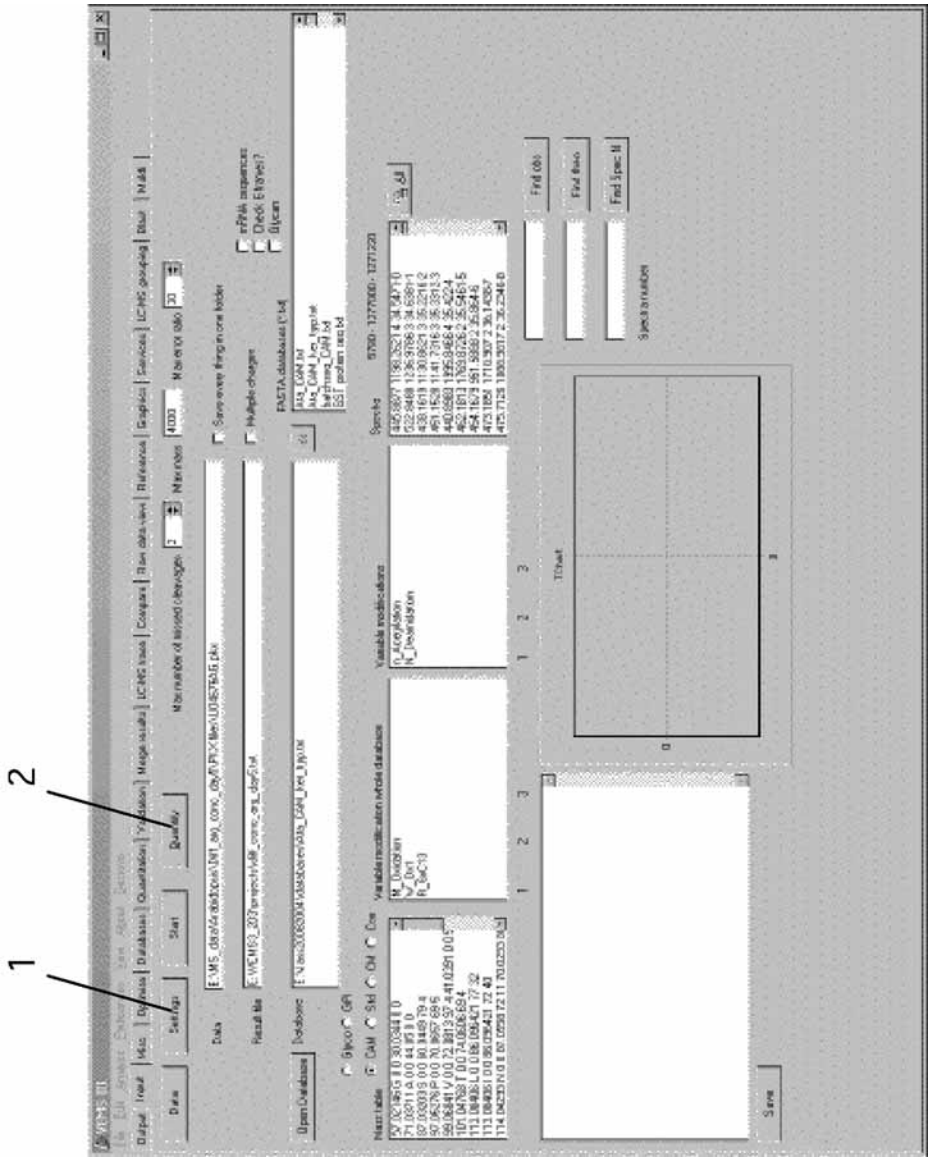


Fig. 2. Input page displaying which files and databases have been loaded and the peptide list.

These modifications are defined in the file “modiAll.txt” located in the VEMS directory to which user-defined modifications can be added, thereby providing great flexibility to incorporate novel reagents in VEMS. The list of modifications is used both for database searching and quantitation of peptides. To choose a SILAC

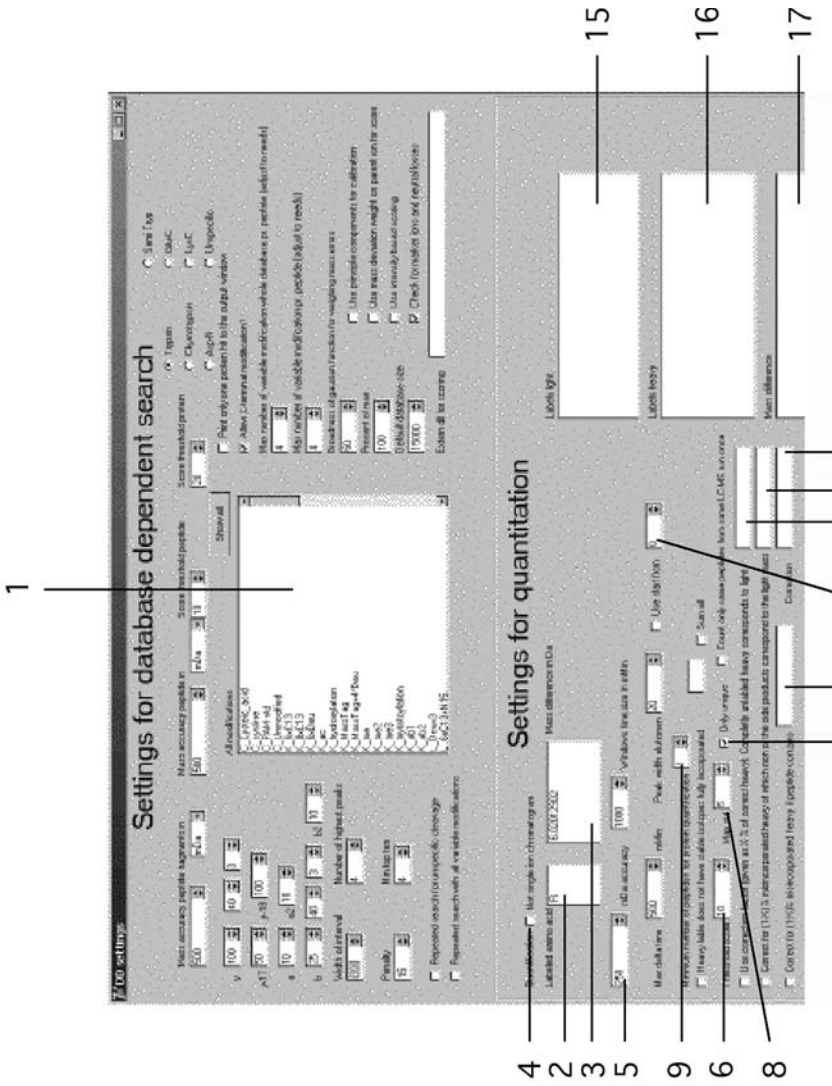


Fig. 3. Setting window where one specifies the search and quantitation settings.

modification for quantitation click on the modification to highlight it. Right-clicking will open a popup menu and choosing the option “Add to quantification” will enter the entry into the fields “Labeled amino acid” (**Fig. 3**, no. 2) and “Mass difference” (**Fig. 3**, no. 3), which displays the mass difference between modified and unmodified amino acids. All modifications that are present in this listbox will be applied to the quantitation of peptides in the results list. It is important that results have been searched with all the modified amino acids used for quantitation. If chemical

labeling is used then two modifications should be specified per tag used. By first highlighting the light version of the tag and choosing "Add light quantification label light." The same procedure is used to specify the heavy version of the tag but using "Add to quantification label heavy." The chosen light and heavy tags are now shown in **Fig. 3**, nos. 15–16 and the mass difference between the tags in **Fig. 3**, no. 17.

5. The fields in the lower part of the settings window determine the parameters used for quantitation. There are two different processing algorithms available, which are chosen by the field "Quantification □ Not single ion-chromatogram" (**Fig. 3**, no. 4). Setting a tick mark means that the program will analyze mass peaks from all MS spectra of the LC elution profile for the corresponding peptide. This will minimize the interference from intensity from co-eluting peptides with similar masses. If the tick mark is not set then the quantitation will be based on single-ion chromatograms.
6. The parameters of peak detection and peptide selection can also be adjusted in VEMS. The value in the field "mDa accuracy" (**Fig. 3**, no. 5) defines the maximal mass difference between the theoretical and observed mass of the identified peptides. The "Threshold score" (**Fig. 3**, no. 6) specifies the minimum score of a peptide to be included in the quantitative analysis and ticking "Only unique" (**Fig. 3**, no. 7), excludes nonunique peptides (*see* **Note 3**). In the field "Max Std" (**Fig. 3**, no. 8) the maximum value for the standard deviation between quantitation of the different peaks for a given peptide can be entered (*see* **Note 4**). The minimum number of peptides per protein (**Fig. 3**, no. 9) defines how many peptides are necessary for a protein to be quantitated and reported in the result list (*see* **Note 5**). Peptides in the search result list are numbered and quantitation can be started at a certain peptide number by entering this figure in the field "Use start from" and ticking off the check-box (**Fig. 3**, no. 10).
7. The remaining fields can be used for automatic correction of ratios, where appropriate. For example, if in a SILAC experiment the incorporation of an isotopically labeled amino acid has been incomplete, the calculated ratios between intensities of heavy and light peaks do not represent the actual relative peptide amounts in the sample. This is because part of the peptides from the labeled sample will be unmodified and the light peptide, therefore, will be over-represented. This can be corrected for by specifying the percentage of isotopic labeling (**Fig. 3**, no. 11). VEMS will then adjust the peak intensities of heavy and light peptides accordingly and calculate the correct ratios. In cases where incomplete labeling will not increase the amount of the light peptides (e.g., using chemical modification of peptides), the box below should be used (**Fig. 3**, no. 12).
8. The bottom fields on the settings pages deal with satellite masses that can occur in some cases. An example for this is the appearance of a peak with a mass 1 Da lower than the monoisotopic mass of the labeled peak in SILAC experiments. This satellite peak is caused by the small percentage of ^{12}C atoms in the isotope-encoded amino acids and its intensity varies between different amino acid batches. It is also known that excess of [$^{13}C_6$]-arginine can be converted by the cells to [$^{13}C_5$]-proline ***(17)***, which causes a satellite peak with a mass of +5 Da in proline-containing

peptides of the stable isotope-encoded sample. In this case all peptide quantitations of peptides containing proline need to be corrected. First, one needs to specify, using capital letters, the amino acids that affect quantitation (**Fig. 3**, no. 13). Next, one gives an estimate of the intensity of the satellite peak given as X% of the heavy peak for a peptide having only one occurrence of one of the affected amino acids (**Fig. 3**, no. 14).

3.2. Quantitation With VEMS and Analysis of Results

When the quantitation settings are specified, and search results and raw data files have been loaded, quantitation is started by pressing the "Quantify" button (**Fig. 2**, no. 2). A counter will appear on the top bar of the program informing about the progress of quantitation. If all peptides are quantitated, results can be listed by clicking on "View quantification" (**Fig. 1**, no. 1) in the output page. Two additional columns will appear in this report showing the values for quantitated peptides (**Fig. 1**, no. 2) and their standard deviation (**Fig. 1**, no. 3). Quantitation results for individual peptides are given as the percentage ratio H/(H+L) × 100% (*see* **Note 2**). Protein quantitation is performed by clustering the ratios of all quantified peptides and subsequent removal of outliers. Quantitated peptides with a standard deviation larger than a threshold value are not considered for protein quantitation, even if the values for their ratios are displayed in the results list. The threshold for outliers is defined by the maximal allowed peptide standard deviation in the settings page (**Fig. 3**, no. 8). The average of all quantitated peptides is calculated and displayed in the bottom line for each protein along with the standard deviation for protein quantitation and the number of quantitated peptides (**Fig. 1**, no. 4). Note that the standard deviation for protein quantitation is a measure of the differences between the peptides and is different from the standard deviation for individual peptides.

3.3. Validation of Peptide Quantitation

VEMS offers several options for the analysis and validation of peptide and protein quantitation results. "Analysis → Quantification → Plot quantification" will open a window that lists all quantified proteins and displays their ratios graphically. The minimum number of quantitated peptides per protein can be specified and graphs are updated by clicking on the "Plot" button. "Analysis → Quantification → Quantification overall" displays the statistics of the analysis in the page-tab "Misc," such as the number of quantitated peptides, outliers, and proteins.

An important feature of VEMS enables the manual validation of peptide quantitation. It is accessed by right-clicking on a highlighted peptide and choosing "View peptide for quantification" from the popup menu. The "Quantification" page (**Fig. 4**, no. 1) will open and display the extracted peptide peaks from the

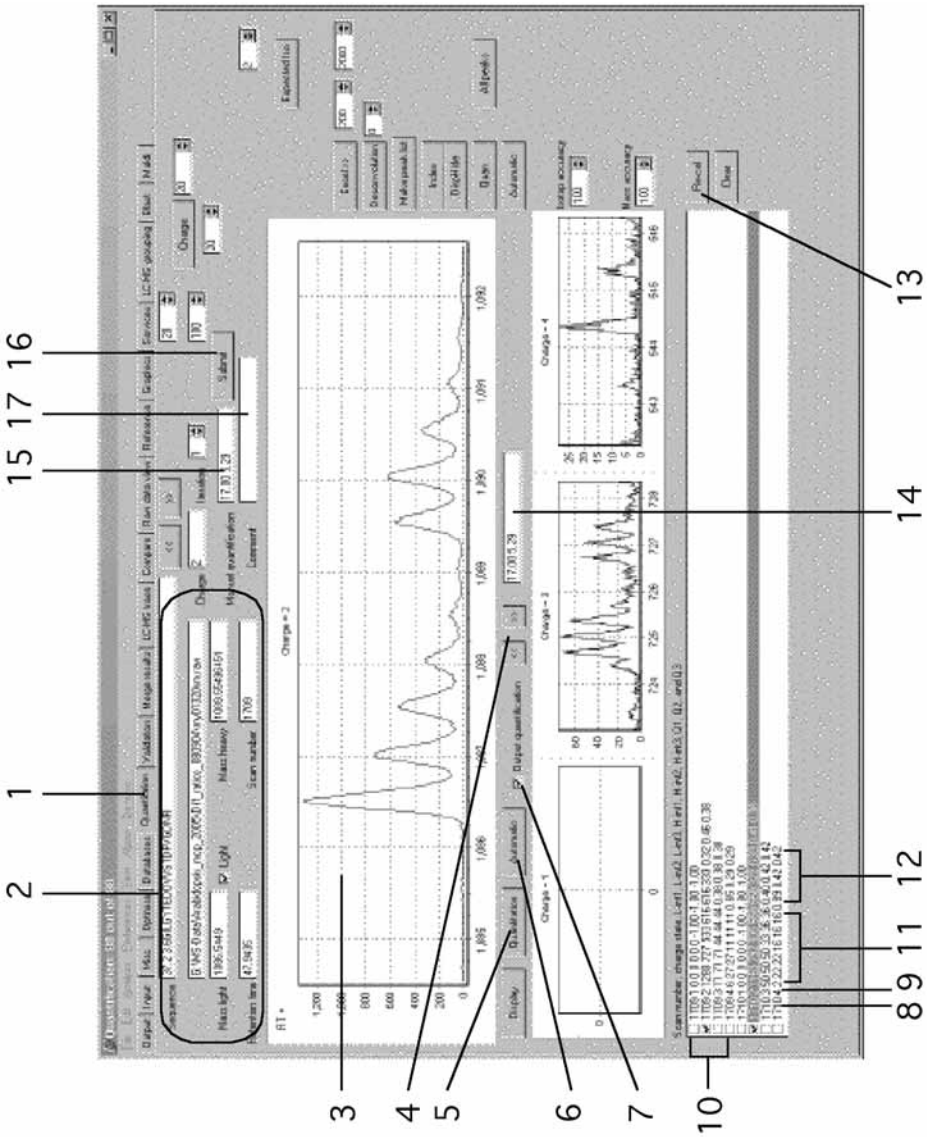


Fig. 4. Virtual expert mass spectrometrist (VEMS) window for validating peptide quantitations.

raw data file. The fields in the upper-left corner of the window display peptide sequence and masses for heavy and light peaks, retention time, scan number, and the corresponding file names (**Fig. 4**, no. 2). The large graph displays the peptide peaks in the charge state of the sequenced peptide (**Fig. 4**, no. 3), the other graphs

display the peptide peaks in other charge states, if they are present. Peaks are extracted from the mass spectra corresponding to the fragment spectrum of the identified peptide or to the spectra with the highest peak intensity, if the function "Find elution maximum" has been run. With the arrow buttons (**Fig. 4**, no. 4) one can move to the preceding or succeeding mass spectra. This feature allows one to inspect the peptide peaks for peak overlap, background noise, and isotope distribution and to ensure that peak pairs are correctly matched. It is also possible to quantitate peptides in this window, either for single spectra or over the whole elution profile, by clicking the "Quantitation" (**Fig. 4**, no. 5) or "Automatic" (**Fig. 4**, no. 6) buttons, respectively. The results are displayed as a matrix in the field below if the checkbox "Output Quantification" has been marked (**Fig. 4**, no. 7). The first column (**Fig. 4**, no. 8) contains the scan number of the mass spectra followed by the charge state of the peptides (**Fig. 4**, no. 9). There is one line for charge states one to four (**Fig. 4**, no. 10). The next six columns (**Fig. 4**, no. 11) display the intensities for three peaks (the monoisotopic mass peak and the two following isotope peaks) of the light peptide and the corresponding heavy peptide, followed by three values for the ratios H/(H+L) for each of the isotope peak pairs (**Fig. 4**, no. 12). Results for individual mass spectra and charge states can now be selected by ticking the box at the beginning of each row. Clicking on "Re-cal" (**Fig. 4**, no. 13) will calculate the average of all selected peptide quantitations and display them together with the standard deviation (**Fig. 4**, no. 14). These values can then be entered in the field "Manual Quantification" (**Fig. 4**, no. 15). Clicking on "Submit" (**Fig. 4**, no. 16) will enter these values in the results list replacing the existing values. If a peptide is to be excluded from quantitation from peptide overlaps for example, "na" can be submitted.

In the field "Comments" (**Fig. 4**, no. 17) an explanation for the correction can be entered. These functions of VEMS provide the possibility of easy and flexible editing and correction of quantitation results and simplify the manual validation of peptide quantitation.

3.4. Extended Analysis Functions

VEMS contains a number of features designed for the analysis of complex experiments consisting of more than one LC–MS analysis. Examples for this type of analysis are multidimensional chromatography, where the sample is separated into several fractions, or experiments with several samples for the different time points. VEMS keeps track of the source file for each peptide and provides a function for grouping individual LC–MS runs.

In order to compare individual samples, the sample source has to be added to the comment section of each peptide. This is done by choosing the "Analysis → Validation → Sample to comment" function. After updating the results view, the name of the PKX file is displayed at the end of each line (*see* **Note 6**).

The command "Analysis → Quantification → Quantification report over samples" will list the results for each sample in the "Compare" window. There is a line for each identified protein with quantitation results, standard deviation, and the number of quantitated peptides in each sample. This list can be exported to Microsoft Excel by right-clicking in the window and choosing "Export to Excel."

The quantitation of sample groups can be analyzed in the same manner. LC–MS runs can be grouped in the "LC–MS grouping" window. Clicking on "Import" will display a list of all loaded raw data files. The active group number is shown in the top-left spinedit box. Right-clicking on a highlighted file name allows it to be added to a group. It will appear in the window on the right side with its group number after the file name. After assigning all (or selected) files to groups, the grouping function is activated by checking "Use grouping." The group annotation must then be entered to the comments field of the results list using "Analysis → Validation → Group to comment" (*see* **Note 7)**. The data for protein quantitation of the individual groups are displayed in the "Compare" window using "Analysis → Quantification → Quantification report over samples" as described for individual samples. If the result is saved then the grouping information is also saved.

4. Notes

1. PLGS is a commercial program for data processing, data searching, quantitation, and data storage. VEMS interfaces to some of the data processing tools in PLGS.
2. Quantitation results in VEMS are given as the ratio between the intensity of the isotope-labeled peptide and the sum of the intensities of the labeled and nonlabeled peptides expressed as percentage $Q = H/(H + L) \times 100\%$. This value can easily be converted to the ratio H:L by the formula $H/L = Q/(1 - Q)$. A value of $Q = 50\%$ corresponds to equal amounts of isotope labeled and unlabeled peptides. $Q = 66.6\%$ corresponds to a twofold excess of the stable isotope-labeled form, and $Q = 33.3\%$ to a twofold excess of the unlabeled peptide.
3. The uniqueness of a peptide is database dependent, because redundant databases (including the NCBI nonrepetitive database) may contain several versions of the same protein, thereby rendering peptides from these proteins nonunique.
4. The ratio between heavy and light peptide peaks is calculated by comparing the intensities of the two monoisotopic mass peaks and the two following isotope peaks of the heavy and light peptide to each other. Because of peak overlap with unrelated peptides of similar mass and background peaks or shifts in the isotopic distribution between heavy and light peptide peaks, the ratios can differ for the three peak pairs. Large discrepancies indicate errors in the quantitation and should therefore be manually validated. The specified value for the peptide standard deviation defines the maximal difference between the three ratios.
5. Every peptide of the results list will be processed, meaning that if a peptide has been sequenced and quantitated several times, it will also count several times, if it

has not been removed from the results list previously. Likewise, if both the heavy and light forms have been identified, they will count as two quantitated peptides. In order to only count the same peptide once for each LC–MS run, go to settings and click the checkbox "Count only same peptide from same LC–MS run once."

6. Updating views can be done by "Analyze → Refresh view," or by clicking the radiobuttons "View quantification" or "View search."
7. If the some of the peptides have comments already attached then these should be removed by "Edit → Remove comments" before running "Analysis → Validation → Group to comment."

Acknowledgments

R. M was supported by grants from EU TEMBLOR and by Carlsberg Foundation Fellowships.

References

1. Matthiesen, R., Bunkenborg, J., Stensballe, A., Jensen, O. N., Welinder, K. G., and Bauw, G. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* **4,** 2583–2593.
2. Goshe, M. B. and Smith, R. D. (2003) Stable isotope-coded proteomic mass spectrometry. *Curr. Opin. Biotechnol.* **14,** 101.
3. Lill, J. (2003) Proteomic tools for quantitation by mass spectrometry. *Mass Spectrom. Rev.* **22,** 182–194.
4. Blagoev, B., Ong, S. E., Kratchmarova, I., and Mann, M. (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat. Biotechnol.* **22,** 1139–1145.
5. Gruhler, A., Olsen, J. V., Mohammed, S., et al. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell Proteomics* **4,** 310–327.
6. Chaurand, P., Schwartz, S. A., Reyzer, M. L., and Caprioli, R. M. (2005) Imaging mass spectrometry: principles and potentials. *Toxicol. Pathol.* **33,** 92–101.
7. Veenstra, T. D., Conrads, T. P., Hood, B. L., Avellino, A. M., Ellenbogen, R. G., and Morrison, R. S. (2005) Biomarkers: mining the biofluid proteome. *Mol. Cell Proteomics* **4,** 409–418.
8. Andersen, J. S., Lam, Y. W., Leung, A. K., et al. (2005) Nucleolar proteome dynamics. *Nature* **433,** 77–83.
9. Langen, H., Takacs, B., Evers, S., et al. (2000) Two-dimensional map of the proteome of Haemophilus influenzae. *Electrophoresis* **21,** 411–429.
10. Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96,** 6591–6596.
11. Ong, S. -E., Blagoev, B., Kratchmarova, I., et al. (2002) Stable istotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386.

12. Zhu, H., Pan, S., Gu, S., Bradbury, E. M., and Chen, X. (2002) Amino acid residue specific stable isotope labeling for quantiative proteomics. *Rapid Comm. Mass Spectrom.* **16,** 2115–2123.
13. Leitner, A. and Lindner, W. (2004) Current chemical tagging strategies for proteome analysis by mass spectrometry. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **813,** 1–26.
14. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999.
15. Ross, P. L., Huang, Y. N., Marchese, J. N., et al. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3,** 1154–1169.
16. Listgarten, J. and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **4,** 419–434.
17. Ong, S. -E., Kratchmarova, I., and Mann, M. (2003) Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *J. Proteome Res.* **2,** 173–181.

9

Sequence Handling by Sequence Analysis Toolbox v1.0

Christian Ravnsborg Ingrell, Rune Matthiesen, and Ole Nørregaard Jensen

Summary

The fact that mass spectrometry have become a high-throughput method calls for bioinformatic tools for automated sequence handling and prediction. For efficient use of bioinformatic tools, it is important that these tools are integrated or interfaced with each other. The purpose of sequence analysis toolbox v1.0 was to have a general purpose sequence analyzing tool that can import sequences obtained by high-throughput sequencing methods. The program includes algorithms for calculation or prediction of isoelectric point, hydropathicity index, transmembrane segments, and glycosylphosphatidyl inositol-anchored proteins.

Key Words: Isoelectric point; hydropathicity index; transmembrane segments; glycosylphosphatidyl inositol-anchored proteins.

1. Introduction

There are many bioinformatic methods for prediction and calculation based on sequence information available on the internet. However, most of them work through a web interface that is not optimal for high-throughput batch submission. Although web-based programs are, in general, user-friendly and can be reached from almost anywhere, they have a number of drawbacks such as speed limitation, instability of links, the dependency on the internet connection, and the stability of the host server.

Another issue is that many of the methods for sequence analysis are scattered out in many different programs making it a tedious task to make an extensive sequence analysis approach. In sequence analysis toolbox (SAT) v1.0 a number of already published methods for sequence analysis have been included. A number of the methods have been improved compared with the original published

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

methods. The program accepts sequences in batch modes making it compatible with high-throughput methods.

1.1. Isoelectric Point Prediction

The isoelectric point (pI) of a polypeptide is defined as the proton concentration ([H^+]) at which the polypeptide has no net charge. A protein usually has many different ionizable groups that take on different charges at different pH (pH = –log[H^+]).

The Henderson-Hasselbach **Eq. 1** ***(1)*** is used to derive the partial charge of an amino acid with a known pK_a value. The partial charge is defined as the fraction of a molecule with a positive or negative charge at a given pH value ***(2)***:

$$pH = pK_A + \log\frac{[A^-]}{[HA]} \tag{1}$$

Rearranging and isolating the partial charge in **Eq. 1** for the acidic residues gives **Eq. 2** ***(3)***

$$faC_{aa}(pH) = \frac{10^{pH-pK_{aa}}}{10^{pH-pK_{aa}}+1} \tag{2}$$

Obtaining a simpler expression is done by multiplying **Eq. 2** by the factor 10^{pKaa}

$$faC_{aa}(pH) = \frac{10^{pH}}{10^{pH}+10^{pK_{aa}}} \tag{3}$$

And for basic residues the partial charge becomes ***(3)***

$$fbC_{aa}(pH) = \frac{1}{10^{pH-pK_{aa}}+1} \tag{4}$$

The total partial positive charge (c^+) in a polypeptide is given by the sum of fractions of positive-charged amino acid:

$$c^+(pH) = \sum_{aa \in pAA} n_{aa} \cdot faC_{aa}(pH) \tag{5}$$

A similar expression for the total fraction of negative charges (c^-) can be deduced:

$$c^-(pH) = \sum_{aa \in nAA} n_{aa} \cdot fbC_{aa}(pH) \tag{6}$$

In **Eqs. 5** and **6**, n_{aa} denotes the number of the amino acid species, *aa*, in the polypeptide. *pAA* denotes the set of positive-charged amino acids: *pAA* = (lysine,

arginine, histidine, N-terminal residue), and *nAA* the set of negative-charged amino acids. nAA = (asparatate, glutamate, cysteine, tyrosine, phosphor serine/ threonine, C-terminal residue), because there is only one N- and C-terminal in a polypeptide, $n_{Nter} = n_{Cter} = 1$.

Subtracting **Eq. 6** from **Eq. 5** yields **Eq. 7**, which describes the total partial charge of a polypeptide:

$$g(pH) = c^{+}(pH) - c^{-}(pH) \tag{7}$$

Setting **Eq. 7** to zero and solving for the pH, yields the pH, at which the net charge is zero—i.e., the pI.

Analytically this equation is hard to solve ***(3)***, but different numerical methods can be applied to solve it. In Weiller et al. ***(3)***, they scan through the pH scale with a step size of 0.1 to find a zero solution.

Another approach is to use the bisection method ***(4)***, which half the searched interval for each iteration. This can be more formally stated as:

Given a function *f* and two points x_k and y_k that satisfy the constrain, given in **Eq. 8**:

$$f(x_k) \cdot f(y_k) < 0 \tag{8}$$

Let X_{k+1} be the midpoint between x_k and y_k as written in **Eq. 9**

$$x_{k+1} = \frac{1}{2}(x_k + y_k) \tag{9}$$

Choose $x_{k+1} = x_k$ or $y_{k+1} = y_k$, so that the condition in **Eq. 10** is satisfied.

$$f(x_{k+1}) \cdot f(y_{k+1}) < 0 \tag{10}$$

Continue this approach with finding a new midpoint, until the error ($e_k = |x_k - y_k|$) becomes satisfactory.

The error in iteration k (e_k) of the bi-section method can be estimated by **Eq. 11**.

$$|e_k| \le 2^{-k} |x_0 - y_0| \tag{11}$$

An interesting feature about a numerical method is to know how many iterations it requires to reach a given precision. Rearranging and solving for *k* in **Eq. 11** yields **Eq. 12** that can be used to estimate the number of iteration *k* needed to reach a precision within $|e_k|$.

$$k \le \frac{\lg\left(\frac{|e_k|}{|x_0 - y_0|}\right)}{\lg 2} \tag{12}$$

To obtain a result within the error range of ±0.1 ($e_k = 0.1$) in the pH-scale 0–14.0 ($x_o = 0$, $y_o = 14$), eight iterations are needed according to **Eq. 12**.

The bisection method is easily applied to **Eq. 7** because the problem of finding two points satisfying **Eq. 10** is practically nonexistent. At a pH close to zero the polypeptide will always be positively charged and likewise at a pH close to 14.0 it will have a net negative charge.

The linear approach applied in ***(3)*** yields a time-complexity of $O(n)$, whereas the bisection method gives a time-complexity of $O(\log_2 n)$. For large n (a large dataset) the difference between those two approaches become significant. The big-O notation was adapted from **ref.** *5*.

The model that the pI prediction is based on has some limitation. First of all, this model was deduced to predict pI in a denaturizing environment ***(6)***. In this environment the polypeptide are approximated as linear with no secondary structures and, thereby, no intramolecular interactions between charged groups. Second, the model depends (significantly) on which pK values are used. In this work the pK values are adapted from **ref.** *7*.

The N- and C-terminal of the polypeptide are amine and carboxylic acid groups respectably, with pK values depending on the residue. Thus, for the model to hold the N- and C-terminal has to be unmodified (i.e., no acetylation, methylation, or amide formation). This constraint also applies to the rest of the polypeptide, which means for the model to hold, no posttranslational modification must be present. However, the implemented algorithm here releases this constrain by allowing phosphorylation on serine, threonine, and tyrosine.

1.2. Prediction of Glycosylphosphatidylinositol-Anchored Proteins

Glycosylphosphatidylinositol-anchored proteins (GPI-AP) belong to a subset of membrane-associated proteins without internal transmembrane spanning regions (TSRs). GPI-APs are found in all eukaryotic organisms and exhibit functions as receptors, adhesion molecules, ectoenzymes, differentiation antigens, and adaptors ***(8)***.

After translation of a protein targeted for GPI anchoring, a preformed GPI glycolipid is attached to the target protein in a transamidation process that leads to the release of the C-terminal peptide from the target protein.

The GPI glycolipid is synthesized on the luminal side of the endoplasmatic reticulum, which causes the GPI-AP to face the exterior of the plasma membrane ***(9)***. Proteins targeted for GPI modification show some common traits that can be utilized in the identification of these based on predictions from their primary structure:

1. N-terminal signal sequence that addresses the protein to the endoplasmatic reticulum.
2. Hydrophobic C-terminal region.

3. Fairly conserved ω-site (the C-terminal cleavage site) sequence motif.
4. No internal TSR.

Because GPI-APs are a wide-spread phenomenon in the eukaryotic kingdom, and display a set of constrained properties, they possess an interest in a bioinformatic view.

The algorithm used for SAT v1.0 for the prediction of GPI-AP is adapted from detection of GPI (DGPI) ***(10)***. The algorithm uses the six properties listed next ***(11)*** to determine whether a given protein is attached to the membrane by a GPI anchor. If this is the case, then it determines the C-terminal cleavage site (ω-site). The DGPI algorithm is build on five rules:

1. The protein must contain an N-terminal secretion signal. The probability of the presence of a secretion signal is predicted using a method described in **ref.** ***12***. This method is based on the formation of weighted matrix calculated from an annotated set of secretory signal sequences.
2. The protein's C-terminal must contain a hydrophobic region of a minimum length of 13 amino acids. In order to determine the C-terminal hydrophobic region of the protein, the kd-index together with the sliding window principle is used. From an annotated GPI-AP dataset (Swiss-Prot) it was determined that a window-size of 15 was appropriate for predicting the presence of a hydrophobic region in the GPI-anchored proteins ***(10)***.
3. Between the hydrophobic C-terminal and the ω-site there has to be a hydrophilic region with a minimal length of three amino acids ***(10)***. This region is determined with the same parameters as described in rule 3 (i.e., kd-index, window size 15).
4. According to **ref.** ***10*** the ω-site can be found seven amino acids before the C-terminal hydrophobic region.
5. This rule is an improvement in the prediction of the ω-site from rule 4. From a Swiss-Prot dataset consisting of 172 annotated GPI-AP with a known ω-site, the amino acids distribution neighbouring (ω+1 and ω+2) the ω-site was calculated. This allows one to verify rule 4 and determine the most potential ω-site based on the above calculated distribution.

However, the DGPI algorithm does not make use of the fact that GPI-AP does not contain any internal TSR ***(11)***. So in this project the GPI predictor makes use of this fact and thereby adds an extra rule that states:

6. No internal TSR in the proteins. TSRs are predicted using THMHMM 2.0 ***(13)*** or the local TSR predictor using the kd-index with the sliding window principle.

1.3. Testing the GPI-AP Prediction

The objective with this section is to give an overview of the quality of using the implemented GPI-AP prediction tool, DGPI. DGPI has been tested on a comprehensive dataset ***(10)***. The dataset consists of 20,000 proteins from the

Swiss-Prot database, where 349 were annotated to be GPI-AP. The results from this test are as follows:

1. 977 proteins were predicted with DGPI to be GPI-AP.
2. 270 of the 977 proteins where predicted correctly, meaning that the 270 proteins were annotated GPI-AP in the Swiss-Prot database.
3. 349 – 270 = 79 GPI-AP were not detected with DGPI.

In other terms this yields 270 true-positives (*TP*), 20,000 – 977 = 19,023 true negatives (*TN*), 79 false-negatives (*FN*), and 977 – 270 = 707 false-positive (*FP*). Taking the size of the dataset into account, an informative error analysis can be carried out.

From the previously listed four values a rough estimate shows that DGPI (rule 1–5; *see* **Subheading 1.2.**) captures about three out of four GPI-AP, but on the other hand, only predicts a little fraction as being GPI anchored when they were not.

A more formal way of measuring the effectiveness of a prediction method is to divide its accuracy into specificity and sensitivity. Specificity is defined as ***(14)***:

$$specificity = \frac{TN}{TN + FP} \tag{13}$$

Applying **Eq. 13** to the DGPI results yield a specificity of 96.4%. Consequently, the DGPI-prediction method is highly specific, i.e., relatively rarely overpredicts.

Sensitivity is measured using ***(14)***:

$$sensitivity = \frac{TP}{TP + FN} \tag{14}$$

Sensitivity is the term used for a prediction method to measure how well the method captures the *TP*. Applying **Eq. 14** to the DGPI results yields a sensitivity of 77.4%. This implies that the DGPI method does not capture all real GPI-AP.

These results were obtained without using rule six in **Subheading 1.2.** In SAT v1.0, rule 6 has been added to the GPI-AP-prediction algorithm. The accuracy of the transmembrane predictor used for rule 6 is also of importance. SAT v1.0 uses THMHMM 2.0 ***(13)*** for prediction of TSR. The THMHMM 2.0 prediction method has been evaluated and showed an approximate specificity of 80% in general ***(15)***. Approving a specificity of 80% makes it highly unlikely that the THMHMM prediction method will predict nine TSRs as *FP* (the likelihood would be: $(1 - 0/80)^9 = 5.12 \cdot 10^{-7}$).

1.4. Membrane Proteins

Membrane proteins play an important role in the cell in areas including cell-to-cell signaling and the transmembrane transport of ions and solutes. Also, they have achieved interest in the pharmaceutical industry owing to their success as therapeutic targets for medicine *(16)*. Transmembrane proteins do not differ significantly in various organisms *(15,17)*, and it is estimated that 20–30% of all genes in most genomes encode membrane proteins *(13)*. This implies that membrane proteins are widely abundant in most organisms, and therefore, also should be given attention in order to understand their function. However, membrane proteins compromise a major challenge for protein chemistry because of their insolubility in aqueous solution and difficulties in crystallizing for structural analysis *(16)*.

Several approached for predicting TSRs from a naked protein sequence have been proposed *(13,16)*. In SAT v1.0 a simple but strong TSR predictor has been implemented.

2. Material

2.1. Analysis of Sequence Data

1. SAT v1.0 (requirement, http://yass.sdu.dk/SAT/SATv1.0).
2. Java 2 platform (requirement; http://java.sun.com/j2se/1.4.2/download.html).
3. VEMS v3.0 has an interface to SAT v1.0 that makes it possible to export peptide and protein sequences obtained by tandem mass spectrometry searches in VEMS to SAT v1.0 (optional; http://yass.sdu.dk).

2.2. Software Development

1. Eclipse is provided by the Eclipse Foundation. The Eclipse Foundation is a nonprofit corporation formed to advance the creation, evolution, promotion, and support of the Eclipse Platform and to cultivate both an open source community and an ecosystem of complementary products, capabilities, and services (http://www.eclipse.org).
2. Java 2 platform (requirement; http://java.sun.com/j2se/1.4.2/download.html).

SAT v1.0 was developed on the Java 2 platform, standard edition software development kit v1.4.2_04 (J2SE SDK) by Sun Microsystems (Santa Clara, CA) using the eclipse editor. The Java language is a platform-independent object-oriented programming language. Java is compiled to an intermediate byte code that is interpreted on the fly by the Java interpreter. In Java one can chose to make applets or applications. SAT v1.0 is programmed as a java application because this gives more programming freedom owing to security risks with applets.

SAT v1.0 is modular designed in three layers (implemented in java as packages) as illustrated in **Fig. 1**. The user interface layer contains the code, handling

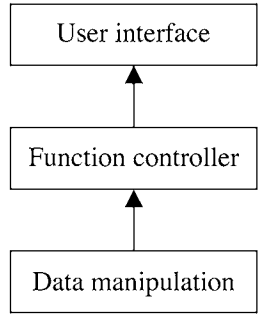


Fig. 1. The overall structure of sequence analysis toolbox (SAT) v1.0. SAT v1.0 is divided into three minor substructures each with a well-defined area of responsibility.

user interactions with the program. The responsibility of the function controller layer is to control the connection between the user interface and the data manipulation layer, which holds responsibility for programming code doing calculation, retrieval of sequences, prediction, and so on. This modular structure enhances the flexibility of the program, which mean that new functionality easily can by implemented.

3. Methods

3.1. General Overview of SAT v1.0

There are two types of tasks that SAT v1.0 handles. The first one is protein sequence retrieval tasks.

1. Retrieve a set of protein sequences from their respective accession numbers.
   a. Accession numbers can be retrieved from a flat file (a file where each line contains an accession number).
   b. From the list of these accession numbers the respective protein sequences can be retrieved from a FASTA file, Swiss-Prot file, or the ExPASy www-service.
   c. A proteomic experiment containing peptide and protein sequences can be imported from VEMS v3.0 ***(18)*** (*see* Chapter 7).
2. Perform sequence predictions/calculations like:
   a. pI/mass.
   b. Hydropathicity index.
   c. TSR/Import prediction performed by THMHMM ***(13)***.
   d. GPI-AP predictions.
   e. Secretion signal peptide prediction.
   f. Secondary structure prediction.
   g. Amino acid composition.

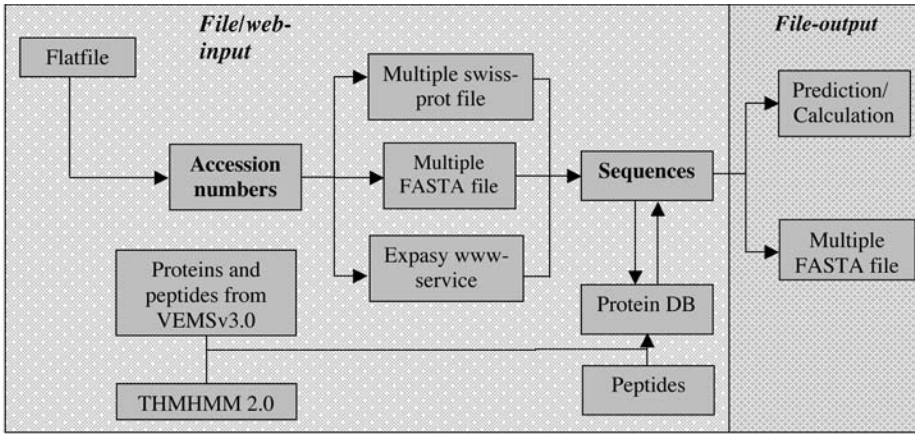


Fig. 2. The information flow for sequence handling in sequence analysis toolbox v1.0. Accession numbers can be retrieved from a flat file and protein sequences are retrieved from multiple Swiss-Prot files, multiple FASTA files, or the ExPASy www-service. Different predictions/calculations can be applied or the retrieved sequence can be written to a file. Another approach is to save the obtained sequences to the protein database (DB) and if available also save peptides to their respective proteins. Last, proteins identified by the mass spectrometry interpretation program Virtual Expert Mass Spectrometrist v3.0 ***(38)*** can be imported. Prediction of transmembrane spanning regions by THMHMM 2.0 can be imported to the protein DB.

These tasks can be undertaken by using the main frame (window) of SAT v1.0. The information flow for sequence handling in SAT v1.0 is illustrated **Fig. 2**.

SATv1.0s main window of its graphical user interface (GUI) is shown in **Fig. 3**. In the following text, the numbers mentioned will refer to numbers in **Fig. 3**. When specifying how to extract information for further processing you use the combo box 1, 2, and 7. Afterwards, you specify where the information is to be retrieved from (which files) in the fields 4 and 5—this is done by using the file-choosers that appear when pressing the buttons left of the fields (4 and 5). With combo box 3, you choose which property you want to examine, and set possible parameters to it, with the spinners at number 8. The elliptic area mark number 6 is used to specify the output file, in which the result from the prediction/calculation are written to (results can be further processed by, e.g., Microsoft Excel). Another approach, when using SAT v1.0, is to add the sequences to the internal data structure "protein DB" and use the operators in combo box 7 to extract different information about the stored sequences. The dominating white-space that fills the bottom half of **Fig. 3** is used to display whether a given task has been completed correctly or not.

The second type of task is the administration of multiple sequences for which SAT v1.0 has implemented a data structure (protein DB) to hold protein

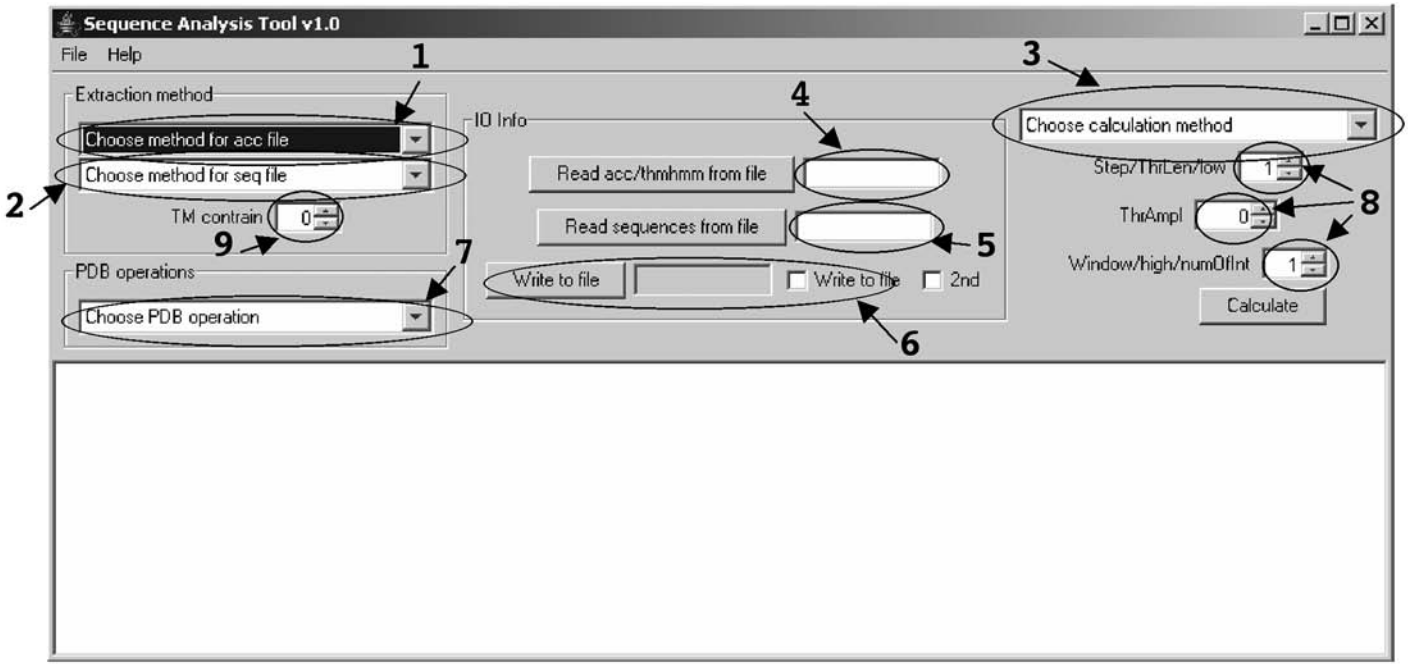


Fig. 3. The main window of the SAT v1.0 application. (1) In this combo box you choose with which method you want to extract the accession numbers from the file defined in field 4. In combo box 2 you choose from what and how you want to extract sequences (peptide or protein) from the file defined in field 5. The file specifications of field 4 and 5 can be chosen by pressing the buttons left to the fields and using the file chooser that appears. Combo box 3 is used to choose which prediction/calculation (e.g., pI prediction, transmembrane spanning region [TSR] prediction, etc.) you want to apply on the obtained sequences specified from 1, 2, 4, and 5. Field 6 is used to define if and where you want to save the results from the prediction/calculation applied. The spinners, 8, are used to set prediction/calculation parameters (e.g., window size for kd-index used in TSR prediction). Combo box 7 is used to operate the protein DB. The spinner, 9, is used to constrain sequence retrieval from the protein DB with a specified number of how many predicted TSRs a protein must have.

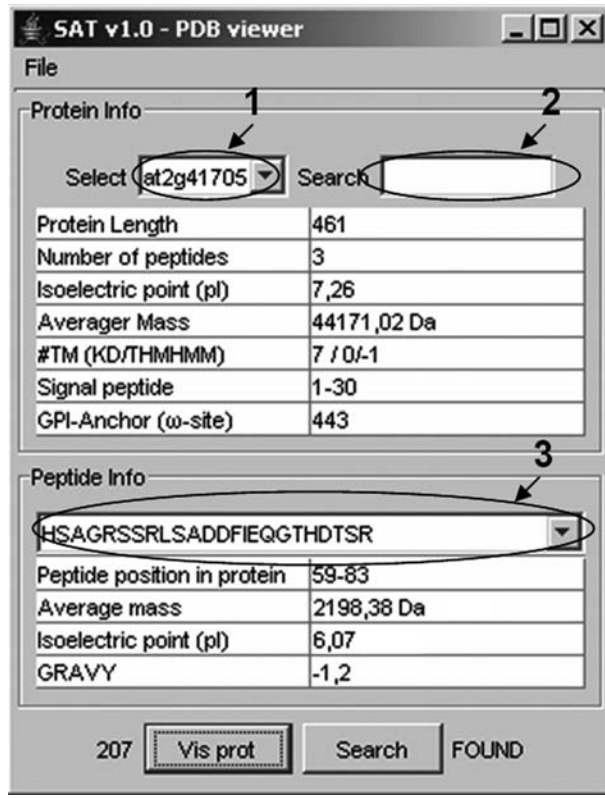


Fig. 4. The Protein database (PDB) viewer lists the physicochemical properties of a chosen protein and its belonging peptides in the PDB. In combo box 1 you can select between proteins in PDB from their accession number (or search them in field [2]). In combo box 3 you can chose between peptides respective to the selected protein in combo box 1.

and peptide information obtained from a peptide mass fingerprint-based proteomic experiment. This allows one to browse amongst the identified proteins and their respective peptides with the aid of SAT v1.0's protein DB viewer component (*see* **Fig. 4**). In the PDB viewer, prediction and calculation of physicochemical properties are shown like in **Fig. 4**. Together with the protein DB-viewer component there is a visualization tool (Protein Vis) that shows the following properties (**Fig. 5**): hydropathicity plot using the kd-index, peptides position in the protein, predicted TSR segments, secondary structure prediction (*see* **Appendix D**), predicted secretion signal peptide, and predicted GPI-anchoring site.

To open the PDB viewer click on the file menu from the main screen and choose "PDB-viewer" from the list. In order to visualize the protein with

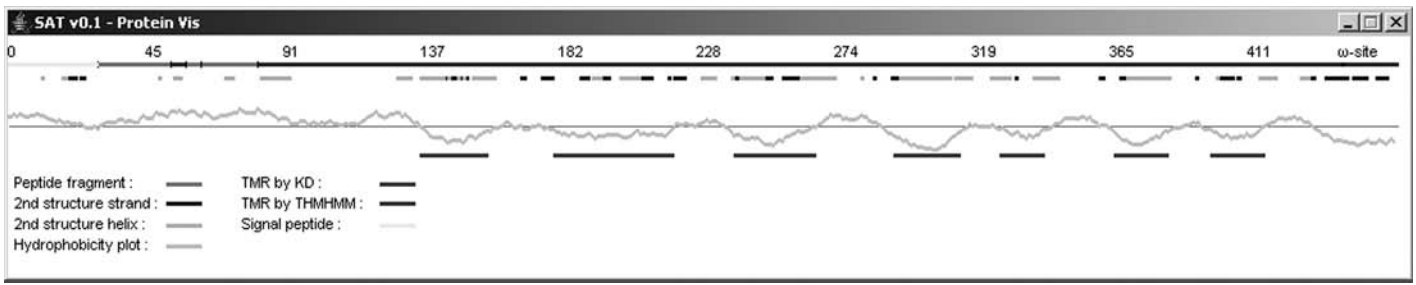


Fig. 5. The “Protein Vis” component is a tool to visualize physicochemical properties and predictions of a protein. The protein shown here is an unknown protein with TGIR gene *ID At2g41705* from *Arabidopsis thaliana*.

Fig. 6. A flat file opened with notepad (Microsoft Windows XP). Each line in the file represents a peptide sequence.

"Protein Vis" click on the "Vis prot" button in the PDB viewer. External prediction made by THMHMM 2.0 ***(13)*** can also be added to the protein DB.

3.2. Practical Guidelines for Predicting/Calculating Physicochemical Properties

3.2.1. Isoelectric Point, GRAVY, and Molecular Weight

1. First step includes importing sequences from the desired source. The source could be a search result from VEMS v3.0 (*see* Chapter 7), multiple FASTA file, Swiss-Prot file, or simply a list of peptides in a text file as shown in **Fig. 6**.
   a. To open a text file in SATv1.0 you press the button "Read sequences from file" (*see* **Fig. 3**). Now a file-chooser box will appear. If the selected file is in a flat-file format (*see* **Fig. 6**) you set combo box nr. 2 to "all seq from flat file."
   b. It is also possible to open files in other formats like Swiss-Prot or FASTA. Combo box 1 (*see* **Fig. 3**) can be used to constrain which sequences are extracted from the sequence file specified in field nr. 4 (*see* **Fig. 3**). If combo box 1 is set to "acc. from flat file" then only sequences with accession numbers present in the file specified in field nr. 5 (*see* **Fig. 3**) will be imported for further processing (for this option to be used the sequence files have to be in multiple FASTA or Swiss-Prot format, which is specified by combo box 2 with the options "acc → seq in multiple FASTA file" and "acc → seq in multiple Swiss-Prot file").
   c. Another option for retrieving sequences is to have a flat-file containing Swiss-Prot accession numbers without having the sequences in hand. This file is selected with the file chooser that opens by clicking the button named "Read

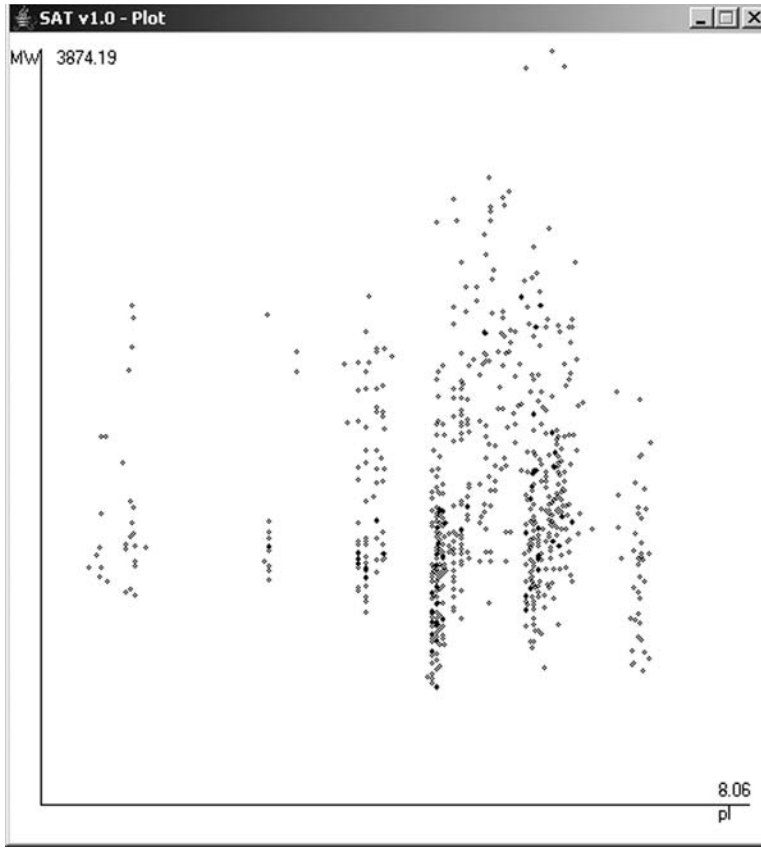


Fig. 7. Predicted isoelectric point vs calculated molecular weight of a peptide dataset obtained in a phosphoproteomic study ***(19)***. The plot window gives a firsthand impression of data.

acc/thmhmm from file." In order to specify that the sequences are to be requested from the Swiss-Prot database combo box 2 must be set to "acc → seq from expasy."

d. Amino acid sequences can also be imported from search results obtained by the VEMS v3.0 program. This is done in two steps. First, the sequences have to be imported to the protein DB, which is done by setting combo box 7 (*see* **Fig. 3**) to "Import sequences from VEMSv3.0" and clicking the calculate button (*see* **Fig. 3**). Now the protein and peptide sequences can be viewed with the PDB viewer by opening the "file menu" and choosing "pdb manager." Then a window appears like in **Fig. 4**, where in order to activate the imported sequences you select the "file menu" and click on "new." To use the sequences in the protein DB for calculation/prediction, set combo box 2 to "all seq from proteinDB."

2. Second step is to choose the prediction/calculation method that should be performed, which is done by combo box 3 (*see* **Fig. 3**). For prediction/calculation of pI, molecular weight, and GRAVY (grand average of hydropathicity) combo box 3 should be set to "pI, mass and GRAVY."
3. Third step includes specifying where to write the results of the calculation/prediction. To specify the location of the result file press button nr. 6 (*see* **Fig. 3**) and a file chooser will be opened. The result file is formatted in a comma separated way, which means that each line represents prediction/calculations for one amino acid sequence and the three physicochemical properties are separated with a comma.
4. The last step is to perform the calculation/prediction, which is done be pressing the calculate button (*see* **Fig. 3**). After the processing of the data a result window will be opened (*see* **Fig. 7**) where pI vs mass will be plotted. If further analysis is necessary the result file can be used.

Acknowledgments

R. M. was supported by grants from EU TEMBLOR and by Carlsberg Foundation Fellowships. O. N. J. is a Lundbeck Foundation Research Professor and the recipient of a Young Investigator Award from the Danish Natural Science Research Council.

References

1. Zumdahl, S. S. and Zumdahl, S. A. (2000) Application of aqueous equilibra, *Chemistry Zumdahl, 5th ed.*, (Zumdahl, S. S. and Zumdahl, S. A., eds.), Houghton Mifflin Company, Boston, MA, pp. 735.
2. Pawson, T. and Scott, J. D. (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278,** 2075–2080.
3. Weiller, G. F., Caraux, G., and Sylvester, N. (2004) The modal distribution of protein isoelectric points reflects amino acid properties rather than sequence evolution. *Proteomics* **4,** 943–949.
4. Christiansen, E. (2003) Numerisk analyse. *Institut for Matematik og Datalogi*, University of Southern Denmark, Odense M, Denkark 1–5.
5. Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001) *Introduction to Algorithms, 2nd ed.* The MIT Press, Cambridge, MA.
6. Bjellqvist, B., Hughes, G. J., Pasquali, C., et al. (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequence. *Electrophoresis* **14,** 1023–1031.
7. Dawson, R. M. C., Elliot, D. C., Elliot, W. H., and Jones, K. M. (eds.) (1986) *Data for Biochemical Research,3rd ed.*, Oxford Science Publications, Oxford, UK, pp. 1–31.
8. Elortza, F., Nühse, T. S., Foster, L. J., Stensballe, A., Peck, S. C., and Jensen, O. N. (2003) Proteomic analysis of glycosylphosphatidylinositol-anchored membrane proteins. *Mol. Cell. Proteomics* **2,** 1261–1270.

9. Voet, D. and Voet, J. G. (1995) Other Pathways of carbohydrate metabolism, *Biochemistry, 2nd ed.*, (Voet, D. and Voet, J. B., eds.), Wiley, Boston, MA pp. 617.
10. Kronegg, J. and Buloz, D. (1999) Detection/prediction of GPI cleavage site (GPI-anchor) in a protein (DGPI). http://129.194.185.165/dgpi/. Last accessed Oct 2006.
11. Borner, G. H. H., Sherrier, D. J., Stevens, T. J., Arkin, I. T., and Dupree, P. (2002) Prediction of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A genomic analysis. *Plant Physiol.* **2,** 486–499.
12. von Heijne, G. (1986) A new method for predicting signal sequence cleavage sites. *Nucl. Acids Res.* **14,** 4683–4690.
13. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580.
14. Brazma, A., Jonassen, I., Ei-dhammer, I., and Gilbert, D. (1998) Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.* **5**, 279–305.
15. Wallin, E. and von Heijne, G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7,** 1029–1038.
16. Möller, S., Croning, M. D. R., and Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17,** 646–653.
17. Stevens, T. J. and Arkin, I. T. (2000) Do more complex organisms have a greater proportion of membrane proteins in their genomes. *Proteins* **39,** 417–420.
18. Matthiesen, R., Bunkenborg, J., Stensballe, A., Jensen, O. N., Karen, G., W., and Bauw, G. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* **4,** 2583–2593.
19. Nuhse, T. S., Stensballe, A., Jensen, O. N., and Peck, S. C. (2003) Large-scale analysis of in vivo phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. *Mol. Cell Proteomics* **2,** 1234–1243.

10

Interpretation of Collision-Induced Fragmentation Tandem Mass Spectra of Posttranslationally Modified Peptides

Jakob Bunkenborg and Rune Matthiesen

Summary

Tandem collision-induced dissociation (CID) mass spectrometry (MS) provides a sensitive means of analyzing the amino acid sequence of peptides. Modern MS instrumentation is capable of rapidly generating many thousands of tandem mass spectra, and protein database search engines have been developed to cope with this avalanche of data. In most studies, there is a schism between discarding perfectly valid data and including nonsensical peptide identifications—this is currently a major bottleneck in data analysis and it calls for manual evaluation of the data. Especially for posttranslationally modified peptides, there is a need for manual validation of the data because search algorithms seldom have been optimized for the identification of modified peptides and because there are many pitfalls for the unwary. This chapter describes some of the issues that should be considered when interpreting and validating low-energy CID tandem mass spectra and gives some useful tables to aid this process.

Key Words: Proteomics; posttranslational modifications; mass spectrometry; database searching.

1. Introduction

Proteins can be viewed as linear biopolymers composed of the only 20 different genetically encoded amino acids, but the functionally active form of a protein is often quite different from that of the linear nascent polypeptide chain. Posttranslational modifications (PTMs) are covalent alterations of the polypeptide chain that change the structure. A bewildering number of changes can occur: intramolecular bonds can be formed by cystine disulfide bridges, sequences can be removed from the polypeptide chain by enzymatic cleavage, and most amino acids can be modified. Many cellular functions are regulated

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

by reversible phosphorylation, acetylation, glycosylation, or other enzymatically catalyzed modifications of proteins. In addition to the biologically significant PTMs it is a well-known and most often ignored fact that proteins are modified during storage and sample handling. When trying to analyze proteins it is very important to be aware of the mundane posttranslational modifications, such as deamidation, oxidation, backbone cleavage, and other common spontaneous modifications, as well as those that occur in consequence of the sample handling, such as alkylation with acrylamide from sodium dodecyl sulfate-polyacrylamide gel electrophoresis separated proteins or carbamoylation through the use of urea as a denaturant.

Mass spectrometry (MS) measures the mass-to-charge ratio (*m/z*) of charged ions and has become a major tool for protein identification. Soft ionization techniques, like electrospray ionization (ESI) and matrix-assisted laser desorption/ ionization (*see* Chapter 1), allow large biomolecules to be transferred to the gas phase and ionized. The direct analysis of proteins is most often not feasible and for identification purposes proteins are typically digested to peptides by a specific protease (e.g., trypsin that cleaves after arginine and lysine residues) and the masses of each peptide are determined. Sequence information for each peptide can then be gained by collision-induced dissociation (CID) tandem mass spectrometry (MS/MS), where a peptide ion is selected and collided with an inert gas. The *m/z* of the resulting fragment ions are then measured. The nomenclature for the peptide fragmentation ***(1–3)*** is illustrated in **Fig 1**. This chapter focuses on low-energy CID (less than 100 eV) of protonated peptides in ion traps, triple quadrupoles, or QqTOFs where there is very little side-chain fragmentation. Normally the low-energy, collision-induced cleavage of the peptide backbone occurs at the amide bond giving rise to y- and b-type ions that contain the C- and N-terminal part of the peptide, respectively. *See* **ref.** ***4*** for a recent review. Additional ions can be generated by the loss of small neutral molecules like water (a –18 Da ion series usually denoted with a superscripted o—e.g., y_i^o) or ammonia (a –17 Da ion series usually denoted with a superscripted asterisk * —e.g., y_i^*). Often, it is also possible to find immonium ions from the individual amino acids in the low *m/z*. CID fragmentation of multiply charged ions can also give rise to multiply charged fragment ions and it is not uncommon for multiply charged ions to lose one or several N-terminal amino acid residues as neutral fragments ***(5)***.

MS instrumentation is rapidly improving and diversifying but there are still different limitations in mass range, mass accuracy, and fragmentation efficiency for each type of instrument, and it is necessary to consider these limitations when analyzing the data. For example, the mass accuracy needs to be better than 1 Da when trying to detect deamidation and better than 0.0330 Da when trying to distinguish phenylalanine from oxidized methionine by mass. Other techniques for inducing peptide fragmentation like electron capture dissociation

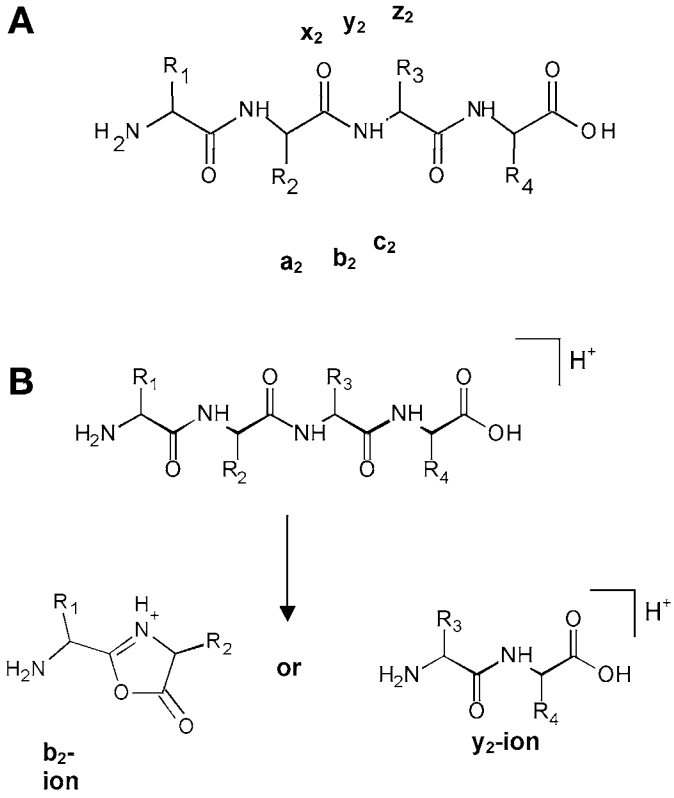


Fig. 1. **(A)** Peptide fragment ion nomenclature. The indices of the amino terminal-containing a-, b-, and c-ions denote the number of residues counted from the N-terminus and the indices of the carboxy terminal x-, y-, and z-ions are counted from the C-terminus. **(B)** Collision-induced dissociation fragmentation of the peptide amide bonds to produce N-terminal b-ions and C-terminal y-ions. Linear b-ions are unstable and cyclize to form an oxazolone structure involving the carbonyl group of the adjacent residue. The b_1-ion is rarely observed because it cannot form the stabilizing oxazolone structure but it does occur if the N-terminal is derivatized with a carbonyl-containing group (e.g., acetylated). The b_1-ion can be seen for arginine, lysine, histidine, and methionine that can form alternative stabilizing structures by means of their side chains ***(39)***.

(6,7) and electron transfer dissociation ***(8)*** are not commonplace yet, but are very powerful tools for analyzing especially labile PTMs as glycosylation ***(9–11)***, phosphorylation ***(8,12)***, and γ-carboxyglutamic acid ***(13)***. A large number of programs are available that can identify the peptide by comparing the experimental MS/MS spectrum with the theoretical one calculated for each peptide in a protein database. The mass changes associated with PTMs make MS ideally suited for the analysis. Modifications that occur stoichiometrically

Table 1
Posttranslational Modifications[a]

Modification	Modified amino acids	Monomer composition	Mono-isotopic mass	Diagnostic ions	Neutral loss
N-terminal acetylation	N-term	+ C2H3O	+ 43.01839		
N-terminal carbamido-methylation	N-term	+ C2H4NO	+ 58.02929		
N-terminal carbamo-ylation	N-term	+ CH2NO	+ 44.01364		
N-terminal Pyro-glutamine	Q (N-term)	–NH2	–16.01872		
N-terminal Pyro-glutamic acid	E (N-term)	–HO	–17.00274		
N-terminal Pyro-carbamido-methylcysteine	C (N-term)	–NH2	–16.01872		
Propionamide	C	C6H10N2O2S	174.04630	147.059	
Oxidation (Methionine sulfoxide)	M	C5H9NO2S	147.03540	120.048	63.998
Oxidation (hydroxy-tryptophan)	W	C11H10N2O2	202.07423	175.087	
Oxidation of carbamido-methylated C	C*	C5H8N2O3S	176.02556		107.004
Oxidation (2-Oxohistidine)	H	C6H7N3O2	153.05383	126.066	
Dioxidation (Methionine sulphone)	M	C5H9NO3S	163.03031	136.043	
Dioxidation (*N*-formylkynurenine)	W	C11H10N2O3	218.06914	191.082	
Kynurenine	W	C10H10N2O2	190.07423	163.087	
Deamidation	N	C4H5NO3 (~D)	115.02694		
Deamidation	Q	C5H7NO3 (~E)	129.04259		

(Continued)

Table 1 ***(Continued)***

Phosphorylation	S	C3H6NO5P	166.99836		97.977
					79.966
Phosphorylation	T	C4H8NO5P	181.01401		97.977
					79.966
Phosphorylation	Y	C9H10NO5P	243.02966	216.042	79.966
Methylation	K	C7H14N2O	142.11061	115.123	
				98.096	
				84.081	
Methylation	R	C7H14N4O	170.11676	143.129	73.064
				115.087	56.0374
				112.087	31.0422
				74.071	
				70.065	
Dimethylation	K	C8H16N2O	156.12626	129.137	
				84.081	
Dimethylation (asymmetric)	R	C8H16N4O	184.13241	157.149	87.0796
				115.087	70.0531
				112.087	45.0579
				88.087	
				71.060	
Dimethylation (symmetric)	R	C8H16N4O	184.13241	157.149	87.0796
				115.087	70.0531
				112.087	31.0422
				88.087	
				71.060	
Trimethylation	K	C9H18N2O	170.14191	143.154	59.073
				84.081	
Trimethylation	R	C9H18N4O	198.14806	171.160	
Acetylation	K	C8H14N2O2	170.10553	143.118	
				126.091	
				84.081	
γ-Carboxylation	E	C6H7NO5	173.0324		43.990
Nitration	Y	C9H8N2O4	208.04841	181.061	
Hydroxyproline	P	C5H7NO2	113.0477	86.060	

[a]This short list contains some of the modifications we look for depending on the biological problem (e.g., histones or serum proteins) and the sample preparation method (e.g., sodium dodecyl sulfate-polyacrylamide gel electrophoresis-separated proteins or an in-solution digestion of urea-denatured proteins). The N-terminal modifications are added as the N-terminal term m_N in the equations and the composite masses as the residue term m_i. Often an iterative search strategy is used, where posttranslational modifications of the identified proteins are looked up in annotated protein databases (e.g., Swiss-Prot or HPRD—*see* **Subheading 2.2.4.**) and included in a second iteration.

are often denoted fixed modifications (because all residues of a given type are modified)—as an example the derivatization of cysteines with iodoacetamide occurs with almost 100% efficiency on all cysteine residues. Modifications that occur substoichiometrically (only a few residues are modified) are called variable modifications—oxidation of methionine is rarely quantitative unless an oxidizing agent is utilized (i.e., all the residues do not oxidize completely) and methionine sulfoxide is most often included in the analysis as a variable modification of methionine residues.

It is often a difficult problem to assess if the peptide retrieved by the search engine really is the correct sequence because the MS/MS data is often far from perfect. Two major limitations apart from data quality usually apply for the identification of PTMs using different search algorithms. Virtually all search engines produce a best-fit solution to a user-defined problem, but it is not always possible to get the correct solution either because the database protein sequence is a variant (or wrong) or because the peptide is modified in an unforeseen way not accounted for by the search parameters. This problem is even worse for many modified peptides where very labile groups can lead to less informative fragmentation patterns from which limited sequence information can be gained. From the mass of the peptide ion it is possible with high mass accuracy to distinguish between PTMs that are nearly isobaric. Fortunately tandem mass spectra give additional analytical handles on modifications where characteristic neutral losses, composite mass increments in the peptide sequence ions, and diagnostic ions in the low mass region can lead to the correct interpretation and identification of the peptide. It is beyond the scope to this chapter to present a comprehensive list of all PTMs, but a selection of the most common and well studied is listed in **Table 1**. The manual interpretation and validation of tandem mass spectra of posttranslationally modified peptides can often be aided by comparing to the MS/MS of the nonmodified sequence, a synthetic version of the modified peptide, or a chemically modified version of the same PTM peptide (e.g., acetylating the peptide).

Modern MS instrumentation directly coupled to liquid chromatography is capable of generating overwhelming amounts of data. Validating the unmodified peptide identifications resulting from database searches of this data is a very time consuming and at times difficult task, but it is currently necessary unless very stringent identification thresholds are imposed. It is not always true what the search engine retrieves and there are many less fortunate examples where the assignments have not been adequately checked (*see* Johnson et al. ***[14]*** for a discussion). This chapter aims at giving some helpful tools and rules for interpreting and validating tandem mass spectra of peptides in general and of modified peptides in particular.

Table 2
List of Accurate Elemental Mass Values for the Most Commonly Occurring Elements (Including Stable Isotope Labeling) in Peptides and Posttranslationally Modified Peptides

H-1	1.007825035
H-2	2.014101787
C-12	12.000000000
C-13	13.003354826
N-14	14.003074002
N-15	15.000108970
O-16	15.994914630
O-18	17.999160300
P-31	30.973762000
S-32	31.972070700
Electron	0.000548580
Proton	1.007276455

2. Materials

2.1. Useful Tables

Table 2 has the exact masses of the most common elements in posttranslational protein modifications—the composite monoisotopic mass of the modified residue can then be calculated once the chemical composition is known if the modification cannot be found among the PTMs in **Table 1**.

2.2. Web Resources

A collection of links to software resources to help the process of peptide identification by CID MS/MS.

2.2.1. Protein Database Search Engines

1. Mascot: http://www.matrix-science.com/.
2. X-tandem: http://www.thegpm.org/.
3. Virtual Expert Mass Spectrometrist (VEMS): http://yass.sdu.dk/.

2.2.2. Protein Analysis Tools

1. GPMAW: http://www.gpmaw.com.

2.2.3. Protein Sequence Databases

1. nr: ftp://ftp.ncbi.nih.gov/blast/db/FASTA/.
2. IPI (human): ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/ipi.HUMAN.fasta.gz.

3. UNIPROT (human): ftp://ftp.ebi.ac.uk/pub/databases/SPproteomes/fasta/proteomes/25.H_sapiens.fasta.gz.

2.2.4. Protein Sequence Annotations

1. Swiss-Prot: http://us.expasy.org/sprot/.
2. Human Protein Reference Database: http://www.hprd.org/.

2.2.5. Protein Modification Resources

1. Delta-mass: http://www.abrf.org/index.cfm?method=dm.home.
2. Resid: http://pir.georgetown.edu/cgi-bin/resid.
3. Unimod: http://www.unimod.org/.

3. Methods

3.1. Generating a Theoretical Spectrum

Although most search engines return an annotated spectrum it can be of immense value to generate theoretical spectra of other possible solutions. Especially with labile modifications as glutamic acid γ-carboxylation (and to some extent serine and threonine phosphorylation) it can be difficult to pinpoint and assign the exact position of the modified residue. Often, there are multiple residues that can be modified in a sequence, and distinguishing between the possible modification sites requires careful comparison of the experimental data and the theoretical spectra for each potential modified residue.

The mass of the protonated parent ion MH^+ is given by the sum

$$MH^+ = m_N + \Sigma m_i + m_C + m(H^+) \text{ (usually } \Sigma m_i + 19.01784 \text{)}$$

where m_N is the mass of the N-terminating group (usually a hydrogen), Σm_i is the sum of masses of i amino acid residues, m_C is the mass of the C-terminating group (usually a hydroxyl group), and $m(H^+)$ is the mass of a proton. The monoisotopic masses, m_i, of the common amino acids can be found in **Table 3**. **Table 1** lists the masses of the most common posttranslational modifications. The commonly observed fragment ion masses are then given by the following equations where the index is counted from the N-terminus for a- and b-type ions and from the C-terminus for y-type ions.

$$a_i = m_N + \Sigma m_i - m(e^-) - m(CO) = b_i - m(CO) \text{ (usually } \Sigma m_i - 26.98654).$$

$$b_i = m_N + \Sigma m_i - m(e^-) \text{ (usually } \Sigma m_i + 1.007276).$$

$$y_i = \Sigma m_i + m_C + m(H) + m(H^+) \text{ (usually } \Sigma m_i + 19.01784).$$

For a peptide with n residues the sum of the two corresponding ions can be calculated as:

Table 3
Masses of the 20 Common Amino Acid Residues and Carbamidomethylated Cysteine

Amino Acid	Abbreviation	Code	Monomer composition	Monoisotopic mass
Glycine	Gly	G	C_2H_3NO	57.021464
Alanine	Ala	A	C_3H_5NO	71.037114
Serine	Ser	S	$C_3H_5NO_2$	87.032028
Proline	Pro	P	C_5H_7NO	97.052764
Valine	Val	V	C_5H_9NO	99.068414
Threonine	Thr	T	$C_4H_7NO_2$	101.047679
Cysteine	Cys	C	C_3H_5NOS	103.009185
Isoleucine	Ile	I	$C_6H_{11}NO$	113.084064
Leucine	Leu	L	$C_6H_{11}NO$	113.084064
Asparagine	Asn	N	$C_4H_6N_2O_2$	114.042927
Aspartic acid	Asp	D	$C_4H_5NO_3$	115.026943
Glutamine	Gln	Q	$C_5H_8N_2O_2$	128.058578
Lysine	Lys	K	$C_6H_{12}N_2O$	128.094963
Glutamic acid	Glu	E	$C_5H_7NO_3$	129.042593
Methionine	Met	M	C_5H_9NOS	131.040485
Histidine	His	H	$C_6H_7N_3O$	137.058912
Phenylalanine	Phe	F	C_9H_9NO	147.068414
Arginine	Arg	R	$C_6H_{12}N_4O$	156.101111
Cysteine - cbm		C*	$C_5H_8N_2O_2S$	160.030648
Tyrosine	Tyr	Y	$C_9H_9NO_2$	163.063329
Tryptophan	Trp	W	$C_{11}H_{10}N_2O$	186.079313

$$b_i + y_{n\text{-}i} = MH^+ + m(H^+) = MH^+ + 1.007276$$

For example, once the b_2-ion has been identified (from the prominent a_2–b_2 pair) the mass of y_{n-2} can be calculated. A list of possible b_2 ions can be found in **Table 4**. Multiple fragmentations of the backbone can occur (but less frequently than single fragmentations) and it is not uncommon to observe internal fragments in the lower mass region (less than 700 Da), especially if there is a proline in the sequence. The theoretical masses of internal ions can be calculated as the sum of residue masses plus the mass of a proton. The information contained in the low mass region can be used to patchwork the proposed sequence ***(17)***. There is also useful information in the low mass region from immonium and related fragment ions that are characteristic of specific amino acids (*see* **Tables 1** and **5**). In addition, some amino acid residues undergo the

Table 4
Amino Acid Residue Composition and Masses of b_2-Ions With Carbamidomethylated Cysteine[a]

115.05	GG	203.10	TT	231.12	MV	258.14	EK	286.15	ER
129.07	AG	205.10	FG	232.08	AC	258.16	RT	286.16	VW
143.08	AA	209.10	AH	233.10	MT	259.09	EE	288.13	TW
145.06	GS	211.14	[I/L]P	235.11	AY	260.11	CV	288.15	MR
155.08	GP	212.10	NP	235.11	FS	260.11	MQ	289.10	CQ
157.10	GV	213.09	DP	235.12	HP	260.14	KM	289.13	CK
159.08	AS	213.16	[I/L]V	237.13	HV	261.09	EM	290.08	CE
159.08	GT	214.12	NV	239.11	HT	261.12	PY	292.08	CM
169.10	AP	214.13	GR	242.15	[I/L]Q	261.16	F[I/L]	292.13	QY
171.11	AV	215.10	DV	242.19	[I/L]K	262.09	CT	292.17	KY
171.11	G[I/L]	215.14	[I/L]T	243.11	NQ	262.12	FN	293.11	EY
172.07	GN	216.10	NT	243.13	E[I/L]	263.09	MM	294.17	HR
173.06	DG	216.10	QS	243.15	KN	263.10	DF	295.11	MY
173.09	AT	216.13	KS	244.09	DQ	263.14	VY	295.14	FF
175.07	SS	217.08	DT	244.09	EN	265.12	TY	298.10	CH
185.09	PS	217.08	ES	244.11	GW	266.12	HQ	300.17	[I/L]W
185.13	A[I/L]	218.06	CG	244.13	DK	266.16	HK	301.13	HY
186.09	AN	219.08	MS	244.14	RS	267.11	EH	301.13	NW
186.09	GQ	219.11	AF	245.08	DE	269.11	HM	302.11	DW
186.12	GK	221.09	GY	245.13	FP	270.19	[I/L]R	304.18	FR
187.07	AD	225.10	HS	245.13	[I/L]M	271.15	NR	308.11	CF
187.07	EG	226.12	PQ	246.09	MN	272.14	DR	311.14	FY
187.11	SV	226.16	KP	247.07	DM	274.12	SW	313.21	RR
189.07	GM	227.10	EP	247.14	FV	274.12	C[I/L]	315.15	QW
189.09	ST	227.18	[I/L] [I/L]	248.07	CS	275.08	CN	315.18	KW
195.09	GH	228.13	[I/L]N	249.12	FT	275.13	HH	316.13	EW
195.11	PP	228.13	QV	251.10	SY	276.06	CD	317.14	CR
197.13	PV	228.15	AR	251.15	H[I/L]	276.13	FQ	318.13	MW
199.11	PT	228.17	KV	252.11	HN	276.17	FK	320.17	RY
199.14	VV	229.09	NN	253.09	DH	277.12	EF	321.07	CC
200.10	AQ	229.10	MP	254.16	PR	277.15	[I/L]Y	324.10	CY
200.14	AK	229.12	D[I/L]	256.18	RV	278.11	NY	324.15	HW
201.09	AE	229.12	EV	257.12	QQ	279.10	DY	327.13	YY
201.12	[I/L]S	230.08	DN	257.16	KQ	279.12	FM	334.16	FW
201.12	TV	230.11	QT	257.20	KK	284.14	PW	343.19	RW
202.08	NS	230.15	KT	258.09	CP	285.13	FH	347.12	CW
203.07	DS	231.06	DD	258.11	EQ	285.17	QR	350.15	WY
203.08	AM	231.10	ET	258.12	AW	285.20	KR	373.17	WW

[a]The order of residues does not matter and [I/L] denotes an isoleucine or leucine residue; these two residues have exactly the same mass. An intense pair of ions separated by 28 Da (a CO group) in the low mass region is usually the signature of an a_2–b_2 pair, and the composition of the b_2-ion can be looked up in this table. An a_2–b_2 pair that does not fit with the masses in this table could be because of modification at the N-terminus. It is not uncommon to observe a number of internal dipeptides from multiple fragmentations that give rise to additional a_2–b_2 pairs.

Table 5
Low Mass Fragment Ions and Neutral Losses From the 20 Common Amino Acid Residues *(15,16)*[a]

Amino acid	Code	Immonium and fragment ions	Neutral loss
Glycine	G	30.03	
Alanine	A	44.05	
Serine	S	60.04	18.01 (H2O)
Proline	P	70.07	
Valine	V	72.08	
Threonine	T	74.06	18.01 (H2O)
Isoleucine	I	86.10	
Leucine	L	86.10	
Asparagine	N	87.06, 70.03	17.03 (NH3)
Aspartic acid	D	88.04, 70.03	18.01 (H2O)
Glutamine	Q	129.10, 101.07, 84.04, 56.05	17.03 (NH3)
Lysine	K	129.11, 101.11, 84.08, 56.05	17.03 (NH3)
Glutamic acid	E	102.05, 84.04	18.01 (H2O)
Methionine	M	104.06	48.00 (CH4S)
Histidine	H	110.07	
Phenylalanine	F	120.08	
Arginine	R	129.11, 115.09, 112.09, 87.09, 70.07, 60.06	17.03 (NH3)
Cysteine-cbm	C*	133.04	91.01 (C2H5NOS)
Tyrosine	Y	136.08	
Tryptophan	W	159.09, 132.08, 130.07	

[a]Cysteines are derivatized with iodoacetamide to form carbamidomethylcysteine (C*).

loss of small neutral molecules, like water or ammonia, leading to satellite peaks to the major fragment ion series (*see* **Tables 1** and **5**).

A number of software packages offer convenient tools for calculating theoretical masses for peptide digests of proteins, molecular masses of the peptides, and fragment ion masses. One of the most user friendly and versatile packages is general protein/mass analysis for Windows (GPMAW) (*see* Peri et al. for a brief description ***[18]***). **Figure 2** illustrates a typical simple application. After importing the protein sequence into GPMAW and doing a theoretical proteolytic digest, the masses for all ions in a theoretical tandem mass spectrum of an oxidized methionine peptide can be calculated. The experimental spectrum is shown in **Fig. 3** as it is returned from a database search using VEMS.

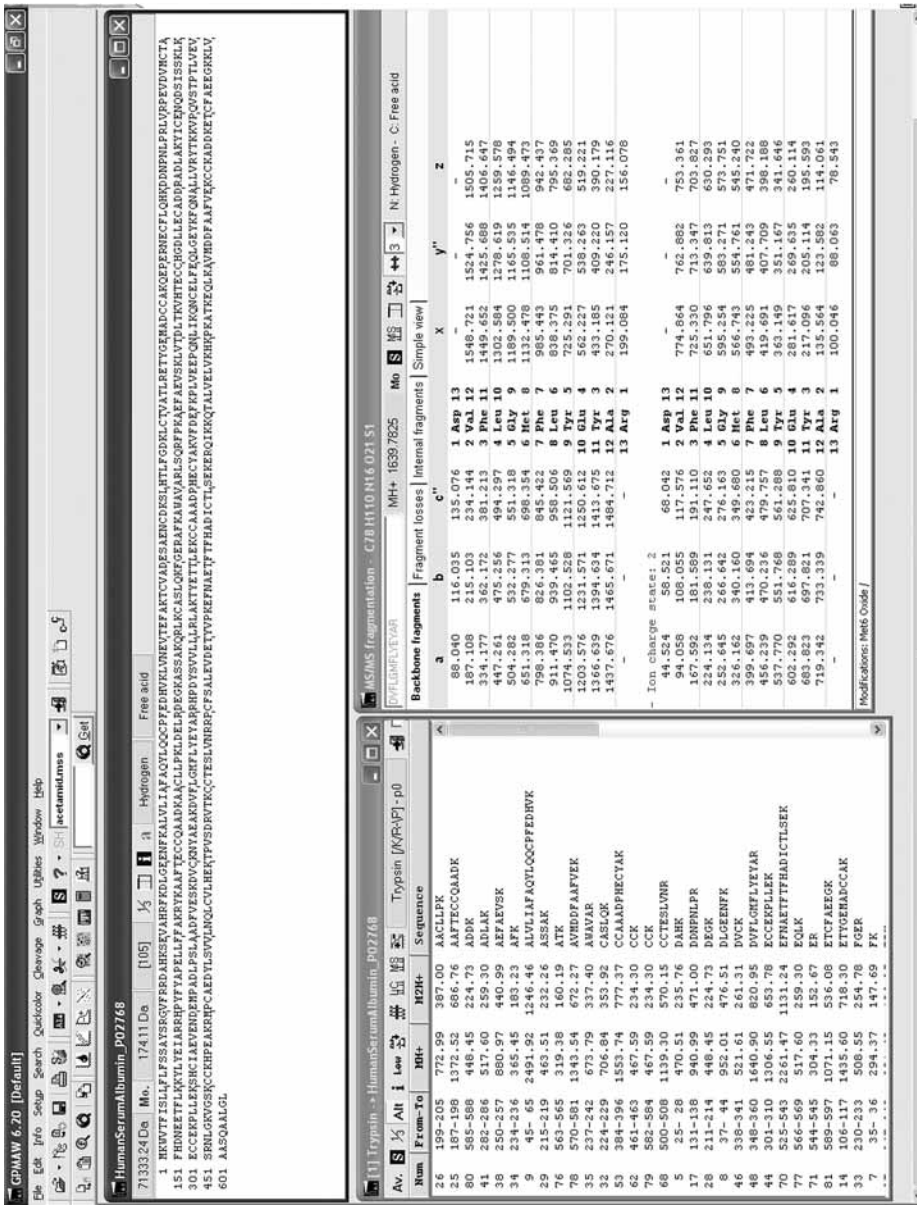


Fig. 2. The program GPMAW is a very versatile tool for protein analysis. In this simple example the protein sequence of human serum albumin is loaded (upper panel) and an in silico tryptic digestion is performed and the theoretical mass-to-charge ratios of the singly and doubly charged ions are calculated (lower-left panel). The theoretical masses of both singly and multiply charged fragment ions can then be calculated for each peptide (lower-right panel). The methionine in this example is flagged as being oxidized and the masses are adjusted throughout the calculations.

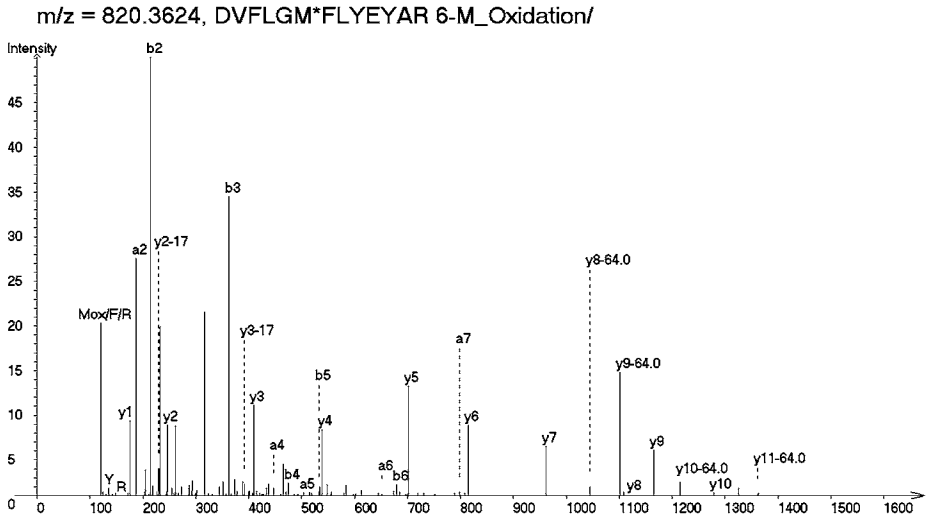


Fig. 3. Electrospray ionization QqTOF tandem mass spectrum of a doubly charged ion at *m/z* 820.36. Database searching with Virtual Expert Mass Spectrometrist (VEMS) against the IPI human database returns the protein human serum albumin and the peptide DVFLGMFLYEYAR with an oxidized methionine. The y-ion series from y_8 and up displays an additional series of rather intense peaks arising from the neutral loss of CH_3SOH from oxidized methionine residue (*see* **Note 5**). The figure was prepared using VEMS.

3.2. Database Searching

A number of software tools have been developed to identify the proteins from CID tandem mass spectra of the peptides. Different search engines yield somewhat complementary results and we often search the same dataset with two or three different programs ***(19–21)*** (*see* **Subheading 2.2.1.**). The VEMS software is freely available and offers a number of tools for database searching and validation. VEMS can search even very large datasets with any number of variable modifications, but care must be taken when defining the search parameters and interpreting the outcome. Depending on the organism, a database has to be selected (*see* **Subheading 2.2.3.**). The detailed use of MS-driven database search engines falls outside the scope of this chapter.

3.3. Assignment and Validation Steps

Most MS-based protein database search engines return a score as a measure of how well the theoretical spectrum matches the experimental data. High-scoring peptides are mostly correct, but there is a large gray area where only critical manual validation can sort out false-positive identifications. Identifying peptides that deviate in some way from what is expected should be examined closely

with the mantra that "extraordinary results require extraordinary proof." The following questions are mostly empirical but can be used to probe the confidence of the identified peptide.

3.3.1. Peptide Sequence

1. Does the peptide sequence conform to the experiment? In most proteomics studies the proteins have been digested into peptides with a sequence-specific protease prior to analysis. It has been shown that the most commonly used protease, trypsin, cleaves the peptide backbone to the C-terminal side of arginine or lysine residues with a very high degree of specificity ***(22)***. Therefore, the amino acid residue preceding the peptide should be arginine or lysine and the peptide C-terminal residue should be K or R. It is possible that there is nonspecific cleavage or that the protein has been processed prior to analysis, but nontryptic peptides in a tryptic protein digest should be examined critically. In many cases there is additional sequence information that can add confidence to seemingly semitryptic peptides where only one end of the peptide follows a tryptic cleavage pattern (for example as a consequence of the cleavage of signal peptides or that the peptide forms the C-terminus of the protein). Likewise, a peptide containing many internal lysines and arginines is unlikely to survive the incubation with trypsin (unless the peptide is modified so that the cleavage sites are masked).
2. Does the number of basic groups (H, K, R, and N-terminal amino group) correspond to the charge state? The charge state of peptide in ESI depends on several factors and the above rule of thumb should only be taken as a rough estimate. Three major interrelated factors are the basicity of the peptide residues, the conformation of the peptide, and the Coulomb repulsion between multiple protons ***(23)***. In many peptides the maximum charge state can be estimated by the number of basic residues (R, K, H, and the N-terminal amino group) but for longer peptides protonation can take place at the next most basic sites (P, Q, and W) ***(24)***. When a parent ion with a low-charge state is assigned a potential peptide sequence with several internal basic residues, it is necessary to carefully examine the spectra because the assignment could be erroneous. For example, the fragmentation at glycine residues to the C-terminal side is often of low intensity (*see* **Subheading 3.3.2.**, **step 11**), the assigned lysine (128.0950 Da) could be glutamine (128.0586) or the dipeptide AG (128.0586 Da), and arginine (156.1011 Da) could be the dipeptide VG (156.0742 Da).
3. How well does the peptide mass match the experimental mass? The mass accuracy of the instrumentation depends on the type of mass analyzer. A rough estimate would be that the mass accuracy of an ion trap or triple quadrupole should be better than 1.5 Da, and that of a QqTOF better than 0.15 Da. The mass analyzers usually perform a lot better than these values and the experimental setup should be checked if a large proportion of the peptide identifications are made with mass accuracies outside these values (*see* **Note 1**).
4. Are residues consistently modified? An obvious question to rise is if the observed residues are consistent with the experimental procedure. If cysteines have been

alkylated with iodoacetamide it should be worrisome if peptides were identified with nonmodified cysteine residues (or as being modified by another reagent).

3.3.2. Spectral Features

1. Is there a consecutive series of ions from sequential peptide fragments (e.g., > three y-ions)? In the case of tryptic peptides there is usually some continuity in the ion series (with some sequence dependencies, *see* **Subheadings 3.3.2.**, **steps 10** and **11**) and for QqTOF MS/MS spectra the ion series above the parent ion mass-to-charge ratio is usually y-ions. If high *m/z* b-ions are observed on a QqTOF instrument there has to be a basic amino acid residue in the N-terminal fragment. A rough rule of thumb for accepting the proposed peptide sequence is if there is a sequence tag of at least three amino acids. This is a somewhat arbitrary criterion and obviously there should not be glaring inconsistencies between the theoretical and experimental spectrum. For example, it should not be possible to extend the sequence tag ion series with an amino acid residue mass to include an intense ion in the series if this amino acid residue does not fit the retrieved sequence.
2. Is there an intense a_2–b_2 pair of ions separated by 28 Da? The b-ions can lose CO to form a-ions, and the a_2–b_2 ion pair separated by 28 Da is usually very prominent in the low mass region. By looking for an intense ion pair separated by 28 Da it is possible to guess at the amino acid composition by using **Table 4**. Additional a_2–b_2 ion pairs can arise from internal fragmentation, *see* **ref.** ***17***.
3. Do the observed immonium ions correspond to residues in the peptide (especially V, [I/L], H, F, Y, W)? Depending on the mass range of the instrument it is usually possible to get an idea if the peptide contains any of the listed amino acids. Some amino acid residues give rise to several low mass fragment ions (*see* **Tables 1** and **5**) that can increase confidence in determining the amino acid composition of the peptide.
4. Are the y_1-ions observed and in accordance with enzymatic specificity? Tryptic peptides give rise to intense y_1-ions because of the C-terminal position of the basic residues and this C-terminal residue should be consistent with the proposed sequence (*see* **Note 2**). Tryptic peptides have y_1-ions at *m/z* 147(K) and 175(R). The y_1-ions of chymotryptic peptides are at *m/z* 132 (I/L), 166(F), 182 (Y), and 205 (W).
5. How well do the fragment ion masses fit the theoretical masses? The assignment of fragment ions can be aided by looking at the fragment ion mass differences from the theoretical values. **Figure 4** shows an example from VEMS for the tandem mass spectrum shown in **Fig. 3** where the residual mass deviation (right pane) for correctly assigned peaks is small; larger deviations can be caused by overlapping peaks or poor ion statistics, but the ions assigned to a_7 and y_8 in **Fig. 3** do not fit very well and these assignments should be viewed with some scepticism.
6. Are multiply charged fragment ions observed? ESI MS/MS often gives rise to multiply charged ions and it is common to observe neutral losses of amino acids from the N-terminal end of tryptic peptides ***(5)***. It is important to note the charge state of the ions to avoid misassigning a singly charged ion to a multiply charged ion. A multiply charged ion in the tandem mass spectrum with a higher mass than the parent

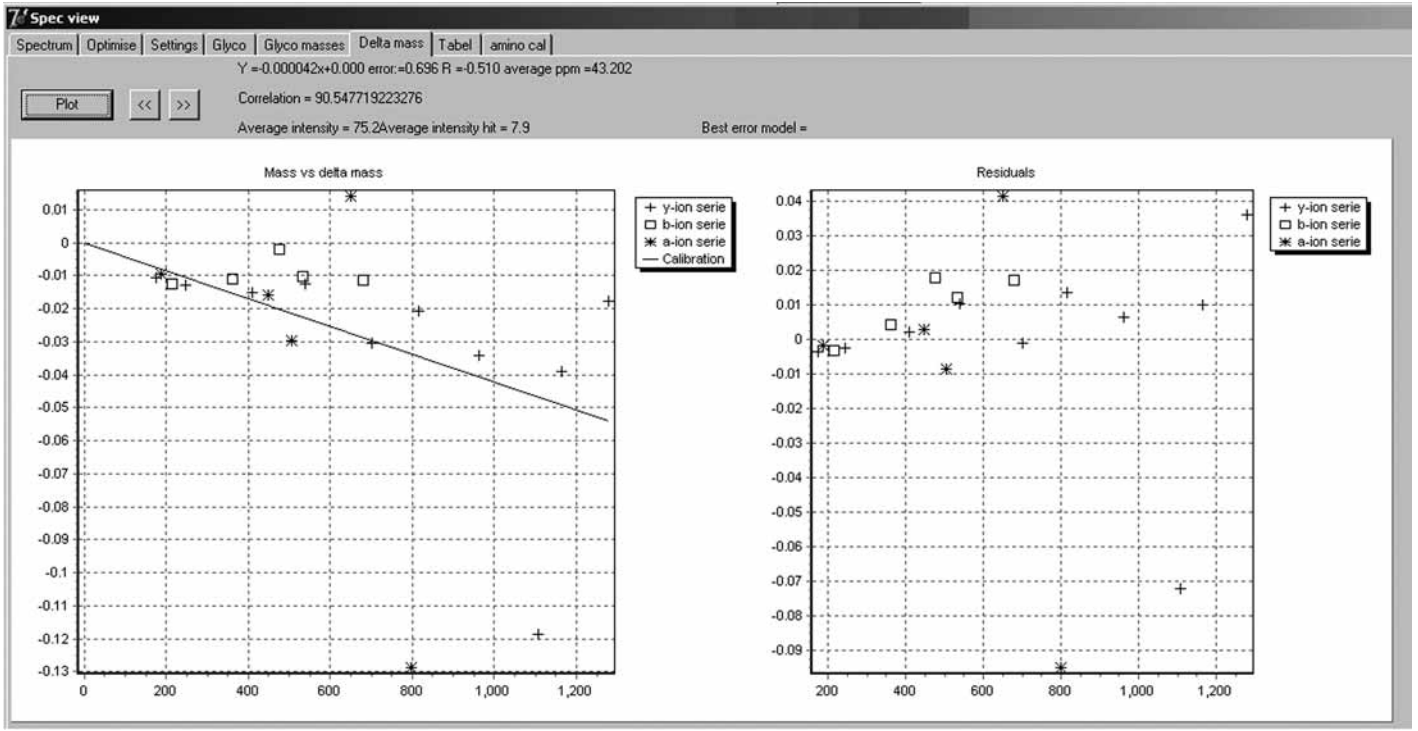


Fig. 4. Displaying the residual masses can often help the assignment of peaks and data validation. The mass accuracy of data from a QqTOF is quite sensitive to temperature fluctuations and a postacqusitional recalibration usually brings the mass deviation residuals below 0.05 Da. If the assigned peaks deviate beyond this the assignment should be checked—going back to the raw data prior to data processing as smoothing and centroiding is often helpful to discover overlapping peaks or other causes for poor mass accuracy. The figure is a screenshot from VEMS.

peptide mass suggests that the charge state of the parent ion has been wrongly assigned (e.g., observing a doubly charged fragment ion at *m/z* 750 when the software reports it has been fragmenting a doubly charged parent ion at *m/z* 600).

7. Are satellite ions observed from the loss of small neutral molecules? Intense ion series often have a series of associated satellite peaks from the neutral loss of small molecules (*see* **Table 5**), especially the loss of water from serine, threonine, aspartic acid, and glutamic acid residues, and ammonia from arginine, lysine, asparagine, and glutamine residues.
8. Can the amino acid sequence be permutated or changed to account for unassigned peaks? If there is a gap in the annotated y-ion series or an unannotated peak in an otherwise continuous ions series it is not uncommon that the order of two amino acid residues or composition can be changed to explain the discrepancy. For example, in an ion series containing an asparagine (114.0429 Da) residue one should check that there is not an ion midway to suggest two glycine residues (also 114.0429 Da).
9. If the sequence contains proline, are there internal fragment ions from multiple cleavages? Multiple peptide fragmentations are especially common for peptides with facile cleavage sites and especially MS/MS of peptides containing proline or aspartic acid display internal fragment ions.
10. Are intense fragment ions observed from cleavage on the N-terminal side of P and the C-terminal side of D? Low-energy CID fragmentation induced by a mobile proton ***(25)*** depends on the charge state of the peptide ion and on how the charges are sequestered by basic residues. Numerous studies have described the facile backbone cleavage N-terminal to proline (*see*, e.g., **ref. *26***) because the tertiary amine is more basic than the other backbone atoms. In cases where charges are sequestered (e.g., when the number of protons is lower than the number of arginine residues) there is an enhanced cleavage at aspartic acid residues ***(25)***.
11. Are fragment ions from the C-terminal side of P and G of low intensity or absent? Statistical analysis of tandem mass spectra databases has shown that there is a bias against fragmentation to the C-terminal side of proline and glycine residues ***(27,28)***.

3.3.3. Spectral Features of Posttranslational Modifications

1. Can the composite mass of the modified residue be found in a consecutive series of ions? The confidence with which one can identify posttranslationally modified residues depends crucially on the data quality and a direct observation of the composite mass from a modified residue in a consecutive series of ions is stronger evidence than just observing an altered parent ion mass that correspond to the PTM(s).
2. Are there characteristic neutral losses? In some cases the posttranslationally modified residue exhibits a characteristic intense loss that increases the confidence in the assignment—often a satellite peak from the parent ion displays the characteristic loss. For example, phosphorylated serine very easily loses a neutral phosphoric acid group and displays an ion series 98 Da less than the expected composite mass ion series; the modified serine can then be identified as a dehydroalanine residue weighing 69 Da.

3. Are there any characteristic low mass ions? Some posttranslationally modified residues are fairly stable under low-energy CID conditions and the immonium ion (and other fragment ions) can increase the confidence of the identification (e.g., phosphorylated tyrosine residues give rise to an immonium ion at 216.04 Da). Some PTMs give rise to multiple diagnostic ions (*see* **Table 1**) and observing all the fragment ions is stronger evidence; e.g., acetyl lysine gives rise to a set of low mass ions at *m/z* 84, 126, and 143.
4. Can the mass increase attributed to the PTM be explained by other means? Because many posttranslational modifications are made of the same few elements as amino acids are composed of, there are many possibilities of interpreting the data wrongly. When separating proteins on a polyacrylamide gel it is not unusual that nonpolymerized acrylamide reacts with cysteine residues ***(29,30)***. The propionamide attached to cysteine has exactly the same composition as an alanine residue and depending on the data quality it can be difficult to distinguish the two possibilities of having a propionamide–cysteine or an unmodified cysteine followed by an alanine residue. The solution to this problem is most often to make an informed guess based on the sample handling procedure. If the proteins have been separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis, reduced, and alkylated with iodoacetamide the former solution is most likely correct (there should be no free cysteine thiol groups after iodoacetamide treatment). Another example is after alkylation with iodoacetamide there is a side product where the N-terminal amino group has been carbamidomethylated—this modification adds exactly the same mass as a glycine residue (having the same elemental composition).
5. Can missed cleavages be rationalized by neighboring modified residues? Tryptic missed cleavages often occur at amino acid stretches with several adjacent basic residues leading to an additional R or K with a terminal position (either C- or N-terminal). Internal missed cleavages for trypsin are often associated with neighboring acidic groups (aspartate, glutamate, and phosphorylated S, T, or Y).
6. Is the modification consistent with known sequence motifs and target amino acids? Many posttranslational modifications only occur on a very small subset of residues determined by the amino acid residue, functional groups, size of the amino acid, or a sequence motif. Looking up information on the various types of modifications can be helpful when analyzing the spectra (*see* **Note 3**). For example, if an asparagine is converted to aspartic acid by deglycosylation treatment with the glycosidase PNGase F, the asparagines can tentatively assigned as being glycosylated. However, the asparagines should appear in the consensus sequon NXS/T where X can be any amino acid except proline, otherwise the N to D conversion most likely is caused by spontaneous deamidation.

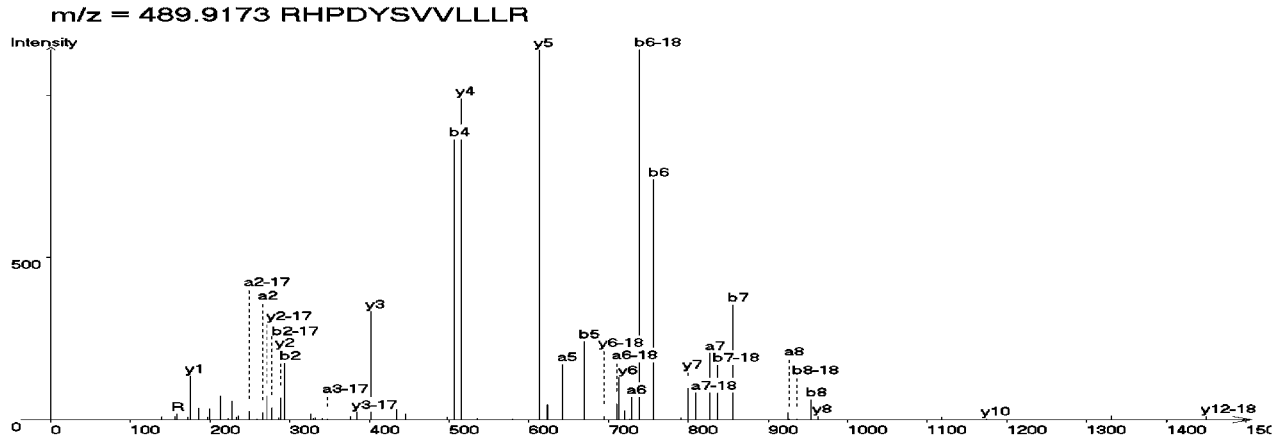
3.3.4. Precursor Ion

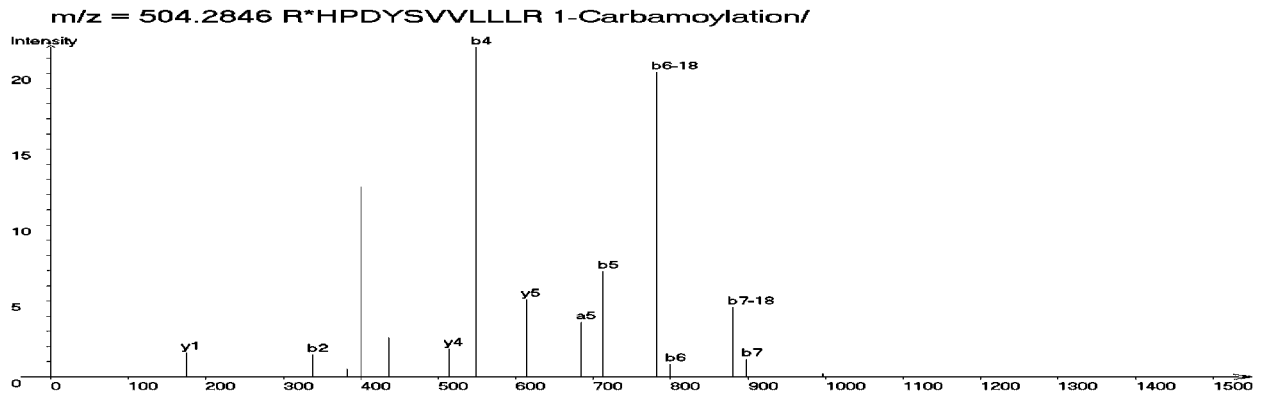
The selection of the parent ion and the determination of mass-to-charge ratio and charge state are crucial for the data quality. The problems listed next usually occur as a consequence of software limitations. Visual inspection of the parent ion in the MS survey spectrum can usually identify these problems.

1. Is the charge state correctly assigned? The assignment of charge states depends on the instrumentation and data quality, but even with high-resolution QqTOF instrumentation, the algorithm fails at times leading to an erroneous parent ion mass—manual inspection of the parent ion can resolve this.
2. Has the correct isotope peak been picked for mass assignment? If the wrong isotope peak has been chosen by the software, the mass assignment is usually off by 1 Da (or an integer number of ^{13}C atoms). This can result in peptides being identified as being deamidated, but the correct mass can be assigned manually.
3. Has more than one precursor ion been transmitted for fragmentation simultaneously? A very complex MS/MS spectrum can be caused by the simultaneous selection of more than one ion for fragmentation. In some cases, it is possible to identify both sequences from the data, but this usually requires manual interpretation of the spectrum.
4. Is the ion selected for MS/MS an in-source-generated fragment of another usually more intense ion? It is not uncommon for ions to undergo some in-source fragmentation, either as a loss of ammonia or a backbone cleavage at proline residues. Comparing the elution profiles of the parent ion and the in-source-generated fragment in liquid chromatoraphy–MS (they should have identical elution profiles) is usually helpful for detecting this problem. The data analysis can be handled by comparing the two tandem mass spectra and annotating the in-source-generated fragment ion spectrum based on similarity.

3.4. Examples

Protein chemistry and biology is overwhelmingly complex and the modifications listed in **Table 1** are only the most common that we look for. For more extensive lists of modifications web resources like Deltamass, Resid, and Unimod (*see* **Subheading 2.2.3.**) hold a wealth of information. A large body of experimental data has been collected on the fragmentation patterns of histone modifications like mono-, di-, and trimethylation and acetylation (*see* **Note 4**). Protein oxidation (*see* **Note 5**) can give rise to a very complex mixture of peptide isoforms and in the case of tryptophan (*see* **Note 6**) more than six oxidized forms have been reported. Sample preparation plays a crucial role in which side product can be formed. If a gel approach has been used for protein purification, the search parameters should reflect that acrylamide can react with cysteines to form propionamide. If the proteins have been digested in the presence of high concentrations of urea there is most likely a lot of carbamoylation. If the alkylation of cysteines with iodoacetamide has not been quenched prior to trypsination there will also be carbamidomethylation of the N-terminal amines. In general sample handling-induced modifications are usually substoichiometric and will mostly affect the proteins with the highest concentration. It is fairly unlikely to identify low-abundance proteins solely based on modified peptides (unless a sample handling-related modification is almost quantitative in which case it should be reflected in the peptides assigned to the most abundant proteins).


m/z = 489.9173 RHPDYSVVLLLR
Intensity
y5
b6-18
y4
b4
b6
500
a2-17
a2
y2-17
b2-17
y2
b2
y1
R
a3-17
y3
y3-17
b5
y6-18
a6-18
a5
y6
a6
y7
a7
b7-18
a7-18
b7
a8
b8-18
b8
y8
y10
y12-18
0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400


m/z = 504.2846 R*HPDYSVVLLLR 1-Carbamoylation/
Intensity
b4
b6-18
20
15
10
b5
y5
5
a5
b7-18
y1
b2
y4
b6
b7
0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500

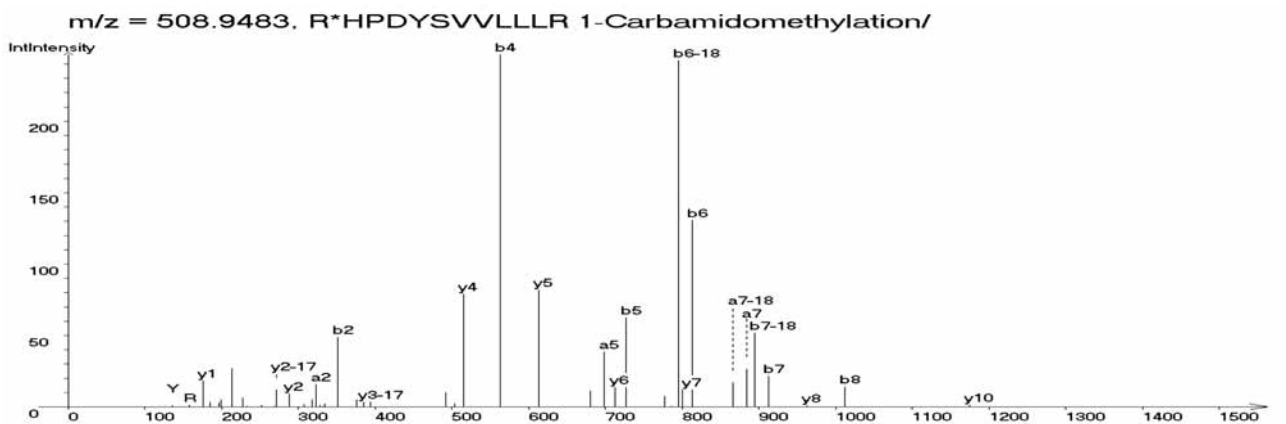


Fig. 5. Three electrospray ionization QqTOF tandem mass spectra of triply charged ion at *m/z* 489.92, 504.28, and 508.95. Database searching with VEMS against the IPI human database returns three different versions of the same peptide from human serum albumin, where the N-terminus has been derivatized with a carbamoyl group (from urea) or a carbamidomethyl group (from iodoacetamide). It is typical to see these unwanted side products of sample handling for the most abundant proteins.

The assignment and validation of tandem mass spectra of posttranslationally modified peptides is not an easy task. This chapter will end with a few examples taken from a human serum sample that was denatured in 8 *M* urea, reduced with dithiothreitol, alkylated with iodoacetamide, diluted to 2 *M* urea, and digested overnight with trypsin at 37°C.

Oxidation of methionine is a commonly occurring phenomenon both in vivo and in vitro. **Figure 3** shows a nice example where the search engine has returned the peptide sequence DVFLGMFLYEYAR with an oxidized methionine. The y_1-ion fits with a C-terminal arginine and the intense a_2–b_2 pair fits with the dipeptide DV. There is a b-ion series in the low mass region and a long continuous y-ion series and most ions are explained by this sequence. There is an ion at *m/z* 120 that could be the immonium ion from both phenylalanine and methionine sulfoxide, but the presence of oxidized methionine is confirmed by the intense neutral loss of 64 Da (CH_3SOH). In some cases, it can be helpful to look at the mass accuracy of the fragment ions to increase confidence in the spectral assignment, an example is shown in **Fig 4**.

Three very similar spectra of the peptide RHPDYSVVLLLR are displayed in **Fig. 5** and assigned to the normal peptide, a carbamoylated and a carbamidomethylated version. The two basic residues, RH, at the N-terminus gives rise to a fairly intense b-ion series. The peptides differ at the N-terminus as can be seen from the identical y-ion series and the shifted b-ion series. The 14-Da mass difference (*see* **Note 7**) between the spectrum of the carbamoylated and the carbamidomethylated peptide could just as well be explained by a methylation of the N-terminal arginine. The parsimonious interpretation is that it is unlikely to have two modifications to the same residue and if the arginine indeed was methylated this peptide containing methylarginine should also be found in a noncarbamoylated form.

4. Notes

1. The most obvious reason for poor mass accuracy is that the calibration has drifted and the instrument needs to be recalibrated. The mass accuracy in an ion trap depends critically on space charging and one could consider reducing the fill time/number of ions per scan or performing a zoom scan on the ions selected for fragmentation to determine charge state and mass. In a QqTOF, the mass accuracy of a well-calibrated instrument depends on having sufficient ion statistics to determine the peak maximum. Some deviation can be caused by other ions or background noise interfering with the isotope pattern peak maxima, but the most common problem is a mass drift in calibration caused by temperature fluctuations (affecting the length of the flight tube). This can be alleviated by postacquisitional recalibration.
2. Ion traps have a low mass cut-off in tandem mass spectra at approximately one-third of the parent ion mass-to-charge ratio. Therefore, it is often not possible to detect the low mass ions.

3. The removal of signaling peptides can be predicted and correlated with the observed sequence. The N-terminus of most proteins is heavily processed and it is estimated that 80–90% of all proteins are acetylated at the N-terminal amino group. Acetylation usually takes place on methionine or if the penultimate residue has a small radius of gyration (G, A, S, C, T, P, or V) the methionine is cleaved off by an aminopeptidase and the penultimate residue is acetylated. Hence, detecting an N-terminally acetylated tryptophan should make the alarm bells go off. A survey of acetylated N-terminal residues has recently been published ***(31)***.
4. Histone modifications have been widely studied and there is a large number of modifications with only the most commonly observed ones listed in **Table 1**. Differentiating between near isobaric-modified residues, such as trimethylated lysine and acetyl-lysine, requires high mass accuracy and careful analysis of the spectra to assign diagnostic marker ions and neutral loss series ***(32)***. Further details of fragmentation pathways and structures for the ions can be found for methylated arginine ***(33)***, dimethylated arginine ***(34)***, and acetyllysine ***(35)***.
5. For a brief review of protein oxidation, *see* Berlett and Stadtman ***(36)***. The fragmentation behavior of oxidized methionine ***(37)*** and oxidized carbamidomethylcysteine ***(38)*** is similar. The mono-oxidized form has a fairly intense neutral loss of 64 Da from methionine sulfoxide and 107 Da from mono-oxidized carbamidomethylcysteine, whereas the neutral loss from the dioxidized is of low abundance.
6. The structures of tryptophan oxidation products can be found on http://www.arbf.org/images/misc/dmass32.jpg.
7. There are many modifications that can give rise to 14-Da mass differences. Fairly conservative amino acid conversions (G to A, V to L, N to Q, D to E, etc.) and other modifications (carbamoylation to carbamidomethylation, carbamidomethylation to propionamide) give rise to 14-Da mass shifts. Therefore, the data should be carefully scrutinized before reporting methylations.
8. The b_1-ion can be seen for arginine, lysine, histidine, and methionine that can form alternative stabilizing structures by means of their side chains ***(39)***.

Acknowledgments

R. M. and J. B. both gratefully acknowledge the Carlsberg Foundation for financial support.

References

1. Roepstorff, P. and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **11,** 601.
2. Biemann, K. (1990) Appendix 5. Nomenclature for peptide fragment ions (positive ions). *Methods Enzymol.* **193,** 886–887.
3. Johnson, R. S., Martin, S. A., Biemann, K., Stults, J. T., and Watson, J. T. (1987) Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Anal. Chem.* **59,** 2621–2625.

4. Steen, H. and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5,** 699–711.
5. Salek, M. and Lehmann, W. D. (2003) Neutral loss of amino acid residues from protonated peptides in collision-induced dissociation generates N- or C-terminal sequence ladders. *J. Mass Spectrom.* **38,** 1143–1149.
6. Cooper, H. J., Hakansson, K., and Marshall, A. G. (2005) The role of electron capture dissociation in biomolecular analysis. *Mass Spectrom. Rev.* **24,** 201–222.
7. Zubarev, R. A. (2004) Electron-capture dissociation tandem mass spectrometry. *Curr. Opin. Biotechnol.* **15,** 12–16.
8. Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. USA* **101,** 9528–9533.
9. Mirgorodskaya, E., Roepstorff, P., and Zubarev, R. A. (1999) Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal. Chem.* **71,** 4431–4436.
10. Hakansson, K., Cooper, H. J., Emmett, M. R., Costello, C. E., Marshall, A. G., and Nilsson, C. L. (2001) Electron capture dissociation and infrared multiphoton dissociation MS/MS of an N-glycosylated tryptic peptic to yield complementary sequence information. *Anal. Chem.* **73,** 4530–4536.
11. Hogan, J. M., Pitteri, S. J., Chrisman, P. A., and McLuckey, S. A. (2005) Complementary structural information from a tryptic N-linked glycopeptide via electron transfer ion/ion reactions and collision-induced dissociation. *J. Proteome Res.* **4,** 628–632.
12. Stensballe, A., Jensen, O. N., Olsen, J. V., Haselmann, K. F., and Zubarev, R. A. (2000) Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Commun. Mass Spectrom.* **14,** 1793–1800.
13. Kelleher, N. L., Zubarev, R. A., Bush, K., et al. (1999) Localization of labile post-translational modifications by electron capture dissociation: the case of gamma-carboxyglutamic acid. *Anal. Chem.* **71,** 4250–4253.
14. Johnson, R. S., Davis, M. T., Taylor, J. A., and Patterson, S. D. (2005) Informatics for protein identification by mass spectrometry. *Methods* **35,** 223–236.
15. Falick, A. M., Hines, W. M., Medzihradszky, K. F., Baldwin, M. A., and Gibson, B. W. (1993) Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass-spectrometry. *J. Amer. Soc. Mass Spectrom.* **4,** 882–893.
16. Papayannopoulos, I. A. (1995) The interpretation of collision-induced dissociation tandem mass-spectra of peptides. *Mass Spectrom. Rev.* **14,** 49–73.
17. Schlosser, A. and Lehmann, W. D. (2002) Patchwork peptide sequencing: extraction of sequence information from accurate mass data of peptide tandem mass spectra recorded at high resolution. *Proteomics* **2,** 524–533.
18. Peri, S., Steen, H., and Pandey, A. (2001) GPMAW—a software tool for analyzing proteins and peptides. *Trends Biochem. Sci.* **26,** 687–689.
19. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567.

20. Matthiesen, R., Bunkenborg, J., Stensballe, A., Jensen, O. N., Welinder, K. G., and Bauw, G. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* **4,** 2583–2593.
21. Craig, R. and Beavis, R. C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **17,** 2310–2316.
22. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell Proteomics* **3,** 608–614.
23. Pallante, G. A. and Cassady, C. J. (2002) Effects of peptide chain length on the gas-phase proton transfer properties of doubly-protonated ions from bradykinin and its N-terminal fragment peptides. *Int. J. Mass Spectrom.* **219,** 115–131.
24. Schnier, P. D., Gross, D. S., and Williams, E. R. (1995) On the maximum charge-state and proton-transfer reactivity of peptide and protein ions formed by electrospray-ionization. *J. Amer. Soc. Mass Spectrom.* **6,** 1086–1097.
25. Wysocki, V. H., Tsaprailis, G., Smith, L. L., and Breci, L. A. (2000) Special feature: commentary—mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **35,** 1399–1406.
26. Breci, L. A., Tabb, D. L., Yates, J. R., and Wysocki, V. H. (2003) Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal. Chem.* **75,** 1963–1971.
27. Kapp, E. A., Schutz, F., Reid, G. E., et al. (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* **75,** 6251–6264.
28. Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H., Lin, D., and Yates, J. R. (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* **75,** 1155–1163.
29. Hall, S. C., Smith, D. M., Masiarz, F. R., et al. (1993) Mass spectrometric and Edman sequencing of Lipocortin-I isolated by 2-dimensional SDS PAGE of human-melanoma lysates. *Proc. Nat. Acad. Sci. USA* **90,** 1927–1931.
30. Hamdan, M., Bordini, E., Galvani, M., and Righetti, P. G. (2001) Protein alkylation by acrylamide, its N-substituted derivatives and cross-linkers and its relevance to proteomics: a matrix assisted laser desorption/ionization-time of flight-mass spectrometry study. *Electrophoresis* **22,** 1633–1644.
31. Polevoda, B. and Sherman, F. (2003) N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. *J. Mol. Biol.* **325,** 595–622.
32. Zhang, K., Yau, P. M., Chandrasekhar, B., et al. (2004) Differentiation between peptides containing acetylated or tri-methylated lysines by mass spectrometry: an application for determining lysine 9 acetylation and methylation of histone H3. *Proteomics* **4,** 1–10.
33. Gehrig, P. M., Hunziker, P. E., Zahariev, S., and Pongor, S. (2004) Fragmentation pathways of N(G)-methylated and unmodified arginine residues in peptides studied by ESI-MS/MS and MALDI-MS. *J. Am. Soc. Mass Spectrom.* **15,** 142–149.

34. Rappsilber, J., Friesen, W. J., Paushkin, S., Dreyfuss, G., and Mann, M. (2003) Detection of arginine dimethylated peptides by parallel precursor ion scanning mass spectrometry in positive ion mode. *Anal. Chem.* **75,** 3107–3114.
35. Kim, J. Y., Kim, K. W., Kwon, H. J., Lee, D. W., and Yoo, J. S. (2002) Probing lysine acetylation with a modification-specific marker ion using high-performance liquid chromatography/electrospray-mass spectrometry with collision-induced dissociation. *Anal. Chem.* **74,** 5443–5449.
36. Berlett, B. S. and Stadtman, E. R. (1997) Protein oxidation in aging, disease, and oxidative stress. *J. Biol. Chem.* **272,** 20,313–20,316.
37. Lagerwerf, F. M., vandeWeert, M., Heerma, W., and Haverkamp, J. (1996) Identification of oxidized methionine in peptides. *Rapid Commun. Mass Spectrom.* **10,** 1905–1910.
38. Steen, H. and Mann, M. (2001) Similarity between condensed phase and gas phase chemistry: fragmentation of peptides containing oxidized cysteine residues and its implications for proteomics. *J. Amer. Soc. Mass Spectrom.* **12,** 228–232.
39. Farrugia, J. M., O'Hair, R. A. J., and Reid, G. E. (2001) Do all b(2) ions have oxazolone structures? Multistage mass spectrometry and ab initio studies on protonated N-acyl amino acid methyl ester model systems. *Int. J. Mass Spectrom.* **210,** 71–87.

11

Retention Time Prediction and Protein Identification

Magnus Palmblad

Summary

Proteins are commonly identified through enzymatic digestion and generation of short sequence tags or fingerprints of peptide masses by mass spectrometry. Separation methods, such as liquid chromatography and electrophoresis, are often used to fractionate complex protein or peptide mixtures and these separations also provide information on the different species, such as molecular weight and isoelectric point from electrophoresis and hydrophobicity in reversed-phase chromatography. These are also properties that can be predicted from amino acid sequences derived from genomic sequences and used in protein identification. This chapter reviews recently introduced methods based on retention time prediction to extract information from chromatographic separations and the applications to protein identification in organisms with small and large genomes. Novel data on retention time prediction of posttranslationally modified peptides is also presented.

Key Words: Liquid chromatography; mass spectrometry; prediction; retention time; peptide; protein identification.

1. Introduction

Proteins can be identified by comparing specific properties predicted from translated genome sequences with the same properties measured by analytical techniques such as mass spectrometry (MS). Ideally, individual proteins are isolated, sequenced, and characterized in a "top-down" fashion, but because of intrinsic limitations in current separation and mass spectrometric technology, proteins are digested by one or more specific enzymes into shorter peptides and corresponding sequence tags or fingerprints of many such peptides are used to identify the protein, or reconstruct the sequence and characteristic modifications of the protein from the "bottom-up." When introducing a complex sample, such as a total protein enzymatic digest, to the mass spectrometer, suppression in ionization and detection limits the dynamic range and the ability to detect

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

low abundant proteins. To reduce the complexity and separate abundant species from less abundant ones, methods such as two-dimensional gel electrophoresis of proteins ***(1,2)*** and liquid chromatography (LC) of peptides ***(3–10)*** have been used with great success. In addition to the primary function of increasing the analytical dynamic range, the separations themselves provide information on the analytes. In sodium dodecyl sulfate-polyacrylamide gel electrophoresis, this is the molecular weight of the protein, in isoelectric focusing the isoelectric point (the pH at which the protein or peptide has no net charge) and in reversed-phase chromatography, the hydrophobic nature of the peptides or proteins ***(11,12)***. These are all properties that can be predicted from peptide sequences, not unlike molecular masses but less accurately, thereby constraining database searches and assisting protein identification, especially in absence of high-quality tandem mass spectra or peptide mass fingerprints.

Given a measured mass of a peptide from an enzymatic digest and mass measurement uncertainty, there is a limited number of possible matching peptides from proteins in any given sequence database that are within the measurement error of the observed mass ***(13,14)***. If mass measurement errors are very small, less than 1 ppm, close to which can be achieved in Fourier transform ion cyclotron resonance (FTICR) MS ***(15)***, there may exist a peptide of each protein that is unique in the organism's proteome within the measurement uncertainty ***(10,14)***. These peptides can then be used as "accurate mass tags" for protein identification ***(14)***. In general, however, mass accuracy is insufficient to identify proteins based on a single tryptic peptide mass, requiring either a pattern of several peptides or additional information on the peptides, such as short sequence tags or otherwise informative tandem mass spectra, for unambiguous protein identification.

The accurate mass measurements, together with information on protein size, peptide or protein isoelectric point, or peptide retention times, form a multidimensional protein-dependent pattern, and as previously stated, these patterns can also be predicted from protein and peptide sequences that in turn are predicted from genome sequence databases. It is this fact that enables protein identification using pattern recognition (*see* **Fig. 1**). Pattern recognition, or the detection of patterns using knowledge of the rules generating those patterns, includes all existing methods for protein identification by MS. When good-quality tandem mass spectra are available, these are in general sufficient to identify the protein (or at least the expressed gene) based on short sequence tags. In absence of such tandem mass spectra, proteins may still be identified based on accurately measured mass and retention times for one or more peptides. Accuracy in retention times refers to the closeness to predicted retention times.

The retention time of peptides in reversed-phase chromatography using linear gradients is often observed to have a linear mass dependence (*see* **Fig. 2**). Nonlinearities, such as from using nonlinear gradients, require at least a nonlinear

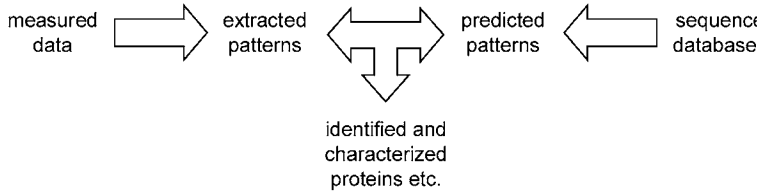


Fig. 1. Generic pattern recognition used for all protein identification by mass spectrometry. The key concept is that any property that can be predicted from sequence databases and measured experimentally can be used in protein identification.

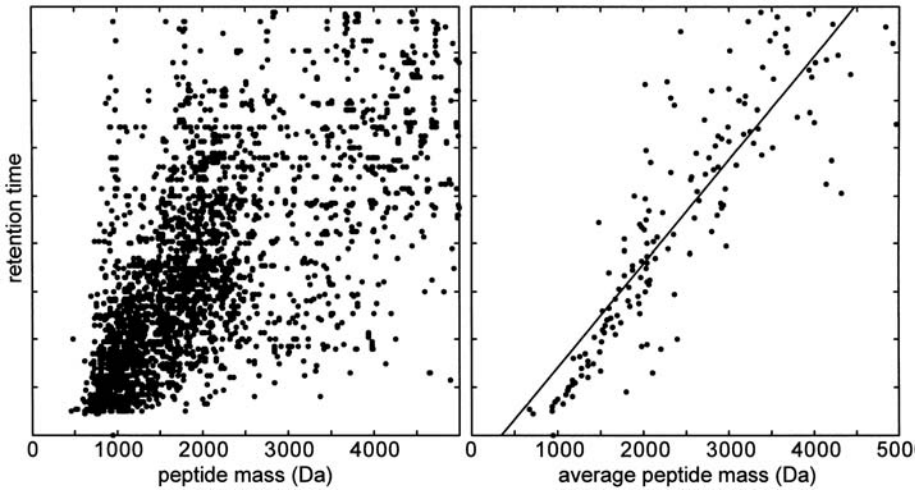


Fig. 2. The retention time of peptides in reversed-phase chromatography using linear gradients is often observed to have a linear mass dependence in the average over all peptide sequences (LC-FTICR data from a *Yersinia pestis* cytosolic protein tryptic digest ***[25]***). Nonlinearities, such as those appearing when using highly nonlinear gradients, require a nonlinear scaling function of either measured or predicted retention times.

scaling function of either measured or predicted retention times. An intuitive and simple model with a small number of parameters for establishing a quantitative structure–activity relationship between amino acid sequences and retention times in LC are thus linear functions of amino acid composition, sometimes with special consideration of terminal residues or other modifications ***(16)***. Using a model similar to that described by Hodges et al. ***(17–19)***, we showed how retention time prediction of peptides can be combined with accurate mass from FTICR MS to improve the statistical confidence in identifications of human proteins ***(16)***. A similar "accurate mass and time tag" approach was subsequently demonstrated on prokaryotic proteomes by LC-FTICR MS by Petritis et al. ***(20)***

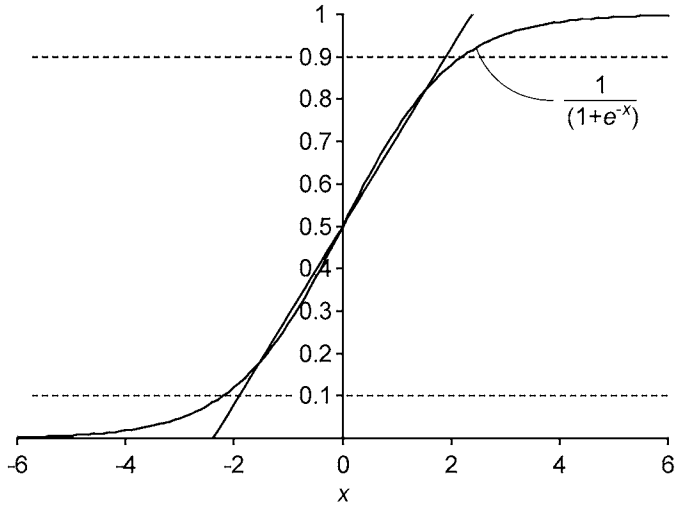


Fig. 3. A linear model and an artificial neural network (ANN) with no hidden layers and a sigmoid output function where retention times are mapped to the most linear part (between dashed lines) are quite similar (within 4%). If a larger training set is available, more input and hidden layer neurons can be added to the ANN and the exact peptide sequence taken into account, improving accuracy of prediction but requiring a much larger training set. The training set should contain many more unique peptides than the number of free parameters in the model to avoid overfitting to measured data.

and LC-time-of-flight MS by Strittmatter et al. ***(21)***, using normalized retention times and an artificial neural network (ANN) taking the amino acid composition as input and with sigmoid transfer and output functions. The improvement from adding nonlinear nodes in a hidden layer in the ANN was found to be small ***(20)***, and the linear model ***(16)*** and the ANN with no hidden layers ***(20)*** are very similar because observed retention times were mapped to the most linear part of the sigmoid output of the ANN (*see* **Fig. 3**). However, optimizing the output function or the part of the output function to which measured data is mapped, as well as using a larger number of input neurons, taking the amino acid sequence into account, should improve the accuracy of the predictor. Chromatographic reproducibility of relative (internally calibrated) retention times sets a lower limit on achievable prediction imprecision.

Retention time prediction can also be used as a discriminant function in SEQUEST tandem mass spectrometry (MS/MS) database searches ***(22,23)***, although the benefit is less significant than when using accurate mass alone.

2. Theory

Retention time data from a number of samples, systems, and experimental protocols ***(16,24,25)*** have been analyzed. Universally, peptide retention times were predicted by a linear combination of unique retention coefficients for each amino acid according to

$$t_{\text{calc}} = \sum_{i=1}^{20} n_i c_i + t_0 \quad (1)$$

where c_i is the retention coefficient of amino acid *i*, of which there are n_i in the peptide, and t_0 compensates for void volumes and any delay between sample injection and acquisition of mass spectra. The coefficients c_i are similar to the weights in the 20-0-1 neural network used by Petritis et al. ***(20)***.

To predict retention times of peptides with some of the most frequently studied posttranslational modifications, four modified amino acids (methionine-*S*-oxide, phosphoserine, phosphothreonine, and phosphotyrosine) were added to the right-hand side of **Eq. 1**:

$$t_{\text{calc}} = \sum_{i=1}^{24} n_i c_i + t_0 \quad (2)$$

The 20 or 24 c_i and t_0 were determined by least-squares fitting t_{calc} to measured retention times of 100–500 peptides from standard proteins, abundant proteins in the samples, or as identified by MS/MS. In the latter case, retention time prediction is subsequently used to increase the confidence in identification of proteins for which no MS/MS data of sufficient quality is available. All software was written in C and run on standard single-processor platforms under Cygwin ***(26)*** or Linux. Experimental LC–MS/MS data on modified peptides was kindly provided by Rune Mathiessen and Shabaz Mohammed using the virtual expert mass spectrometrist program ***(27)*** for the initial data analysis and extraction of retention times.

Proteins are identified by comparing likelihood ratios of the observed masses and retention times against those predicted from the proteome under study and against a large number of unrelated or random proteins of the same size and amino acid distributions (*see* **Fig. 4** and **Table 1**). The random or "decoy" protein database can be generated by filtering sequences with little homology from other organisms or constructed by reversing the sequences in the database or using Markov chains ***(28)***. A conservative measure of statistical significance is thus given by the frequency of random (false) matches to experimental data by random protein sequences. Unlike accurate mass and time tags, complete knowledge of all posttranslational modifications, nontryptic peptides, and contamination from other species is not assumed or necessary.

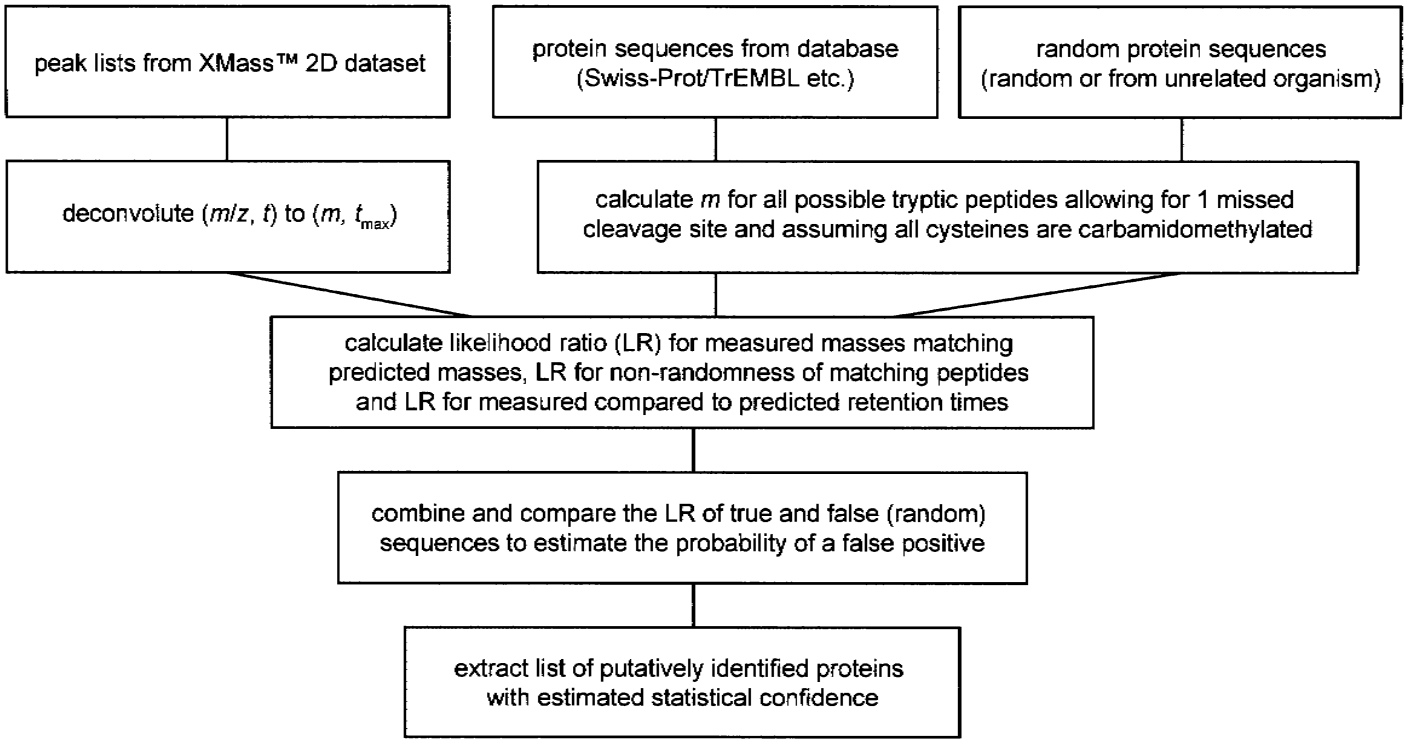


Fig. 4. Protein identification by multimodal pattern recognition (accurate masses, retention time, and nonrandomness of tryptic digestion ***[16,24,32]***). Likelihood ratios are ratios of probabilities of observing measured masses (given the mass measurement error), retention times (given the accuracy of prediction), and "runs" in the sequence coverage for peptides from a particular protein in the database to those probabilities of random proteins. (Reproduced from **ref.** ***24*** with permission.)

Table 1
A Single Example of Improvement in Protein Identification in a Body Fluid (Amniotic Fluid) by Retention Time Prediction *(24)*

Accurate mass +		Nonrandomness	
		–	+
Retention time	**–**	**4**	**28**
prediction	**+**	**15**	**43**

The numbers refer to the number of proteins putatively identified using accurate mass with (+) and without (–) retention time prediction and nonrandomness of enzymatic digestion. Strittmatter et al. ***(22)*** also noted a significant increase in the number of confident peptide identifications by tandem mass spectrometry using peptide retention times.

3. Results and Discussion

In **Fig. 2**, the retention times of peptides from a *Yersinia pestis* cytosolic protein tryptic digest ***(25)*** are shown to exhibit a strongly positive mass dependence. This is particularly emphasized for tryptic peptides, because these have few lysine and arginine residues, the two most hydrophilic amino acids. If tryptic digestion is complete, peptides contain at most one lysine or arginine in the C-terminus, hence longer tryptic peptides have more hydrophobic residues.

The accuracy of the predictor was found to be at best around within 6–7% (*see* **Fig. 5**). Retention time prediction was as accurate for phosphorylated peptides as for unmodified peptides, provided the training set contained a sufficient number of modified peptides (*see* **Fig. 6**). It appears that the training set should contain at least approx 10 peptides incorporating each amino acid and modification for all c_i to converge (*see* **Fig. 7**). The c_i values scale with hydrophobicity as expected from literature, i.e., the more hydrophobic, the higher the retention coefficient c_i, although this is a somewhat circular argument since hydrophobicity is often determined experimentally by reversed-phase chromatography ***(12)***. The retention coefficients are dependent on chromatographic conditions, such as mobile phase composition and pH. The pH directly influences the charge of peptide side chain groups and termini and, hence, the hydrophobicity (charged residues being more hydrophilic) and retention coefficients in reversed-phase chromatography ***(29,30)***. The t_0 values were usually near the observed void time. The overall positive c_i also implies a size dependency of the retention of tryptic peptides, i.e., the longer the peptide, the longer the retention time as shown in **Fig. 2**. The models used so far are not the only models for predicting reversed-phase chromatographic retention of peptides. More sophisticated approaches would account for the actual sequence or even predicted secondary structure ***(31)***.

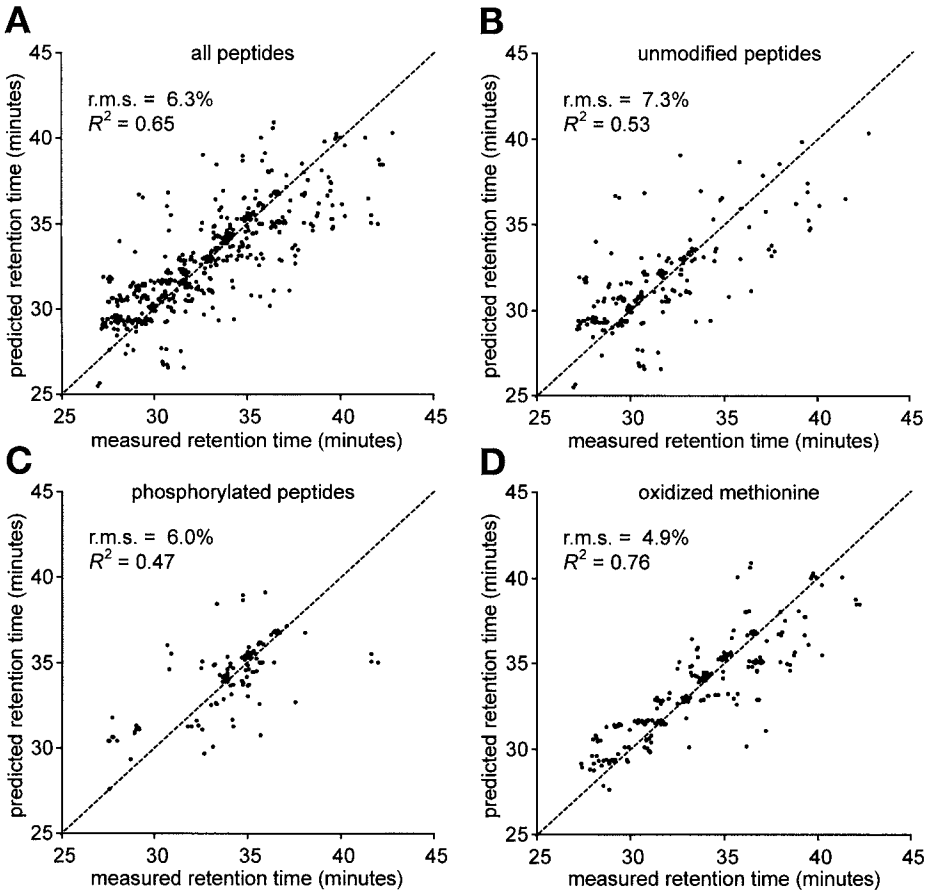


Fig. 5. Round-robin validation of the retention time predictor for modified peptides. Predicted vs measured retention time for 488 nonredundant peptides identified in nine separate liquid chromatography–tandem mass spectrometry runs of samples enriched for phosphopeptides using a linear reversed-phase gradient and least-squares fitting (regression) to the linear model in **Eq. 2**. The retention times of modified peptides were not harder to predict than unmodified peptides. In the round-robin, 90% of the dataset was used for training and 10% for validation. The validation set was then varied 10 times to use the entire dataset (data kindly provided by Rune Matthiesen and Shabaz Mohammed at the University of Southern Denmark, Odense, Denmark).

Information on physicochemical properties of peptides, such as hydrophobicity, that can be predicted from the peptide sequence, can assist or validate protein identification by peptide mass fingerprinting. This information is readily available in LC–MS and LC–MS/MS datasets.

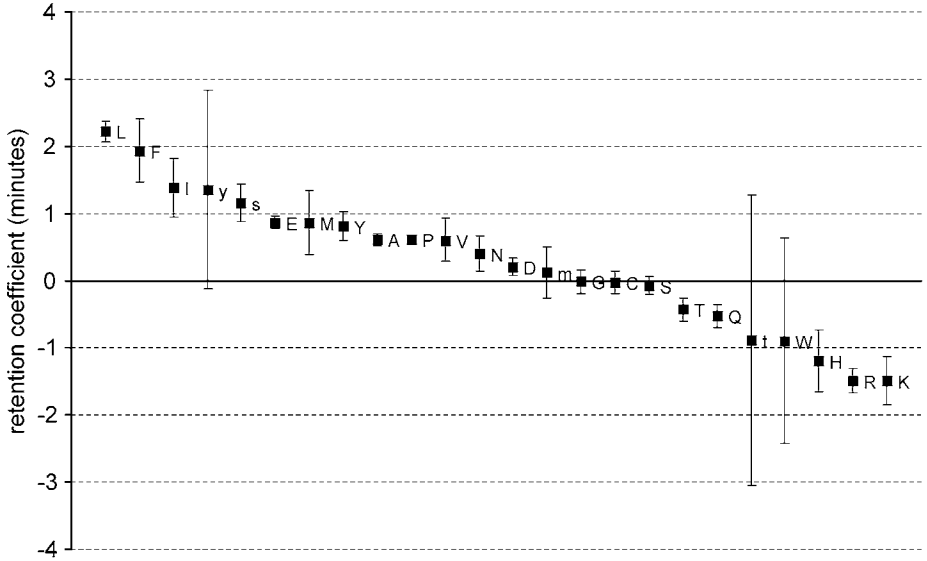


Fig. 6. Distribution of retention coefficients when using one-third of the nine combined liquid chromatography–tandem mass spectrometry datasets in the training set. One-letter amino acid codes: m, methionine-*S*-oxide; s, phosphoserine; t, phosphothreonin; y, phosphotyrosine.

4. Software

The source code for the retention time predictor (rt) is available under the GNU General Public License at http://yass.sdu.dk/RT/rt.htm, along with instructions for compilation and usage. In its simplest form, rt takes one argument on the command-line,

rt <*training set*>

where <*training set*> is a space delimited list of retention times in arbitrary units and peptide sequences in the one-letter code with lowercase letters for the modified amino acids and one measured retention time and peptide sequence pair per row in an ASCII text file, e.g.,

28.536 AHGHSmsDPAISY
32.14 SHLtWFCTMKLD
34.763 AGASyTDVAYK
etc.

The output from rt is a list of amino acids in the one-letter code and the corresponding retention coefficients c_i. The chromatographic retention time peptide[i].t

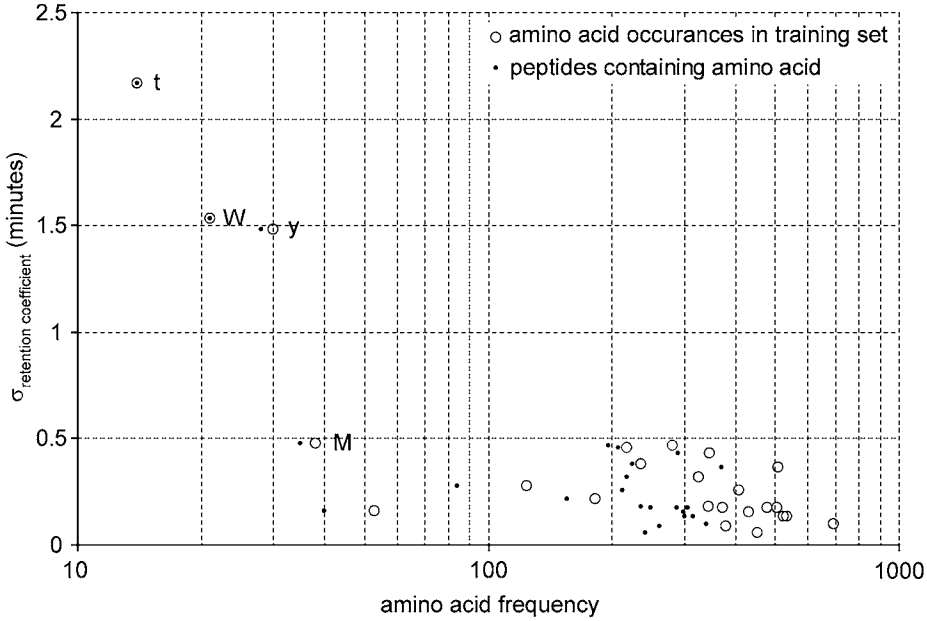


Fig. 7. Retention coefficients converges when at least approx 10 unique peptides containing the corresponding amino acid or posttranslational modification (amino acid frequency in the full dataset, one-third of which was used for training). For rare modifications, synthetic peptides with these modifications could be added to the sample.

of a candidate peptide peptide[i].sequence (as a text string in the one-letter code) is then predicted by applying **Eq. 2**:

```
peptide[i].t=c[24];
for(j=0;j<strlen(peptide[i].sequence);j++)
  {
   peptide[i].t+=c[(24–
   strlen(strchr("ARNDCEQGHILKMFPSTWYVmsty",peptide[i].
     sequence[j])))];
  }
```

where *c[]* is a vector of the retention coefficients c_i, "ARNDCEQGHILKM FPSTWYVmsty" a string defining the order of amino acids (in one-letter code) in *c[]* and *c[24]* is the constant term t_0 in **Eqs. 1** and **2**). Extensions to rt for testing the accuracy of the retention time prediction are also available from the author.

References

1. Klose, J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26,** 231–243.
2. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* **90,** 5011–5015.
3. Whitehouse, C. M., Dreyer, R. N., Yamashita, M., and Fenn, J. B. (1985) Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.* **57,** 675–679.
4. Stacey, C. C., Kruppa, G. H., Watson, C. H., et al. (1994) Reverse-phase liquid chromatography/electrospray-ionization Fourier-transform mass spectrometry in the analysis of peptides. *Rapid Commun. Mass Spectrom.* **8,** 513–516.
5. Voyksner, R. D. (1997) Combining liquid chromatography with electrospray mass spectrometry, in: *Electrospray Ionization Mass Spectrometry,* (Cole, R. B., ed.), John Wiley and Sons, New York, pp. 323–341.
6. Jensen, P. K., Pasa-Tolic, L., Peden, K. K., et al. (2000) Mass spectrometric detection for capillary isoelectric focusing separations of complex protein mixtures. *Electrophoresis* **21,** 1372–1380.
7. Smith, R. D., Pasa-Tolic, L., Lipton, M. S., et al. (2001) Rapid quantitative measurements of proteomes by Fourier transform ion cyclotron resonance mass spectrometry. *Electrophoresis* **22,** 1652–1668.
8. Shen, Y., Tolic, N., Zhao, R., et al. (2001) High-throughput proteomics using high-efficiency multiple-capillary liquid chromatography with on-line high-performance ESI FTICR mass spectrometry. *Anal. Chem.* **73,** 3011–3021.
9. Conrads, T. P., Alving, K., Veenstra, T. D., et al. (2001) Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling. *Anal. Chem.* **73,** 2132–2139.
10. Smith, R. D., Anderson, G. A., Lipton, M. S., et al. (2002) The use of accurate mass tags for high-throughput microbial proteomics. *Omics* **6,** 61–90.
11. Frenz, J., Hancock, W. S., Henzel, W. J., and Horváth, C. (1990) Reversed phase chromatography in analytical biotechnology of proteins, in *HPLC of Biological Macromolecules: Methods and Applications,* (Gooding, M. and Regnier, F. E., ed.), Marcel Dekker, New York, pp. 145–177.
12. Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C. (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **195,** 659–685.
13. Zubarev, R. A., Håkansson, P., and Sundqvist, B. U. R. (1996) Accuracy requirements for peptide characterization by monoisotopic mass measurements. *Anal. Chem.* **68,** 4060–4063.
14. Conrads, T. P., Anderson, G. A., Veenstra, T. D., Pasa-Tolic, L., and Smith, R. D. (2000) Utility of accurate mass tags for proteome-wide protein identification. *Anal. Chem.* **72,** 3349–3354.

15. Bruce, J. E., Anderson, G. A., Wen, J., Harkewicz, R., and Smith, R. D. (1999) High-mass-measurement accuracy and 100% sequence coverage of enzymatically digested bovine serum albumin from an ESI-FTICR mass spectrum. *Anal. Chem.* **71,** 2595–2599.
16. Palmblad, M., Ramström, M., Markides, K. E., Håkansson, P., and Bergquist, J. (2002) Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Anal. Chem.* **74,** 5826–5830.
17. Hodges, R. S., Parker, J. M., Mant, C. T., and Sharma, R. R. (1988) Computer simulation of high-performance liquid chromatographic separations of peptide and protein digests for development of size- exclusion, ion-exchange and reversed-phase chromatographic methods. *J. Chromatogr.* **458,** 147–167.
18. Hearn, M. T., Aguilar, M. I., Mant, C. T., and Hodges, R. S. (1988) High-performance liquid chromatography of amino acids, peptides and proteins. LXXXV. Evaluation of the use of hydrophobicity coefficients for the prediction of peptide elution profiles. *J. Chromatogr.* **438,** 197–210.
19. Mant, C. T., Zhou, N. E., and Hodges, R. S. (1989) Correlation of protein retention times in reversed-phase chromatography with polypeptide chain length and hydrophobicity. *J. Chromatogr.* **476,** 363–375.
20. Petritis, K., Kangas, L. J., Ferguson, P. L., et al. (2003) Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.* **75,** 1039–1048.
21. Strittmatter, E. F., Ferguson, P. L., Tang, K., and Smith, R. D. (2003) Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J. Am. Soc. Mass Spectrom.* **14,** 980–991.
22. Strittmatter, E. F., Kangas, L. J., Petritis, K., et al. (2004) Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *J. Proteome Res.* **3,** 760–769.
23. Qian, W. J., Liu, T., Monroe, M. E., et al. (2005) Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J. Proteome Res.* **4,** 53–62.
24. Nilsson, S., Ramstrom, M., Palmblad, M., Axelsson, O., and Bergquist, J. (2004) Explorative study of the protein composition of amniotic fluid by liquid chromatography electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *J. Proteome Res.* **3,** 884–889.
25. Palmblad, M., Ramstrom, M., Bailey, C. G., McCutchen-Maloney, S. L., Bergquist, J., and Zeller, L. C. (2004) Protein identification by liquid chromatography-mass spectrometry using retention time prediction. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **803,** 131–135.
26. The Cygwin homepage. http://www.cygwin.com. Last accessed 05/26/2006.
27. Matthiesen, R., Bunkenborg, J., Stensballe, A., Jensen, O. N., Welinder, K. G., and Bauw, G. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* **4,** 2583–2593.
28. Finney, G., Merrihew, G., Klammer, A., and MacCoss, M. (2004) Protein False Discovery Rates from MS/MS experiments: Decoy Databases and Normalized

Cross-Correlation. 52nd American Society for Mass Spectrometry conference on Mass Spectrometry, May 23–27, 2004 Nashville, TN.

29. Meek, J. L. (1980) Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. USA* **77,** 1632–1636.
30. Sanz-Nebot, V., Toro, I., Benavente, F., and Barbosa, J. (2002) pKa values of peptides in aqueous and aqueous-organic media. Prediction of chromatographic and electrophoretic behaviour. *J. Chromatogr. A* **942,** 145–156.
31. Rost, B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134,** 204–218.
32. Palmblad, M. (2002) *Identification and characterization of peptides and proteins using Fourier transform ion cyclotron resonance mass spectrometry.* PhD thesis, 05/17/2002, Uppsala Universitet, Uppsala, Sweden.

12

Quantitative Proteomics by Stable Isotope Labeling and Mass Spectrometry

Sheng Pan and Ruedi Aebersold

Summary

The goal of quantitative proteomics is to systematically study static state or perturbation-induced changes in protein profile. Most of the recently developed mass spectrometry (MS)-based quantitative proteomic methods employ stable isotope labeling to introduce signature mass tags to peptides/proteins that can be used by a mass spectrometer to quantify each analyte and to determine the sample from which it originates. In this chapter, we discuss several methods for the introduction of mass tags to proteins and peptides for MS-based quantitative proteomic analysis, including isotope-coded affinity tags, stable isotope labeling by amino acids in cell culture, global internal standard technology, and mass-coded abundance tagging.

Key Words: Quantitative proteomics; isotope-coded affinity tags; ICAT; stable isotope labeling by amino acids in cell culture; SILAC; global internal standard technology; GIST; mass-coded abundance tagging; MCAT.

1. Introduction

Quantitative proteomics aims to identify the proteins contained in complex samples such as cell and tissue extracts or subcellular fractions, and to determine the quantitative difference in abundance for each polypeptide contained in different samples. It is expected that the patterns resulting from such analyses will define comprehensive molecular signatures in health and disease and impact a wide range of biological and clinical research questions, such as the systematic study of biological processes and the discovery of clinical markers for detection, diagnosis, and assessment of treatment outcome ***(1–5)***. The traditional approach for quantitative protein profiling has been based on two-dimensional gel electrophoresis for protein separation, followed by mass

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

spectrometric identification of selected or all detected proteins ***(6,7)***. In the past few years, several mass spectrometry (MS)-based quantitative proteomic methods have been developed ***(1–3)***. Although they differ in sample preparation and other aspects, these methods share the use of stable isotope protein labeling or mass tagging to generate the mass signatures that identify the sample of origin and serve as the basis for accurate quantification. In general, the proteins contained in two different samples are labeled individually to acquire a different isotopic or chemical signature and then combined and analyzed by multidimensional chromatography and tandem mass spectrometry (MS/MS). Relative quantification of each identified protein in the samples compared is accomplished by determining the abundance ratio from the signal intensities of the differentially labeled peptides with identical sequence. These methods provide a broadly applicable means for quantitative proteome experiments, and most significantly, greatly improve the capability to analyze the proteins with low abundance.

Over the last few years, several methods for the incorporation of signature tags into proteins or peptides have been described (*see* **Fig. 1**). These include the use of chemical reactions to introduce an isotopic or chemical tag at specific functional groups on polypeptides ***(8–11)***, metabolic isotope labeling using heavy amino acids ***(12–16)***, and methods that introduce stable isotope tags via enzymatic reactions ***(17,18)***. Each one of these methods has specific strengths and weaknesses. Incorporation of stable isotopes or mass tags via chemical reactions allows great flexibility and selectivity for specifically tagging different reactive groups on proteins or peptides, including side chains of amino acids and specific types of modifications. Because the labeling reactions occur post isolation, specific functional groups and affinity tags can be introduced into essentially any sample to facilitate the selective isolation and analysis of a targeted subset of the proteome. Avoiding possible side reactions is important for the application of this method. For metabolic stable isotope labeling, proteins are labeled using cell culturing in media that are isotopically enriched or isotopically depleted. Because metabolic labeling does not require any chemical reactions, the method is easy to apply and particularly useful for experiments that involves cell culturing. It is, however, limited to cells that can be labeled in vitro and the range of available tags is limited to precursors of polypeptide synthesis that can be effectively metabolized. Complete isotopic labeling in cell culture can be difficult to achieve, even for those proteins with very long half-lives, as the growing cells synthesize new proteins during the cell culture ***(13)***. Stable isotope incorporation via enzyme reaction is straightforward and generates constant and uniform mass difference. However, the mass difference generated by enzymatic ^{18}O incorporation is either 4 or 2 Da, which can be difficult for quantitative analysis without using an algorithm for deconvolution of isotope patterns.

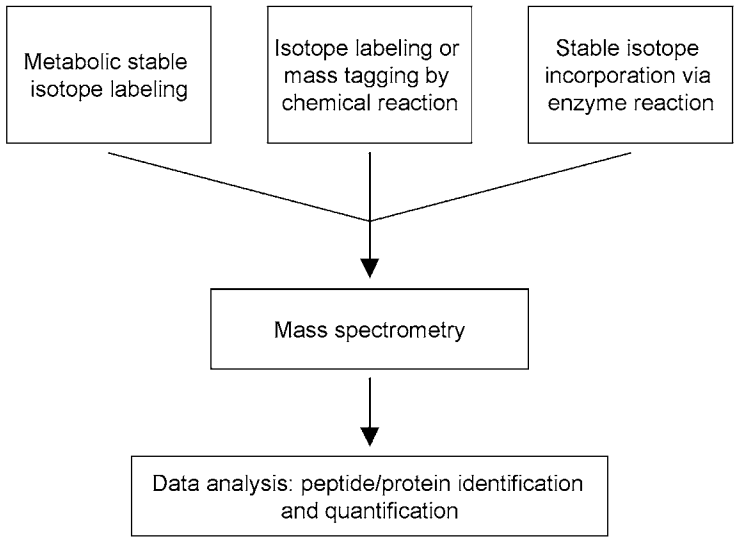


Fig. 1. Schematic illustration of the methods used to introduce stable isotope or mass tags to proteins or peptides for quantitative proteomics.

The most commonly used of these methods is the isotope-coded affinity tag (ICAT) approach ***(8,9)***, in which the proteins in two samples representing different proteomes are labeled separately on the side chain of their reduced cysteinyl residues using one of two chemically identical but isotopically different ICAT reagents. An important feature of this method is the incorporation of a biotin affinity tag into the ICAT reagents, which enables the selective isolation and purification of labeled analytes, thus affording a substantial reduction in sample complexity. Global internal standard technology (GIST) also utilizes chemical reactions to introduce stable isotope tags at specific sites in a polypeptide ***(10,19,20)***. Peptides from the samples to be compared are differentially derivatized on the primary amino groups and the quantification can be done at the MS level as well as MS/MS level using tandem mass spectrometer ***(10)***. The mass-coded abundance tagging (MCAT) method relies on labeling the proteins in the samples to be compared with compounds that are structurally related but different in mass. An example of this is the differential guanidination of C-terminal lysine residues on tryptic peptides ***(11)***. The stable isotope labeling with amino acids in cell culture (SILAC) method takes a different approach, and uses cell culturing to introduce the isotope labeling on proteins by adding ^{12}C- and ^{13}C-labeled amino acids to the growth media of separately cultured cell lines to differentially label the proteins in the growing cells ***(13–16)***. In this chapter, we provide

detailed protocols for the incorporation of signature tags into polypeptides via a number of well-established methods.

2. Materials

2.1. ICAT

1. Cleavable ICAT reagent kit from Applied Biosystems (Foster City, CA). The reagents and cartridges described in **Subheading 3.** are included in the kit.

2.2. SILAC

1. Base cell culture medium: essential medium depleted for certain amino acids (*see* **Note 1**) (Sigma, St. Louis, MO or Invitrogen, Carlsbad, CA) and supplemented with antibiotics and 10% dialyzed fetal bovine serum (*see* **Note 2**) (Invitrogen).
2. Normal (^{12}C-) and ^{13}C-labeled amino acids (*see* **Note 1**) (Cambridge Isotope Laboratories, Andover, MA).
3. Buffer: phosphate-buffered saline (PBS) (Sigma).
4. Lysis buffer: 1% sodium dodecyl sulfate (SDS), 1% Nonidet P-40, 50 m*M* Tris-HCl, pH 7.5, 150 m*M* NaCl, and protease inhibitor (Complete™ tablets; Roche Diagnostics, Mannheim, Germany).
5. Bradford protein assay kit (Pierce, Rockford, IL).
6. SDS-PAGE sample buffer (Invitrogen or Bio-Rad, Hercules, CA).
7. 10% SDS-polyacrylamide gel electrophoresis (PAGE) gel (Invitrogen or Bio-Rad).
8. Silver staining kit (Invitrogen or Bio-Rad).
9. Dithiothreitol (DTT) (Sigma) solution: 10 m*M* DTT and 100 m*M* ammonium bicarbonate.
10. Iodoacetamide (Sigma) solution: 55 m*M* iodoacetamide and 100 m*M* ammonium bicarbonate.
11. Trypsin (Promega, Madison, WI).
12. 20 m*M* ammonium bicarbonate solution.
13. 50 m*M* ammonium bicarbonate solution.
14. 100 m*M* ammonium bicarbonate solution.

2.3. GIST

1. Chemicals: *N*-acetoxysuccinimide, *N*-hydroxysuccinimide, hydroxylamine, tris(hydroxyl methyl)aminomethane (Tris base), tris(hydroxyl methyl) aminomethane hydrochloride (Tris acid), *N*-tosyl-L-phenylalanine chloromethyl ketone (TPCK) treated trypsin (Sigma), $[^2H_6]C_1$ acetic anhydride (Aldrich, Milwaukee, WI), D,L-DTT, 4-vinylpyridine (Bio-Rad), HPLC-grade trifluoroacetic acid (Pierce), HPLC-grade water, and acetonitrile (ACN) (Mallinckrodt Baker, Phillipsburg, NJ).
2. 10 m*M* DTT solution: 10 m*M* DTT and 100 m*M* ammonium bicarbonate.
3. 50 m*M* Tris solution: pH 8.0

4. Labeling buffer: 50 m*M* phosphate buffer, pH 7.5.
5. Sep-Pak cartridge (Waters, Milford, MA).

2.4. MCAT

1. Lysis buffer: 8 *M* urea, 1 m*M* $CaCl_2$, 100 m*M* Tris-HCl, pH 8.5.
2. Digestion buffer: 100 m*M* ammonium bicarbonate, pH 8.5, and 1 m*M* $CaCl_2$.
3. Immobilized TPCK trypsin beads (Pierce).
4. Solid *O*-methylisourea (*S*-methylisothiourea hemisulfate salt) (Sigma-Aldrich, St. Louis, MO).
5. Solid-phase SPEC-PLUS PTC18 cartridge (Ansys Diagnostics, Lake Forest, CA).

3. Methods

3.1. ICAT

1. 100 µg of protein from the test and control sample are dissolved individually using 80 µL of the denaturing buffer.
2. 2 µL of the reducing reagent is added to both the control and test sample. After vortexing and spinning, the samples are placed in a boiling water bath for 10 min.
3. The samples are mixed and spun for 1–2 min to cool. A 1-µL aliquot from each sample is removed and designated "before labeling" for process monitoring (*see* **Note 3**).
4. Cleavable ICAT reagent light and heavy, respectively, is dissolved in 20 µL ACN.
5. The control and test sample are transferred to the vial containing light and heavy ICAT reagent, respectively.
6. The samples are mixed with vortex, spun, then incubated for 2 h at 37°C. A 1-µL aliquot from each sample is removed and designated "after labeling" for process monitoring (*see* **Note 3**).
7. Trypsin is dissolved in 200 µL Milli-Q® water.
8. The labeled control and test samples are combined together in a single vial.
9. Trypsin solution is added to the combined sample. The sample is well mixed and incubated for 12–16 h at 37°C. A 1-µL aliquot from the sample is removed and designated "after digestion" for process monitoring (*see* **Note 3**).
10. The cation-exchange cartridge is assembled according to the manufacturer's (Applied Biosystems) instruction and conditioned with 2 mL cation exchange buffer-load.
11. The sample mixture is diluted with 2 mL of cation exchange buffer-load. The sample is well mixed and the pH is adjusted to between 2.5 and 3.3.
12. The diluted sample is slowly injected (~1 drop/s) onto the cation-exchange cartridge. The flow-through is collected into the original sample tube (*see* **Note 4**).
13. 1 mL of cation exchange buffer-load is injected to wash Tris(2-carboxyethyl) phosphine), SDS, and excess ICAT reagents from the cartridge. The flow-through is collected into the original sample tube (*see* **Note 4**).

14. 500 µL of the cation-exchange buffer elute is injected slowly (~1 drop/s) to elute the peptides. The eluted peptides are collected as a single fraction in a fresh 1.5-mL tube.
15. The cation-exchange cartridge is cleaned and stored according to the manufacturer's (Applied Biosystems) instruction.
16. The avidin cartridge is assembled according the manufacturer's (Applied Biosystems) instruction and conditioned with 2 mL of the affinity buffer elute followed by 2 mL of the affinity buffer load (*see* **Note 5**).
17. The peptide mixture collected after the cation-exchange step is neutralized with 500 µL of the affinity buffer-load and well mixed.
18. The neutralized sample is slowly (~1 drop/s) injected onto the avidin cartridge, followed by 500 µL of the affinity buffer-load. The flow-through is collected and kept until successful loading is confirmed.
19. The avidin cartridge is washed with 1 mL of the affinity buffer wash 1 to reduce the salt concentration.
20. 1 mL of the affinity buffer wash 2 is injected onto the cartridge to remove the nonspecifically bound peptides. The first 500 µL of the flow-through is collected and kept until that the success of sample loading is confirmed.
21. The avidin cartridge is washed with 1 mL of Milli-Q water.
22. 50 µL of the affinity buffer-elute is injected onto the cartridge slowly (~1 drop/s), and the eluent is discarded.
23. Another 750 µL of the affinity buffer-elute is injected onto the cartridge slowly (~1 drop/s) and the eluted peptides are collected.
24. The avidin cartridge is cleaned and stored according to the manufacturer's (Applied Biosystems) instruction.
25. The affinity-eluted peptide mixture is dried down by a centrifugal vacuum concentrator.
26. 100 µL of the combined cleaving reagent (95 µL of cleaving reagent A and 5 µL of cleaving reagent B) is added to the sample. The sample is incubated for 2 h at 37°C.
27. The sample is dried down using a centrifugal vacuum concentrator and resuspended in an appropriate solvent based on the mass spectrometric analysis used.
28. The sample is stored at –80°C until its analysis by MS.

3.2. SILAC

1. ^{12}C- or ^{13}C-labeled amino acid (*see* **Note 1**) are prepared as 250X stock solutions in the PBS, sterile filtered, and added to the base cell culture media for a final concentration of 52 mg/L.
2. Two populations of cells are plated into six dishes each and grown in cell culture medium containing normal (^{12}C-) or ^{13}C-labeled amino acid (*see* **Note 1**) in a humidified atmosphere with 5% CO_2 in air.
3. The cells are harvested when the labeled amino acids are completely incorporated in the proteins (*see* **Note 6**). The cells are washed twice with ice-cold PBS and then scraped in the lysis buffer. The lysate is sonicated for two cycles of 30 s each and centrifuged to pellet cellular debris.

4. The supernatant of the samples compared are collected separately. The protein concentrations are measured using Bradford protein assay (or bicinchoninic acid assay; Pierce Biotechnology).
5. Equal amount of protein from each sample (labeled and unlabeled) are combined and mixed in 1:1 ratio and boiled in the SDS-PAGE sample buffer.
6. The protein mixtures are resolved on a 10% SDS-PAGE gel and stained with silver stain (or Coomassie blue; Bio-Rad) to visualize the gel bands.
7. The gel bands are excised and cut into approx 1-mm^3 cubes, placed in a small Eppendorf tube, washed with 1 mL deionized water twice, then dehydrated in 100 µL ACN for approx 15 min. The ACN solution is removed and the gel is dried by Speed-Vac.
8. 10 m*M* DTT solution is added to the Eppendorf tube until the gel pieces are completely covered. The protein is reduced at 56°C for 1 h.
9. After the sample is cooled to room temperature, the DTT solution is removed and an equal volume of 55 m*M* iodoacetamide solution is added to the gel. The gel is then incubated for 45 min in dark place.
10. The gel pieces are washed with 100 µL of 100 m*M* ammonium bicarbonate solution, then soaked in ACN for 15 min. The step is repeated twice. The gel is completely dried by Speed-Vac.
11. A freshly prepared solution containing 12.5 ng/µL trypsin and 50 m*M* ammonium bicarbonate is added to the sample until the gel pieces are completely covered (*see* **Note 7**). The gel pieces are reswelled at 4°C for 45 min. The sample is then incubated overnight at 37°C for protein digestion.
12. The sample is centrifuged and the supernatant is collected.
13. 20 m*M* ammonium bicarbonate solution is added to the sample until the gel pieces are completely covered. The gel is soaked in the solution for 20 min at which time the sample is centrifuged and the supernatant is collected.
14. The peptides are further extracted by soaking the gel pieces in 5% formic acid/50% ACN for 30 min. The samples are centrifuged and the supernatant is collected. **Step 14** is repeated three times.
15. The supernatants collected from **steps 12** to **14** are combined, dried down by Speed-Vac, and resuspended in a solvent (e.g., 0.4% formic acid, 5% ACN solution) appropriate for mass spectrometric analysis.
16. The sample is stored at –80°C until its analysis by MS.

3.3. GIST

1. 8.0 g of *N*-hydroxysuccinimide in 21.4 g of $[_2H^6]C_1$ acetic anhydride is stirred for 15 h at room temperature.
2. White product crystals (*N*-acetoxy-$[^2H_3]$succinimide) are collected after removing the liquid phase by rotary evaporation.
3. The white crystalline residue is treated with hexane and dried in vacuum. The product's melting point is 133–134°C.
4. 20 mg of protein from test and control samples is denatured separately by adding urea at a concentration of 8 *M*.

5. The proteins are reduced using 10 m*M* DTT solution after 1 h incubation at 50°C.
6. Vinyl pyridine is added at a concentration of 25 m*M* for cysteine alkylation. The reaction is allowed to proceed at room temperature for 30 min.
7. The samples are diluted with the Tris solution to reach a final urea concentration of 1 *M*.
8. TPCK-treated trypsin is added to a ratio of 1:50 (w/w) and the samples are incubated for 24 h at 37°C.
9. The digested proteins from test and control samples are collected on a reversed-phase column and then eluted with 60% ACN containing 0.1% trifluoroacetic acid, separately. Each of the samples was then dried down and resuspended in the labeling buffer.
10. A 50-fold molar excess of *N*-acetoxysuccinimide and *N*-acetoxy-[$^2H_3$]succinimide is added to the test and control sample, respectively. The labeling reagent is added in small amounts over the course of the first hour and the reactions are carried out at room temperature over 4 h with constant mixing.
11. Equal amount of labeled test and control samples are combined and treated with excess hydroxylamine. The pH is adjusted to 11.0–12.0. The incubation of hydroxylamine is allowed to proceed for 10 min (*see* **Note 8**).
12. The derivatized peptide mixture is purified with a Sep-Pak cartridge, dried down by Speed-Vac, then resuspended in a solvent (e.g., 0.4% formic acid, 5% ACN solution) appropriate for mass spectrometric analysis.
13. The sample is stored at –80°C until its analysis by MS.

3.4. MCAT

1. The cell lysates from test and control samples are suspended in the lysis buffer separately and centrifuged.
2. The supernatants are collected and diluted to 2 *M* urea using the digesting buffer.
3. Appropriate amount of immobilized TPCK trypsin beads are added to the extracts from test and control samples separately, according to the manufacturer's (Pierce) instruction. The samples are incubated at 30°C for 2 d with tumbling.
4. Solid *O*-methylisourea is added to the test sample to a final concentration of 1.5 *M*, the pH is adjusted to 10.0 with NaOH, and the sample is incubated at 37°C overnight (*see* **Note 9**).
5. The peptides from test and control samples are combined together. The combined sample is purified by solid-phase SPEC-PLUS PTC18 cartridges according to the manufacturer's (Ansys Diagnostics) instruction, dried down, and buffer-exchanged into a solvent (e.g., 0.4% formic acid, 5% ACN solution) appropriate for mass spectrometric analysis.
6. The sample is stored at –80°C until its analysis by MS.

4. Notes

1. ^{13}C-labeled leucine, arginine, and lysine have been used for SILAC experiments. Other essential amino acids can be used for SILAC experiments as well.

The selection of the type of amino acid for SILAC experiments is based on their abundance and availability.

2. The use of dialyzed serum instead of normal serum is based on the consideration that dialyzed serum does not contain detectable amounts of free amino acids.
3. The process monitoring aliquots (before labeling, after labeling, and after digestion) are analyzed by SDS-PAGE gel using silver stain to visualize the gel bands. After labeling a distinctive shift of the banding patterns toward a higher molecular weight will be detected. The postdigestion sample should be essentially free of proteins bands.
4. The flow-through is kept until the success of the loading on the cation-exchange cartridge is confirmed.
5. Conditioning the avidin cartridge with elute buffer is required to free up low-affinity binding sites on the avidin cartridge.
6. The time needed for a complete incorporation of labeled amino acids in the proteins depends on the type of cell line and experimental conditions. For example, time course studies showed that for NIH 3T3 fibroblasts it took approximately five doublings to reach the full incorporation of labeled amino acids ***(13)***. It is also important to note that even those proteins with very long half-lives would still show approx 97% incorporation of the label in the case of full incorporation as the growing cells synthesize new proteins during the cell culturing ***(13)***.
7. Trypsin works best in the pH range between 8.0 and 8.5.
8. The hydroxylamine treatment is to hydrolyze esters that might have been formed during the acylation reaction.
9. The reaction should be performed in a fume hood.

Acknowledgment

The authors thank Drs. Johan Malmstrom and Wei Yan for critical review of the chapter.

References

1. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207.
2. Patterson, S. D. and Aebersold, R. H. (2003) Proteomics: the first decade and beyond. *Nat. Genet.* **33 (Suppl),** 311–323.
3. Aebersold, R. and Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chem. Rev.* **101,** 269–295.
4. Diamandis, E. P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell. Proteomics* **3,** 367–378.
5. Srinivas, P. R., Srivastava, S., Hanash, S., and Wright, G. L., Jr. (2001) Proteomics in early detection of cancer. *Clin. Chem.* **47,** 1901–1911.
6. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. USA* **97,** 9390–9395.

7. Rabilloud, T. (2002) Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* **2,** 3–10.
8. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999.
9. Zhou, H., Ranish, J. A., Watts, J. D., and Aebersold, R. (2002) Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nat. Biotechnol.* **20,** 512–515.
10. Chakraborty, A. and Regnier, F. E. (2002) Global internal standard technology for comparative proteomics. *J. Chromatogr. A* **949,** 173–184.
11. Cagney, G. and Emili, A. (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol.* **20,** 163–170.
12. Veenstra, T. D., Martinovic, S., Anderson, G. A., Pasa-Tolic, L., and Smith, R. D. (2000) Proteome analysis using selective incorporation of isotopically labeled amino acids. *J. Am. Soc. Mass Spectrom.* **11,** 78–82.
13. Ong, S. E., Blagoev, B., Kratchmarova, I., et al. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386.
14. Foster, L. J., De Hoog, C. L., and Mann, M. (2003) Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors. *Proc. Natl. Acad. Sci. USA* **100,** 5813–5818.
15. Blagoev, B., Kratchmarova, I., Ong, S. E., Nielsen, M., Foster, L. J., and Mann, M. (2003) A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21,** 315–318.
16. Ibarrola, N., Kalume, D. E., Gronborg, M., Iwahori, A., and Pandey, A. (2003) A proteomic approach for quantitation of phosphorylation using stable isotope labeling in cell culture. *Anal. Chem.* **75,** 6043–6049.
17. Mirgorodskaya, O. A., Kozmin, Y. P., Titov, M. I., Korner, R., Sonksen, C. P., and Roepstorff, P. (2000) Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using ^{18}O-labeled internal standards. *Rapid Commun. Mass Spectrom.* **14,** 1226–1232.
18. Yao, X., Freas, A., Ramirez, J., Demirev, P. A., and Fenselau, C. (2001) Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal. Chem.* **73,** 2836–2842.
19. Wang, S., Zhang, X., and Regnier, F. E. (2002) Quantitative proteomics strategy involving the selection of peptides containing both cysteine and histidine from tryptic digests of cell lysates. *J. Chromatogr. A* **949,** 153–162.
20. Ji, J., Chakraborty, A., Geng, M., et al. (2000) Strategy for qualitative and quantitative analysis in proteomics based on signature peptides. *J. Chromatogr. B Biomed. Sci. Appl.* **745,** 197–210.

13

Quantitative Proteomics for Two-Dimensional Gels Using Difference Gel Electrophoresis

David B. Friedman

Summary

Difference gel electrophoresis (DIGE) technology provides a powerful quantitative component to proteomics experiments involving two-dimensional (2D) gel electrophoresis. DIGE allows for the detection of subtle changes in protein abundance with statistical confidence while controlling for gel-to-gel variation, as well as additional variation that is non-biological in origin (e.g., sample preparation error, normal variation in a system). Samples are differentially labeled with spectrally resolvable fluorescent dyes (Cy2, Cy3, and Cy5) and co-resolved for direct quantification within the same 2D gel. Increased statistical confidence is obtained when combining experimental repetition with internal standards such that independent replicate measurements from single- and multivariable analyses can be intercompared with a relatively small number of coordinated DIGE gels.

Key Words: DIGE; difference gel electrophoresis; two-dimensional gel electrophoresis; quantification.

1. Introduction

The proteome is a dynamic entity, constantly changing both in levels of protein expression as well as in posttranslational modification, all of which are completely hidden in the static DNA code. One great challenge in proteomics has been in quantifying the dynamics of protein expression in any given state. Various strategies have been implemented to address the issue of protein abundance quantification. They usually involve either gel-based protein separations using two-dimensional gel electrophoresis (2D-GE), with subsequent protein identification using mass spectrometry (MS) on a subset of proteins of interest, or high-performance liquid chromatography (HPLC) separations of complex peptide mixtures coupled directly in-line with MS. Until recently, quantitative

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen

strategies have been limited to either comparison of total protein stains between separate 2D-GE separations, or by using differential stable isotope labeling of peptide mixtures and performing HPLC coupled with tandem MS experiments (liquid chromatography [LC]–MS/MS). Practical experience demonstrates that both gel-based and LC/MS-based approaches are complementary with some degree of overlap. In general terms, gel-based approaches are often used on a more global scale, whereas LC/MS-based strategies target subproteomes or defined protein mixtures (although there are many examples of the reverse, as well as strategies that take advantages of the complementary aspects of both approaches). 2D gel-based strategies separate intact proteins based on both charge (isoelectric point, pI) and mass, and therefore have the ability to resolve multiply charged isoforms (that may result from phosphorylation or other charged posttranslational modifications) and biologically significant proteolytic products. Subsequent MS can verify that a set of isoforms is, in fact, related without necessarily identifying the modified peptide(s), whereas such changes may be completely overlooked in peptide-based approaches without mass spectral information on the modified peptide(s).

2D-GE is capable of resolving several thousands protein spot features in a single separation ***(1)***, with detection limits of ca. 1 ng. It couples first dimension protein separation by charge (using isoelectric focusing, IEF) with second dimension protein separation by molecular weight (using sodium dodecyl sulfate-polyacrylamide gel electrophoresis [SDS-PAGE]) (*see* **Note 1**). Although single 2D-GE runs can resolve proteins with pI ranges between pH 3.0 and 11.0 and apparent molecular mass ranges between 10 and 200 kDa, higher resolution and sensitivity can be obtained by running a series of medium range (e.g., pH 4.0–7.0, 7.0–11.0) and narrow range (e.g., pH 5.0–6.0) IEF gradients with increasing protein loads, leading to an overall more comprehensive proteomic analysis ***(1,2)*** (*see* **Note 2**). This is analogous to gaining increased resolution and sensitivity in an LC/MS-based strategy by using multiple HPLC columns with different affinity chemistries (e.g., MuDPIT ***[3]***). Specific proteins of interest are then identified using standard MS approaches on gel-resolved proteins that have been excised and proteolyzed into a discrete set of peptides (*see* Chapter 1).

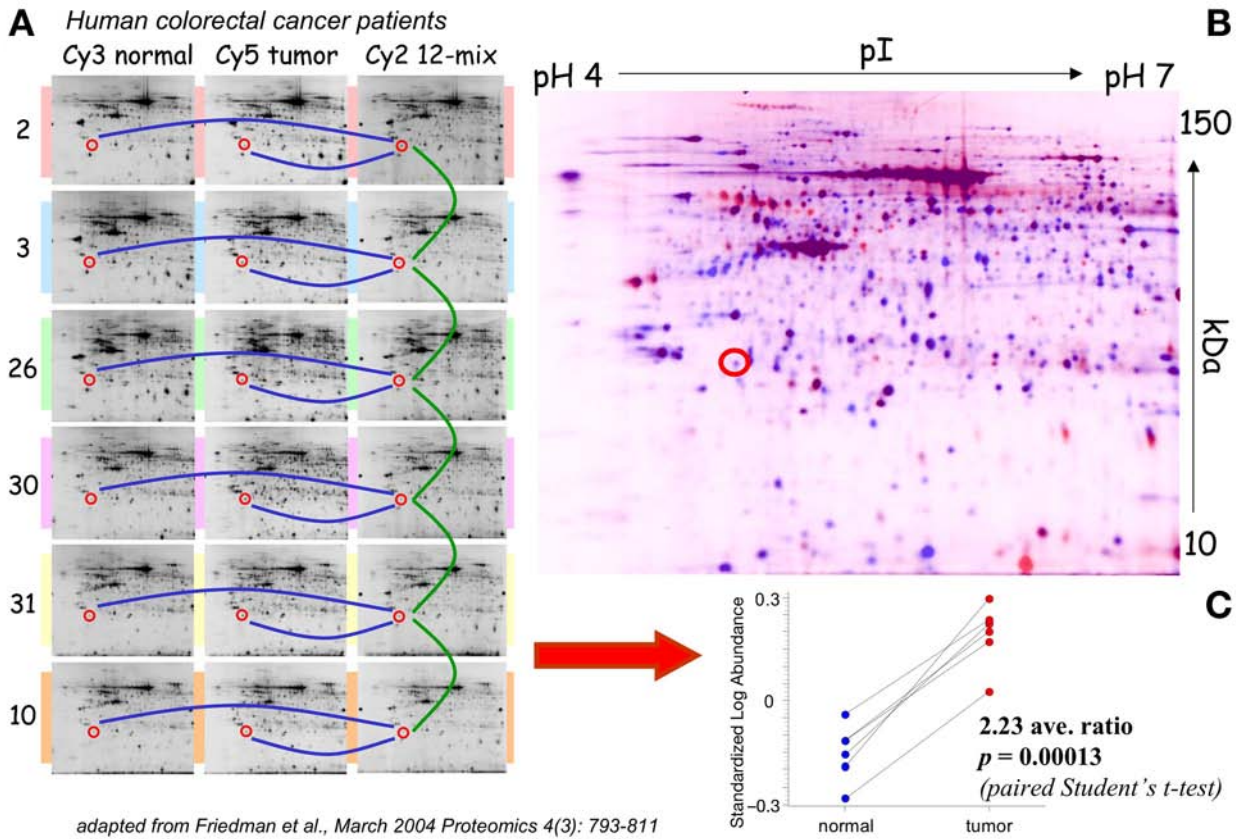
2D-GE has traditionally been a popular method for differential-display proteomics on a global scale, but until recently, these strategies lacked the ability to directly quantify abundance changes in the same fashion as in stable isotope LC/MS-based strategies ***(4–6)***. This has mainly been because of the inability to directly correlate migration patterns and protein staining between gel separations (gel-to-gel variation). Stable isotopes have been used in gel-based proteomics as well, whereby different proteomes have been separately labeled with different stable isotopes (e.g., growing cells using ^{14}N- vs ^{15}N-labeled medium) prior to mixing and running together through the same 2D-GE separation ***(7)***. In this case,

abundance changes can be monitored during the MS stage on individual proteins, but requires the in-gel digestion and MS on every protein present to discover the subset of proteins that is changing.

Difference gel electrophoresis (DIGE) technology was recently introduced (first described by Unlu et al. ***[8]***), which adds an essential quantitative component to 2D-GE and allows for the detection of subtle changes in protein abundance with statistical confidence ***(9–11)***. DIGE uses three spectrally resolvable fluorescent dyes (Cy2, Cy3, and Cy5) to label up to three samples to be run together on the same 2D gel.

The ability to coresolve multiple samples in a single gel is attractive because it allows for direct relative quantification for a given protein without any interference from gel-to-gel variation, removing the need for running replicate gels for each sample to control for variation (similar to stable isotope LC/MS-based strategies). This approach has limited statistical power, however, because confidence intervals are determined based on the overall variation within a population (*see* **Subheading 3.6.**). Many researchers new to DIGE technology are not immediately aware of the increased statistical advantage and multiplexing capabilities of DIGE when combining this approach with a pooled-sample mixture as an internal standard for a series of coordinated DIGE gels ***(12)***. This design will allow for repetitive measurements (vital to any type of experimental investigation), and in such a way as to control both for gel-to-gel variation and provide increased statistical confidence (using Student's *t*-test or ANOVA). In this way, statistical confidence can be measured for each individual protein based on the variance of repetitive measurements, independent of the variation in the population. Incorporating independently prepared replicate samples into the experimental design also controls for unexpected variation introduced into the samples during sample preparation.

This more complex and statistically powerful experimental design is accomplished by using one of the three dyes (usually Cy2) to label a grand mixture of all of the samples in an experiment to serve as an internal standard to be loaded onto every gel. Because this standard is composed of all of the samples in a coordinated experiment, each protein in a given sample should be represented in the standard and thus have its own unique internal standard. Direct quantitative comparisons are made individually for each resolved protein between the Cy3- or Cy5-labeled samples and the cognate protein signal from the Cy2-labeled standard for that gel (without interference from gel-to-gel variation). The individual signals from the internal standard are also used to normalize and compare between each in-gel direct quantitative comparison for that particular protein from the other gels. Using the Cy2-labeled standard in this fashion, therefore, allows for more precise and complex quantitative comparisons between gels, including sample repetition (**Fig. 1**).



adapted from Friedman et al., March 2004 Proteomics 4(3): 793-811

Importantly, the internal standard experimental design allows for the identification of significant changes that would not have been identified if the analyses were performed separately, even when using Cy3- and Cy5-labeled samples on the same DIGE gel ***(13)***. This experimental design also allows for multivariable analyses to be performed in one coordinated experiment, whereby statistically significant abundance changes can be quantitatively measured simultaneously between several sample types (e.g., different genotypes, drug treatments, disease states), with repetition and without the necessity for every pairwise comparison to be made within a single DIGE gel ***(14,15)*** (*see* **Note 3**).

This chapter assumes a solid understanding in 2D-GE, and will focus on the design and implementation of the DIGE method using the pooled-sample internal standard methodology. Sample preparation and the CyDye-labeling reactions using the minimal dye chemistry will also be discussed, with notes provided for the less common saturation-labeling chemistry.

2. Materials

2.1. Cell Lysis Buffers

1. TNE: 50 m*M* Tris-HCl, pH 7.6, 150 m*M* NaCl, 2 m*M* EDTA, pH 8.0, 2 m*M* dithiothreitol (DTT), and 1% (v/v) NP-40.
2. RIPA buffer: 50 m*M* Tris-HCl, pH 8.0, 150 m*M* NaCl, 1% NP-40, 0.5% deoxycholic acid, and 0.1% SDS.
3. 2D-GE sample buffer: 7 *M* urea, 2 *M* thiourea, 4% CHAPS, and 2 mg/mL DTT.

Fig. 1. *(Opposite page)* Schematic illustrating a difference gel electrophoresis (DIGE) analysis using a pooled-sample as an internal standard to coordinate samples across multiple DIGE gels. **(A)** Twelve individual samples are separately labeled with either Cy3 or Cy5, and then one Cy3/Cy5 pair of samples is mixed together along with an equal aliquot of a Cy2-labeled mixed sample internal standard to then be coresolved across six DIGE gels. The Cy2/3/5-specific spot maps are then independently scanned from each gel as shown (Cy2/3/5 images are grouped horizontally for each gel). Each resolved protein from a given Cy3- or Cy5-labeling has an unique internal standard represented in the Cy2-labeled mixture that is loaded equally onto every gel. For each resolved protein feature, direct quantitative measurements are made relative to the signal from the standard within the same gel separation, without interference from gel-to-gel variation. The standard is then used for matching and normalization between gels in a coordinated set. Gel numbers indicate individual human patient samples. **(B)** A representative two-dimensional spot map from patient number 30, with overlayed normal sample (Cy3-red) and the tumor sample (Cy5) for comparison. **(C)** Graphical representation of the normalized log abundances and average ratio between normal and tumor for the protein indicated with the red circle, with associated Student's *t*-test *p* value. This protein was subsequently identified by mass spectrometry as translationally controlled tumor protein P13693. Adapted from **ref.** ***13***.

2.2. SDS-PAGE

1. Immobilized pH gradient strips and accompanying ampholyte mixtures can be purchased from a number of commercial vendors. Strip lengths vary from 7 cm to high-resolution 24-cm strips, and pH ranges vary from wide-range (e.g., pH 3.0–11.0) to high-resolution narrow-range (e.g., pH 5.0–6.0) strips.
2. 50 mL bind silane working solution: 40 mL ethanol, 1 mL acetic acid, 50 µL bind silane solution (Amersham Biosciences), and 9 mL water (*see* **Note 4**).
3. 4X separating gel buffer: 1.5 *M* Tris-base, pH 8.8.
4. 30% Acrylamide:*bis*-acrylamide (37.5:1): *N*,*N*,*N*,*N*'-tetramethylethylenediamine and ammonium persulfate.
5. 1 L 10X SDS-PAGE running buffer: 30.25 g Tris base, 144.13 g glycine, and 10 g 0.1% SDS.
6. 1 L fixing solution for SyproRuby staining: 100 mL methanol, 70 mL acetic acid, 830 mL water. SyproRuby (Invitrogen/Molecular Probes), Deep Purple (GE Healthcare/Amersham Biosciences), and other total-protein stains are available from several commercial sources.
7. Second dimensional equilibration buffer: 6 *M* urea, 50 m*M* Tris-base, 30% glycerol, 2% SDS, and trace bromophenol blue.
8. Water-saturated butanol (*see* **Note 5**).
9. DTT (store dessicated).
10. Iodoacetamide (store dessicated, keep in the dark).

2.3. DIGE-Labeling Materials

1. *N*,*N*-dimethyl formamide (DMF) (*see* **Note 6**).
2. Labeling (L) buffer: 7 *M* urea, 2 *M* thiourea, 4% CHAPS, 30 m*M* Tris-base, and 5 m*M* magnesium acetate (*see* **Note 7**).
3. Rehydration (R) buffer: 7 *M* urea, 2 *M* thiourea, 4% CHAPS, 2 mg/mL 13 m*M* DTT (2%).
4. Cyanine dyes with *N*-hydroxy succinimidyl (NHS) ester chemistry for minimal labeling (Cy2, Cy3, Cy5), and with maleimide chemistry for saturation labeling (Cy3 and Cy5) are available from GE Healthcare/Amersham Biosciences as dry solids.
5. 10 m*M* lysine quenching solution.
6. 200 mg/mL DTT reduction stock solution.

3. Methods

DIGE is a powerful technique for quantitative multivariable differential-display proteomics. However, the quality of the data will only be as good as the quality of the underlying 2D-GE technology on which it is based. The main focus of this chapter is to provide detailed notes on the DIGE technology, however, some key considerations to successful high-resolution 2D-GE are also provided.

3.1. Sample Preparation

The key to success for any analytical measurement begins with sample preparation. This not only includes the buffers and materials used, but also the nature of the samples and the way in which they are procured. The addition of exogenous materials or allowing for uncontrolled manipulation of the sample (such as conditions that may lead to proteolysis) can severely hamper and sometimes completely prevent an analysis. Care should be taken to ensure against common laboratory contaminants (e.g., mycoplasma for tissue culture) that, if present, may be detected as significant changes using DIGE, either resulting from the presence in a subset of samples, or by responding to the experimental perturbation.

1. Protein extracts can be made essentially using any method of preference, as the necessary amount of protein can be subsequently precipitated prior to resuspension in the CyDye-labeling buffer (*see* **Subheading 3.2.**). Ensuring against proteolysis and loss of posttranslational modifications (e.g., phosphorylation) is of monumental importance. However, care should be taken not to use reagents that will resolve on the 2D gel, such as soybean trypsin inhibitor. Small molecule inhibitors, such as aprotinin, leupeptin, pepstatinA, antipain, AEBSF, sodium orthovanadate, okadaic acid, and microcystin, among others, are far better choices.
2. There are myriad ways to extract proteins from biological systems, and a given protocol may work better for certain samples. Standard lysis buffers such as TNE, RIPA buffers, or even the buffers used for 2D-GE (*see* **Subheading 2.1.**) all have the capability of producing high-resolution samples for 2D-GE.
3. Sonication can often prove beneficial in creating high-resolution samples for 2D-GE, most likely because of the disruption of nucleic acids, which are subsequently removed by MeOH/$CHCl_3$ precipitation (*see* **Subheading 3.2.**) along with phospholipids. Both of these nonproteinacious ionic components can obliterate the resolution during IEF. Short bursts with a tip sonicator is suggested, especially in the presence of urea-containing samples that should never be heated (*see* **Note 7**).
4. Protein concentrations should be determined using a system that is compatible for the buffer that the proteins are extracted in. For example, the CHAPS and thiourea in the buffers used for DIGE, although adequately chaotropic, interfere with either the Bradford or bieinchoninic acid (BCA) assays, making the data inaccurate and unreliable. In these cases, aliquots should be precipitated prior to quantification in a suitable buffer.
5. A good working range to aim for is between 1 and 10 mg/mL. Too dilute and it will be difficult to quantitatively recover proteins following precipitation cleanup (*see* **Subheading 3.2.**); too concentrated and it will be difficult to accurately dispense the appropriate volume for the experiment. Freeze-thawing should also be kept to a minimum; freezing samples in 1-mL aliquots or less will usually suffice.

3.2. Sample Cleanup

The desired amount of sample to be used in the experiment should be precipitated prior to labeling. This both removes nonproteinacious ions from the

Table 1
Experimental Design for CyDye Labeling (Minimal Dyes) Using a Pooled-Sample Internal Standard

	Samples						
	Gel 1		Gel 2		Gel 3		
	Control-1	Treated-1	Control-2	Treated-2	Control-3	Treated-3	Pool
Precipitated amount	150 µg	150 µg	150 µg	150 µg	150 µg		
L-buffer	24 µL	24 µL	24 µL	24 µL	24 µL	24 µL	
Aliquot	16 µL	16 µL	16 µL	16 µL	16 µL	16 µL	8 µL (×6)
Cy2 (100 pmol/µL)							6 µL
Cy3 (100 pmol/µL)	2 µL		2 µL			2 µL	
Cy5 (100 pmol/µL)		2 µL		2 µL	2 µL		
	30 min on ice in the dark						
Lysine (quench)	2 µL	2 µL	2 µL	2 µL	2 µL	2 µL	6 µL
	10 min on ice in the dark						
Total vol	20 µL	20 µL	20 µL	20 µL	20 µL	20 µL	60 µL
	For each gel, combine the quenched Cy3- and Cy5-labeled samples and add one-third of the quenched Cy2-labeled pooled mixture						
	20 + 20 + 20 µL		20 + 20 + 20 µL		20 + 20 + 20 µL		
2X R-buffer	60 µL		60 µL		60 µL		
Total	120 µL		120 µL		120 µL		
R-buffer	to V_f		to V_f		to V_f		

This table illustrates a typical DIGE-labeling experiment, as described in **Subbeadings 3.2.** and **3.3.**

sample (e.g., nucleic acids, phospholipids) that can interfere with IEF, as well as transfers the proteins into a labeling buffer optimized for CyDye labeling and subsequent IEF. Determine how much total protein will be on each gel, and precipitate one-half of that amount for each sample to be run on that gel. This is straight-forward for a two-component separation, but also works out for the multigel experiments where one-third of the total protein amount on each gel comes from the pooled-sample internal standard (**Table 1**). Precipitate only what is needed for each sample for the experiment; too much material may create pellets that are difficult to resolubilize completely.

Many precipitation methods are available, the following is a MeOH/$CHCl_3$ protocol that works well for DIGE, and can be easily performed in 1.5-mL tubes (adapted from Wessel and Flugge ***[16]***):

1. Bring up a predetermined amount of protein extract to 100 µL with water.
2. Add 300 µL (3 vol) water.
3. Add 400 µL (4 vol) methanol.

4. Add 100 µL (1 vol) chloroform.
5. Vortex vigorously and centrifuge (10 min, 13*g*); the protein precipitate should appear at the interface.
6. Remove the water/MeOH mix on top of the interface, being careful not to disturb the interface. Often the precipitated proteins do not make a visibly white interface, and care should be taken not to disturb the interface.
7. Add another 400 µL methanol to wash the precipitate.
8. Vortex vigorously and centrifuge; the protein precipitate should now pellet to the bottom of the tube.
9. Remove the supernatant and briefly dry the pellets in a vacuum centrifuge.
10. Resuspend the pellets in a suitable amount of CyDye-labeling buffer (L-buffer, *see* **Table 1**). Maintaining the pH at 8.5 is critical to efficient CyDye labeling, so it is beneficial to check the pH of the resuspended samples with pH paper prior to dye labeling.

3.3. DIGE Experimental Design

1. Preliminary gel. All experiments should start with a preliminary gel on representative samples to ensure equivocal protein amounts between samples, and that the highest resolution and sensitivity are obtained before embarking on a multigel DIGE experiment (*see* **Notes 8** and **9**). The preliminary gel will also show any problems with the sample preparation that may be corrected by adjusting the procurement methods (*see* **Subheading 3.1.**). This step can also be used to optimize the maximal amount of protein that can be loaded without adversely affecting resolution.
   The preliminary gel need only to test one or two of the samples of a much larger experiment. This gel can simply be stained with a total protein stain (e.g., Sypro Ruby or Deep Purple) to visually inspect the resolution and sensitivity. Alternatively, the gel can contain two or three different samples prelabeled with Cy dyes and coresolved (*see* **Note 10**).
2. Choice of pH gradient. Precast IEF strips are commercially available from several vendors. The widest length is currently 24 cm, providing the highest resolving power for a given pH range. Medium-range IEF focusing gradients (e.g., pH 4.0–7.0) offer the best trade-off between overall resolution and sensitivity. Subsequent experiments can then be designed to resolve proteins in the basic range (pH 7.0–11.0) and in narrow pI ranges with commensurate increases in protein loading to gain access to the lower abundant proteins in a given sample (*see* **Note 2**). In this way, a more comprehensive picture of the proteomes under study can be obtained.
3. Incorporation of a pooled-sample mixture internal standard on every DIGE gel in a coordinated experiment. This internal standard, usually labeled with Cy2, is composed of an equal aliquot of *every* sample in the entire experiment, and therefore represents every protein present across all samples in an experiment. The use of this pooled-sample internal standard on every DIGE gel in a coordinated

experiment allows for the facile comparison of independent sample replicates with increased statistical confidence. This experimental design also enables the simultaneous quantitative comparison between multiple variables in a coordinated experiment.

4. Plan out which samples will be labeled with which dyes ahead of time. For minimal dye-labeling chemistry (*see* **Subheading 3.4.**), each gel will contain two individual samples labeled with either Cy3 or Cy5, and an equal amount of the pooled-sample internal standard. The example outlined in **Table 1** is for a two-component comparison repeated in triplicate, with 300 µg total protein loaded onto each of three gels. In this case, 150 µg of each sample should be precipitated (**Subheading 3.2.**), resuspended in L-buffer, and then split 2:1. Two-thirds of each sample (100 µg) will be individually labeled with either Cy3 or Cy5. The remaining one-third of each sample will be pooled together and labeled with Cy2 to serve as an internal standard. By following this, there will be enough of the Cy2-labeled internal standard to have an equal amount as the Cy3 or Cy5 samples loaded onto each gel (*see* **Note 11**).

3.4. CyDye Labeling

The most established DIGE chemistry is the "minimal labeling" method, which has been commercially available since July 2002. CyDye DIGE fluors are supplied as an NHS ester, which reacts with primary amino groups, typically the ε-amine group of lysine side chains. The three fluors are mass matched (*ca*. 500 Da), and carry an intrinsic +1 charge to compensate for the loss of each proton-accepting site that becomes labeled (thereby maintaining the pI of the labeled protein). Each dye molecule also adds a hydrophobic component to the proteins, which along with molecular weight, influences how proteins migrate in SDS-PAGE (yielding apparent molecular mass separations).

However, some proteins may not remain soluble under 2D-GE conditions if too much additional hydrophobicity is introduced, and lysines are quite prevalent in protein sequences. Thus, minimal labeling reactions are optimized for labeling that only 2–5% of the total number of lysine residues are labeled, such that on average, a given labeled protein would contain only one dye molecule. This also allows for direct comparison of fluorescent signals between multiple protein samples labeled in the same fashion. Labeling with CyDye DIGE fluors is very sensitive, comparable to silver staining or Sypro Ruby (ca. 1 ng) using the minimal labeling stochiometry, but with a linear response in protein concentration over five orders of magnitude ***(9)*** (*see* **Notes 12–14**).

All steps are performed on ice. The following protocol is for sample loading via rehydration of immobilized pH gradient (IPG) strips, and assumes incorporation of a pooled-sample internal standard to coordinate many samples across multiple DIGE gels simultaneously. The steps are summarized in **Table 1** (*see* **Note 15**).

1. Resuspend precipitated sample in 24 µL labeling (L) buffer. Remove 8 µL (one-third of sample) and place into a new tube that will contain the pooled-sample internal standard (8 µL from all of the other individual samples will be pooled into this tube) (*see* **Note 16**).
2. CyDyes are purchased as dry solids and should be reconstituted to 10X stock solutions (1 nmol/µL) in fresh DMF. Dilute stock solutions of CyDyes 1:10 in fresh DMF to a final working concentration of 100 pmol/µL (*see* **Note 6**).
3. Label each sample (50–250 µg) with 2 µL (200 pmol) of either Cy3 or Cy5 working dilution for 30 min on ice in the dark. Label the pooled-sample mixture with 2 µL (200 pmol) of Cy2 working dilution for every equivalent amount of sample present in the pooled standard as compared with the individually labeled samples. That is, if 100 µg of each sample is labeled with 200 pmol of Cy3 or Cy5, then 50 µg of each of these samples is present in the pooled standard, and 200 pmol of Cy2 is used for every 100 µg of pooled standard (*see* **Table 1** and **Note 17**).
4. Quench reactions with 2 µL of 10 m*M* lysine for 10 min on ice in the dark.
5. For each gel, combine the quenched Cy3- and Cy5-labeled samples and add one-third of the quenched Cy2-labeled pooled mixture (*see* **Note 18**).
6. To each tripartite mixture, add an equal volume of 2X R-buffer and incubate on ice for 10 min. 2X R-buffer is R-buffer supplemented with an additional 2 mg/mL DTT using the 200 mg/mL DTT stock solution. DTT is omitted from the L-buffer to prevent unfavorable interaction with the CyDyes. Adding an equal volume of 2X R-buffer to the quenched reactions provides the reducing agents to the total reaction volume at a 1X final concentration.
7. Add R-buffer (1X DTT concentration) to a final volume suggested by the manufacturer for the given IPG strip length (e.g., 450 µL for 24-cm strips). Add the appropriate volume of IPG buffer ampholines to 0.5% final (v/v) for IEF. Proceed with rehydration of dehydrated IPG strips for >16 h and proceed with IEF (*see* **Subheading 3.5.3.**).

3.5. 2D-GE and Poststaining

As a result of the minimal labeling, quantification with the CyDyes is carried out on only 2–5% of the proteins and this labeled portion of the protein may migrate at a higher apparent molecular mass than the majority of the unlabeled protein because of the added mass and hydrophobicity of the dyes (exacerbated in lower M_r species). To ensure that the maximum amount of protein is excised for subsequent in-gel digestion and MS, minimally labeled 2D DIGE gels are poststained with a total protein stain such as SyproRuby or Deep Purple. Accurate excision is also ensured by preferentially affixing the 2D gel to a presilanized glass plate during gel casting so that the gel dimensions do not change during the analysis (*see* **Notes 19** and **20**).

These methods assume the use of an Amersham Biosciences 2D electrophoresis system, but is easily adaptable to other commercially available systems. It also assumes usage of high-resolution, 24×20 cm gels.

1. Special gels for 2D SDS-PAGE. Using low-fluorescence glass plates, pretreat one plate for each gel with 3–5 mL bind silane working solution, carefully wiping the entire surface of the plate with a lint-free wipe. Leave treated plates covered with lint-free wipes for several hours to allow for sufficient out-gassing of fumes (that may contain bind silane) before assembling gel plates and casting of the 2D SDS-PAGE gels (*see* **Note 21**).
2. Assemble plates and pour 12% homogeneous SDS-PAGE gel(s) using the appropriate amount of 30% stock acrylamide and 4X separating gel buffer for the volumes needed for the number of gels being poured (*see* **Note 22**). Overlay the gels with water-saturated butanol for several hours to provide a straight and level surface to place the focused IPG strip (*see* **Note 5**).
3. The combined tripartite-labeled samples, brought up to final volume with 1X R-buffer and passively rehydrate IPG strips for longer than 16 h (**Subheading 3.4.7.**), are subjected to IEF (several manufactures available) (*see* **Note 23**).
4. The focused IPG strips are next equilibrated into the 2D equilibration buffer. During this step, the cysteine sulfhydryls in the focused proteins are reduced and carbamidomethylated by supplimenting the equilibration buffer with 1% DTT for 20 min at room temperature, followed by 2.5% iodoacetamide in fresh equilibration buffer for an additional 20-min room temperature incubation (*see* **Note 24**).
5. Place equilibrated IPG strip on top of the SDS-PAGE gels that were precast with low-fluorescence glass plates. Use a thin card or ruler to carefully tamp down the IPG strip to the SDS-PAGE gel, removing air bubbles at the interface (*see* **Notes 25** and **26**).
6. Perform 2D SDS-PAGE at constant wattage, using << 1 W/gel for at least 1 h prior to ramping up to lower than 20 W/gel (*see* **Note 27**).
7. CyDye images are acquired using a fluorescence imager, such as the Typhoon 9400 series (GE Healthcare/Amersham Biosciences) equipped with lasers and filters that are compatible with the emission/excitation spectra of the dyes. Imaging is performed through the glass plates using the intact gel cassette (*see* **Note 28**).
8. After imaging the gels, carefully remove the plate that was untreated with bind silane. The gel will remain affixed to the treated plate, and can be stained with SyproRuby "open-faced" in the fixation/staining solutions a total protein stain recommended by the various manufactures. Acquire images using a fluorescence imager (*see* **Note 29**).

3.6. DIGE Analysis

3.6.1. Software Algorithms

Many bioinformatics tools are commercially available for the comparison of multiple 2D gel-separated protein spot patterns (*see* **Note 30**). Some free internet-based utilities (e.g., www.lecb.ncifcrf.gov/flicker/) provide simple alternation between two-spot patterns, whereas most of the commercial products contain

Table 2
Statistical Applications of DeCyder Biological Variation Analysis Module

Average ratio	Calculated for each protein-spot feature between two groups or experimental conditions. Derived from the log standardized protein abundance changes that were directly quantified within each DIGE gel relative to the internal standard for the protein-spot feature.
Student's *t*-test	This common test is used to calculate the statistical significance of an abundance change between two groups or experimental conditions. The null hypothesis being tested is that there is no significant change in the protein abundance between experimental groups. *P* values reflect the probability that the observed change has occurred from stochastic chance alone. With DIGE, *p* values of <0.01 are often observed. Assumes normal distributions of protein abundance. Can be performed either unpaired or paired.
One-way ANOVA	Tests for differences in standardized abundance across all groups of a multicomponent analysis. Indicates that one group is significantly different from another in the group. For two-group comparisons, this test will generate the same values as the *t*-test.
Two-way ANOVA	Calculates the significance of the difference between multiple groups with the same condition, where multiple conditions are analyzed. For example, two drugs (condition 1) at three time-points (condition 2), each with four independent biological replicates.

proprietary algorithms for protein-spot detection, intergel matching, protein-spot quantification, and even utilities for building web-based tools for data dissemination. Many include the ability to average replicate patterns into a single virtual pattern to be used in a comparative study. They are all designed to compare multiple spot patterns and quantify abundance changes for individual proteins between experimental conditions.

The DeCyder suite of software tools (GE Healthcare/Amersham Biosciences) were specifically developed to support the DIGE platform, especially for those experiments that incorporate the internal standard approach, and is therefore used as an example here. The differential in-gel analysis (DIA) module of DeCyder is used for direct quantification of protein-spot volume ratios between the triply codetected signals emanating from each resolved protein, and can be used for the simplest form of a DIGE experiment for pairwise comparisons with N = 1. The more advanced DIGE experiments that use the internal standard to cross-compare replicate samples from pairwise and multivariable analyses (N>3) are handled by the biological variation analysis (BVA) module of DeCyder. In a BVA experiment, the signals emanating from the internal standard are used

both for direct quantification within each DIGE gel in a coordinated set (using DIA module), as well as for normalization and protein-spot pattern matching between gels (*see* **Note 31**). This allows for the calculation of Student's *t*-test and ANOVA statistics for individual abundance changes (*see* **Subheading 3.6.2.**; **Table 2**). BVA is also used to match patterns between SyproRuby- and CyDye-stained images to facilitate protein excision for subsequent MS (*see* **Notes 19**, **20**, and **29**).

3.6.2. Experimental Design and Statistical Confidence

In the simplest form of a DIGE experiment, two or three samples are separately labeled with one of the three dyes and separated in the same gel for direct pairwise comparisons. In this case, the software first normalizes the entire signal for each Cy-dye channel and then calculates the protein-spot volume ratio for each protein pair. A normal distribution is modeled over the actual distribution of protein pair volume ratios, and two standard deviations of the mean of this normal distribution represent the 95% confidence level for significant abundance changes.

This N = 1 type of experiment has only limited statistical power, because the 95th percentile confidence interval is determined based on the overall distribution of changes within the population (two standard deviations of the mean approximates the 95th percentile confidence interval; *see* **Note 32**). Many more changes in abundance of much lesser magnitude can be detected with much greater statistical confidence (Student's *t*-test and ANOVA; **Table 2**) by incorporating independent replicate samples into the experiment (*see* **Note 33**). The number of replicates required in a study depends on the amount of variation in the system being investigated. Increasing the number of replicates will increase confidence in smaller changes in expression. Subsequent classification and hierarchical clustering analysis can also be performed on the results ***(17,18)***, which are beneficial in extending DIGE applications for diagnostic and prognostic uses (now available in DeCycler v6.5).

With replicate samples, the Student's *t*-test and ANOVA statistics are measuring the significance of the variation of a specific protein change, independent of the overall distribution of abundance changes in the population. Incorporating replicate samples into the experimental design also controls for unexpected variation introduced into the samples during sample preparation. This design not only allows for the identification of abundance changes that are consistent across multiple replicates of an experiment, but can also identify significant abundance changes that would not have been identified even if the analyses were performed using Cy3- and Cy5-labeled samples on the same gels, but without the pooled-sample internal standard to coordinate them ***(13)***.

4. Notes

1. Both hydrophobicity and molecular weight influence how proteins migrate during SDS-PAGE, yielding information on apparent molecular mass.
2. The use of hydroxyethyl disulfide (commercially available as "DeStreak reagent"), combined with anodic cup loading, should be used for enhanced resolution for IEF greater than pH 8.0 ***(19)***.
3. Repetition not only enables the identification of subtle differences with statistical confidence, it is also vital to control for nonbiological variation. Thus, it is important that each replicate sample is derived from an independent experiment, ideally performed on different occasions as perhaps using different media. The independent samples can then be analyzed coordinately using the pooled-sample internal standard methodology. *See* **Table 1** for an example of this design.
4. All solutions should be prepared using water that has a resistivity of 18.2 MΩ-cm; this is referred to as "water" throughout the text.
5. Mix equal parts butanol and water and shake vigorously. Let the two phases separate overnight, and use the butanol phase for overlay. Butanol that is not completely water saturated can extract water from the top of the gel. A 0.01% SDS solution can also be used if carefully overlayed or sprayed as a fine mist, but the gel/overlay interface will not be as obvious.
6. DMF can degrade, producing amines that can react with the NHS ester CyDyes. DMF stocks should be kept fresh (<3 mo) and anhydrous to ensure optimal labeling conditions.
7. For buffers that contain urea, care should be taken to ensure the urea is fresh and free of the natural breakdown product isocyanate, which will add ionic content and carbamylate free amines and thereby neutralize the protonatable ε-amine groups of lysines residues. This is problematic for several reasons, the foremost being the fact that this gives rise to artificially charged train isoforms in the first dimension IEF. Heating samples above 37°C should also be avoided, as this facilitates the conversion to isocyanate.
8. For example, 500 µg of material may be loaded onto a pH 4.0–7.0 gel, but because of the overall distribution of proteins in the sample, as well as a sometimes unusually high abundance of a subset of proteins, it may result in much less material actually resolving between the electrodes. A good rule to follow is to load the desired amount based on the protein concentrations, and then adjust the load by eye as necessary.
9. Running every DIGE gel with the maximal amount of protein (without adversely effecting first-dimension resolution) not only enables detection of lower abundance proteins, but also provides more material for subsequent protein identification using MS. This makes every gel in a coordinated DIGE experiment a "pick-able" gel, without the need to run subsequent preparative gels with increased protein load that then have to be carefully matched to a lower-abundant, analytical gel. When combined with narrow-range IEF, maximizing the protein amount also allows interrogation of the lower abundant proteins in a sample.

10. This is DIGE in its most simplistic form, and can show differences between the samples without interference from gel-to-gel variation, but provides limited statistical power to help distinguish true biological variation from background, such as artificial noise introduced during sample preparation.
11. Employing a dye-swapping approach will control for any dye-specific effects that may result from preferential labeling or different fluorescence characteristics of acrylamide at the different wavelengths of excitation for Cy2, Cy3, and Cy5, especially at low protein-spot volumes. This is easily incorporated into any DIGE analysis where repetitive samples are used (along with the internal standard to compare across multiple DIGE gels). The fluorescent properties of Cy3 and Cy5 are more similar than either is to Cy2, which is why Cy2 should be chosen for the internal standard to minimize dye bias for individual samples.
12. Saturation labeling. A second labeling chemistry became available in 2003, which uses a thiol-reactive maleimide group to label cysteine sulfhydyl groups. In general, the overall lower cysteine content in proteins allows for labeling these residues to saturation without increasing the overall hydrophobicity of the proteins to cause insolubility problems. Thus this chemistry takes more advantage of the increased sensitivity of the CyDyes (150–500 pg), and even more so for proteins with high cysteine content. However, the saturation chemistry is only available for Cy3 and Cy5, requires additional optimization, and is blind to the small but significant population of noncysteine-containing proteins. For these reasons, saturation DIGE is usually reserved for experiments where samples are limited, where the advantage of the increased sensitivity outweigh these additional considerations.
13. In comparison, commonly used silver or colloidal Coomassie blue (*ca*. 5–10 ng sensitivity) stains typically exhibit a dynamic range of less than two orders of magnitude ***(9,20)***. The CyDye labeling system is compatible with the downstream processing commonly used to identify proteins via MS and database interrogation, which involves the generation of tryptic peptides within excised gel plugs. Trypsin cleaves the peptide bonds at the C-terminal side of lysine and arginine residues, but peptide generation is mostly unhindered as so few lysine residues are modified by dye labeling.
14. The saturation dyes are only available for Cy3 and Cy5. The pooled sample internal standard methodology can still be employed with only two dyes, however this will require doubling the number of gels used in the experiment, where each gel contains a single individual sample (satCy5) along with an equal aliquot of the pooled standard (satCy3).
15. For saturation chemistry, general methods and considerations are the same as for the minimal chemistry, but there are several unique features to also consider for the saturation chemistry. First, careful optimization of the labeling conditions must be carried out for each new sample set to ensure complete reduction of cysteine residues. Insufficient labeling will lead to multiple spots in the second dimension because of molecular weight and hydrophocity shifts. Overlabeling results in side reactions with the ε-amine groups of lysine side chains, but because

the maleimide dyes do not carry compensatory charge, this results in the overall loss of a charge, which creates a series of isoelectric forms in the first dimension ("charge trains").

16. To compensate for small pipettor error, adding extra volume (1–2 μL) is acceptable. L-buffer volume can be increased, if necessary, for complete resolubilization, although 100–250 μg or more should resolubilize readily in this volume. The volume of labeling buffer used for resolubilization should not exceed 40 μL per sample when using cup loading for sample entry to ensure that the final volumes will not exceed the capacity of the cup loading (*ca*. 100–150 μL).
17. These methods are provided assuming that all gels to be run will be used both for analytical (quantification), as well as preparative (providing material for subsequent mass spectrometry), purposes. Current recommendations from the manufacturer are to label 50 μg of sample with 400 pmol CyDye. Sufficient amount of unlabled sample can be added to the quenched reactions to achieve final protein amounts to facilitate subsequent MS. Alternatively, many have found that the ratios can be adjusted to label increasing amounts of sample (up to 200–300 μg with 200 pmol dye) without adversely affecting the overall labeling reaction (presented here), but may increase the background noise for low-abundance proteins.
18. If samples are to be introduced using anodic cup loading, simply bring this mixture up to 100 μL in R-buffer and proceed with cup loading. R-buffer can always be supplemented with additional DTT using the 200 mg/mL DTT stock solution. In the presence of Destreak reagent for focusing in pH ranges greater than pH 8.0, the addition of an equal volume 1X R-buffer should provide a sufficient amount of DTT without interfering with the Destreak reagent.
19. Comparison of minimally labeled protein 2D maps with unlabeled protein maps is generally not a problem, as the addition of only one dye molecule does not generally prevent the facile matching of small alterations in protein mobility between the 2–5% labeled protein and the remaining unlabeled protein that will provide enough material for MS.
20. Poststaining is not necessary with saturation DIGE because an unlabeled population with potentially different migration characteristics will not exist.
21. This treatment binds the gel to one of the glass plates and, therefore, prevents shrinking/swelling during poststaining and protein excision processes, thereby facilitating accurate robotic protein excision. Nothing should be placed on top of wipes that are covering bind silane-treated plates, as this may leave impressions that are detected during the scanning phase. Assembly and casting too soon may create a binding surface on the opposite glass plate, preventing the gel to be subsequently poststained and picked. Automated protein excision can be facilitated for certain systems by placing fluorescent alignment reference targets on the plate, which can be performed at this stage.
22. A stacking gel is not required for 2D-GE, as the proteins are effectively "stacked" to the height of the IPG strip. SDS is also not essential in the separating gel, as the SDS associated with the proteins during the equilibration step, and present in the running buffer, is sufficient.

23. Samples of similar nature should always be focused simultaneously for optimal reproducibility. Focusing programs vary for some pH gradients: a typical program for many ranges is 500 V for 500 V-hours, stepping to 1000 V for 1000 V-hours, followed by a final step to 8000 V until more than 40 V-hours have been reached. Check recommendations from specific vendors.
24. Volume of equilibration buffer should be large to ensure sufficient removal of ampholines and other components of the 1D run.
25. Carefully wash out any remaining liquid on top of the SDS-PAGE gel. Prewet the IPG strip with 1X running buffer and place the strip between the gel plates with the plastic backing adhering to the inside surface of one of the glass plates. The prewetted running buffer will facilitate the manipulation of the IPG strip down the inside surface of the plate and on top of the SDS-PAGE gel.
26. An agarose overlay, used by many protocols, is not absolutely necessary to ensure proper contact between the IPG strip and the 2D SDS-PAGE gel. Using a thin card or ruler to carefully tap down the pre-wetted plastic backing of the IPG strip to the gel is usually sufficient and removes the added problems associated with the overlay, such as trapped air bubbles in the solidified agarose.
27. Running gels at much less than 1 W/gel can improve resolution in the high molecular weight regions of the 2D gel. Use wattage appropriate for the 2D unit being used. Many different gel units can accommodate increased power by compensating for the increased heat.
28. Absorption/emission maxima in DMF are 491/506 for Cy2, 553/572 for Cy3, and 648/669 for Cy5; although care must be taken to scan in regions of each spectrum that do not contain absorbance or emission in the other spectra, which may mean using a nonmaximal region of a given spectrum.
29. Comparison of the 2D spot maps between saturation-labeled samples and minimal labeled or unlabeled samples is impossible, as proteins containing multiple cysteine residues may appear as significantly larger M_r species when labeled with the saturation dyes, which of course will be impossible to predict without first knowing the protein identity.
30. The roots for these various software packages can be found in astronomical software designed to facilitate the search for near-Earth objects in a constant star field ***(20,21)***.
31. Almost all software packages for 2D electrophoresis involve matching of protein-spot patterns between gels. For DeCyder, it is used in the BVA module to match the quantitative data obtained from the triply coresolved protein signals from each gel in the DIA module (where gel-to-gel variation does not come into play). Manual verification of the matching is almost always required with any software package.
32. There are many “all-or-none” type of experiments where the single gel comparison may be valid, and subtle changes are not expected. Nevertheless, using independent replicates and the pooled-sample internal standard methodology is still needed to control for nonbiological sample preparation error.
33. The multigel approach allows many data points to be collected for each group to be compared. Spots of interest can be selected by looking for significant change

across the groups. Student's *t*-test and ANOVA probability scores (*p*) indicate the probability that the observed change occurred from stochastic, random events (null hypothesis). Probability values less than 0.05 are traditionally used to determine a statistically significant difference from the null hypothesis. As this represents 50 potential false-positives for 1000 resolved proteins, confidence intervals within the 99th percentile ($p < 0.01$) are arguably more valid, and can be attained using DIGE ***(13,23–27)***.

References

1. Gorg, A., Postel, W., Domscheit, A., and Gunther, S. (1988) Two-dimensional electrophoresis with immobilized pH gradients of leaf proteins from barley (Hordeum vulgare): method, reproducibility and genetic aspects. *Electrophoresis* **9,** 681–692.
2. Gorg, A., Obermaier, C., Boguth, G., et al. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21,** 1037–1053.
3. Wolters, D. A., Washburn, M. P., and Yates, J. R., 3rd. (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73,** 5683–5690.
4. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999.
5. Mason, D. E. and Liebler, D. C. (2003) Quantitative analysis of modified proteins by LC-MS/MS of peptides labeled with phenyl isocyanate. *J. Proteome Res.* **2,** 265–272.
6. Ross, P. L., Huang, Y. N., Marchese, J. N., et al. (2004) Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics* **3,** 1154–1169.
7. Vogt, J. A., Schroer, Holzer, K., et al. (2003) Protein abundance quantification in embryonic stem cells using incomplete metabolic labelling with 15N amino acids, matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry, and analysis of relative isotopologue abundances of peptides. *Rapid Commun. Mass Spectrom.* **17,** 1273–1282.
8. Unlu, M., Morgan, M. E., and Minden, J. S. (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **18,** 2071–2077.
9. Tonge, R., Shaw, J., Middleton, B., et al. (2001) Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics* **1,** 377–396.
10. Von Eggeling, F., Gawriljuk, A., Fiedler, W., et al. (2001) Fluorescent dual colour 2D-protein gel electrophoresis for rapid detection of differences in protein pattern with standard image analysis software. *Int. J. Mol. Med.* **8,** 373–377.
11. Gade, D., Thiermann, J., Markowsky, D., and Rabus, R. (2003) Evaluation of two-dimensional difference gel electrophoresis for protein profiling. Soluble

Proteins of the marine bacterium PIRELLULA sp. strain 1. *J. Mol. Microbiol. Biotechnol.* **5,** 240–251.
12. Alban, A., David, S. O., Bjorkesten, L., et al. (2003) A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics* **3,** 36–44.
13. Friedman, D. B., Hill, S., Keller, J. W., et al. (2004) Proteome analysis of human colon cancer by two-dimensional difference gel electrophoresis and mass spectrometry. *Proteomics* **4,** 793–811.
14. Gerbasi, V. R., Weaver, C. M.. Hill, S., Friedman, D. B., and Link, A. J. (2004) Yeast Asc1p and mammalian RACK1 are functionally orthologous core 40S ribosomal proteins that repress gene expression. *Mol. Cell Biol.* **24,** 8276–8287.
15. Sitek, B., Apostolov, O., Stuhler, K., et al. (2005) Identification of dynamic proteome changes upon ligand activation of Trk-receptors using two-dimensional fluorescence difference gel electrophoresis and mass spectrometry. *Mol. Cell Proteomics* **4,** 291–299.
16. Wessel, D. and Flugge, U. (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138,** 141–143.
17. Seike, M., Kondo, T., Fujii, K., et al. (2004) Proteomic signature of human cancer cells. *Proteomics* **4,** 2776–2788.
18. Yokoo, H., Kondo, T., Fujii, K., Yamada, T., Todo, S., and Hirohashi, S. (2004) Proteomic signature corresponding to alpha fetoprotein expression in liver cancer cells. *Hepatology*. **40,** 609–617.
19. Olsson, I., Larsson, K., Palmgren, R., and Bjellqvist, B. (2002) Organic disulfides as a means to generate streak-free two-dimensional maps with narrow range basic immobilized pH gradient strips as first dimension. *Proteomics* **2,** 1630–1632.
20. Lilley, K. S., Razzaq, A., and Dupree, P. (2002) Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Curr. Opin. Chem. Biol.* **6,** 46–50.
21. Rabinowitz, D. (1991) Detection of Earth-approaching asteroids in near real time. *Astron. J.* **101,** 1518–1559.
22. Scotti, J. (1993) Computer aided near-Earth object detection. In: *Asteroids, Comets, Meteors* (Milani, DiMartino, and Cellino, eds.) Kluwer, Dordrecht, The Netherlands, pp. 17–30.
23. Knowles, M. R., Cervino, S., Skynner, H. A., et al. (2003) Multiplex proteomic analysis by two-dimensional differential in-gel electrophoresis. *Proteomics* **3,** 1162–1171.
24. Wang, D., Jensen, R., Gendeh, G., Williams, K., and Pallavicini, M. G. (2004) Proteome and transcriptome analysis of retinoic acid-induced differentiation of human acute promyelocytic leukemia cells, NB4. *J. Proteome Res.* **3,** 627–635.
25. Prabakaran, S., Swatton, J. E., Ryan, M. M., et al. (2004) Mitochondrial dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Mol. Psychiatry* **9,** 684–697.

26. Zhang, Y. Q., Matthies, H. J., Mancuso, J., et al. (2004) The Drosophila fragile X-related gene regulates axoneme differentiation during spermatogenesis. *Dev. Biol.* **270,** 290–307.
27. Zhang, Y. Q., Friedman, D. B., Wang, Z., et al. (2005) Protein expression profiling of the drosophila fragile X mutant brain reveals up-regulation of monoamine-synthesis. *Mol. Cell Proteomics* **4,** 278–290.

14

Proteomic Data Exchange and Storage

Using Proteios

Per Gärdén and Rikard Alm

Summary

Proteios (http://www.proteios.org) is an initiative for the development of a comprehensive open source system for storage, organization, analysis, and annotation of proteomics experiments. The Proteios platform is based on existing principles for proteomics data publishing and data exchange.

Key Words: Mass spectrometry; 2D gel; data repository; proteomics data collection; laboratory information management system; LIMS; free; open source; Java; mzData; MIAPE.

1. Introduction

Proteomics as a discipline in medical science is comparatively new and involves both biology and chemical engineering. The successful development of technologies for the analysis of DNA sequence information, especially DNA microarray, have helped to shift the manner in which biological systems are studied. As a result of technological advances in expression profiling methods, differential gene and protein expression can now be studied.

Proteomics experiments arise from issues in practice related to understanding a disease, the development of a new drug, or understanding basic biological processes. A prerequisite is to identify a specific molecular target, usually a protein or pathway. The goal of such a study is to know which proteins may be differentially expressed in response to a stimulus or between different disease states.

The successful integration of a proteomics workflow into a biochemical engineering laboratory interfaces with both biochemistry, analytical chemistry, and computer science (e.g., development of efficient algorithms). The amount of data from the experiments demand data management tools.

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

In spite of several efforts, proteomics lacks standardized software tools and data formats to manage experimental data. Proteios offers an all-encompassing proteomics experiment data model and aims to be the glue holding different steps of the experiment together.

2. Data Handling in Proteomics

2.1. Data Model

When studying a problem domain from a computer science point of view, a common approach is to try to define how a software is to be used in the context of a specific setup, a *use case*. Use cases describe sequences of actions which make up workflows. A software model to hold use case data is developed and tested against each case. In proteomics the basic use case is an experiment. Capturing the data from different proteomics experiments in a meaningful way demands a model consistent with all occurring use cases. And it must be foreseen in the model how these use cases can be merged into complete experiments.

Recently the Microarray Gene Expression Data Society, the Human Proteome Organization, and the Reporting Structure for Biological Investigations have started the Functional Genomics Experiment (FUGE) ***(1)*** effort to create overall models, as well as specific use cases for different bioinformatics fields, including proteomics. Proteios, like its initial source of inspiration PEDRo ***(2)***, uses a model specifically created for proteomics experiments.

As a data model, Proteios spans the whole proteomics experiment, from hypothesis to actual protein identifications. Proteios is the software tool based on this model, managing sample information, raw data, images, analysis results, as well as connectivity to protein identification, data viewing, and analysis tools. The organization and interface of Proteios is designed to closely follow the natural workflow of the proteomics researcher, and is compatible with both liquid chromatography (LC)–tandem mass spectrometry and two-dimensional (2D)-gel experiments. Being an open source software, Proteios can be used independently of equipment manufacturers, and be extended or modified to fit local needs.

2.2. Data Repository

Typical experiment data management involves functionality to store results from experiments, retrieve previously stored data, as well as the ability to run analysis on existing data. This is *repository* functionality. Current software tools to retrieve and maintain data in repositories are equipment or method centered and do not cover a complete proteomics experiment. Basing a repository and eventually a laboratory information management system (LIMS) on such tools risks tying the user to a specific environment prescribed by the equipment manufacturer.

Information management will thus be difficult to adapt to different experiment setups. As there is a need for a public and generic data repository, one of the ambitions with Proteios is that it should be possible to use as a generic repository ***(3)***.

Few proteomics tools handle automatic reading and validation of data. Data amounts are huge, most of which is not meaningful or feasible to add manually. Automation is the trend in gel handling, chromatography, and mass spectrometry (MS). A realistic proteomics repository tool must be able to do automatic data import. Also, as typical experiment workflow consists of several distinct steps, it must be possible to import experiment data in parts, some of which are manual whereas others are machine generated. Data will come in different formats and originate from different equipment. Proteios aims to smoothen out these differences into a generic repository. Automatic and semiautomatic composition of experiment steps is possible and tracking provides LIMS functionality.

2.3. Data Formats

Handling the large amounts of data arising from proteomics experiements requires specific file formats. Proteomics data formats are abundant with some formats being equipment specific, whereas others deal only with parts of experiments, typically in conjunction with some treatment or analysis equipment. Additionally, there are specific formats to handle sample processing, MS, and identification results from database searches.

The Proteomics Standards Initiative ***(4)*** have recognized the need for standard formats and are developing the "minimum information about a proteomics experiment" (MIAPE) to cover the larger experiment context and serve as a publication standard. Further data format standards are expected to emerge from generic data modeling work within FUGE. Sample processing data formats are specific to LC tools and gel-picking robots (e.g., Ettan, ProPic, GelPix).

There are plenty of software tools to convert between MS equipment specific formats (e.g., ReAdW, MassWolf, mzStar, mzBruker). As a publication standard for MS experiments and as a part of MIAPE mzData ***(5)*** has been developed in cooperation with instrument manufacturers. Although mzData is not a substitute for vendor specific formats, it is expected that MS machine vendors will provide software transforming their raw files to mzData. The raw data itself is stored directly in mzData in binary format and can also be put in a referenced file.

The mzXML format from the Seattle Proteome Center is a raw data format turned into XML in that it accepts and contains all spectrum data as binary. There are converters from several machine specific formats to mzXML ***(6)***. Both mzData and mzXML can also be used to store equipment settings.

Analyzing the MS result means trying to identify specific proteins, or in the case of MS/MS specific peptides, in the spectrum pattern against databases of

known such patterns with database search tools like Mascot ***(7)*** or Sequest ***(8)***. The result of an invoked search is returned in a data format specific to the search tool being used. Usually search data must be submitted to a web page. Application programmer's interfaces (API) exist in some cases (e.g., Mascot) but usually not. Current Proteomic Standards Initiative activities include the development of mzIdent, which is intended to represent search paramaters and analysis results from protein identification software tools.

Proteios can import database search results from the PIUMS ***(9)*** tool, based on MS. Also, as Proteios is completely compatible with the Virtual Expect Mass Spectrometrist (VEMS) ***(10)*** data model, VEMS searches can readily be imported. It is also our ambition to handle existing and upcoming MIAPE standard data formats as they become availiable. Currently Proteios supports mzData.

3. The Proteios Application

3.1. Core Functionality

Proteios is a client–server application, with a many-to-many relationship between clients and servers. This architecture makes it possible to share data between users worldwide. The Proteios data model is a *graph* of elements, which represent real experiment objects or data containers, e.g. Sample, Gel2D. The model is mapped to XML schemas and database tables in the Proteios application. Proteios is implemented in Java, and is thus platform independent. Specifically, the Proteios client runs as a Java application on virtually any workstation. Connections to server database(s), like Oracle and MySQL, are done through the Hibernate ***(11)*** middleware. This gives the user a wide variety of database management systems to use as a Proteios back-end server.

Proteios also provides LIMS functionality in that it assists the user in managing and connecting data from heterogeneous sources with the aim to track all relevant information from an experiment—sample, processing, MS, and protein identification. This sets it apart from other applications, most of which either focus on MS (e.g., Sashimi ***[6]***, OPD ***[12]***, PROTEOME-3D ***[13]***) or do not enable automized data capture (e.g., PEDRo ***[2]***).

The Proteios data model is designed to map the steps of a proteomics experiment. We have specifically considered the fact that some parts of an experiment are generated automatically, whereas others are manual. Also, we take into account different points in time and different locations, which corresponds to a typical researcher's work situation.

The prerequisites for installing Proteios is Java and access to a relational database. Although installation of Java and the Proteios application itself is fairly simple a system administrator may be needed to install a database.

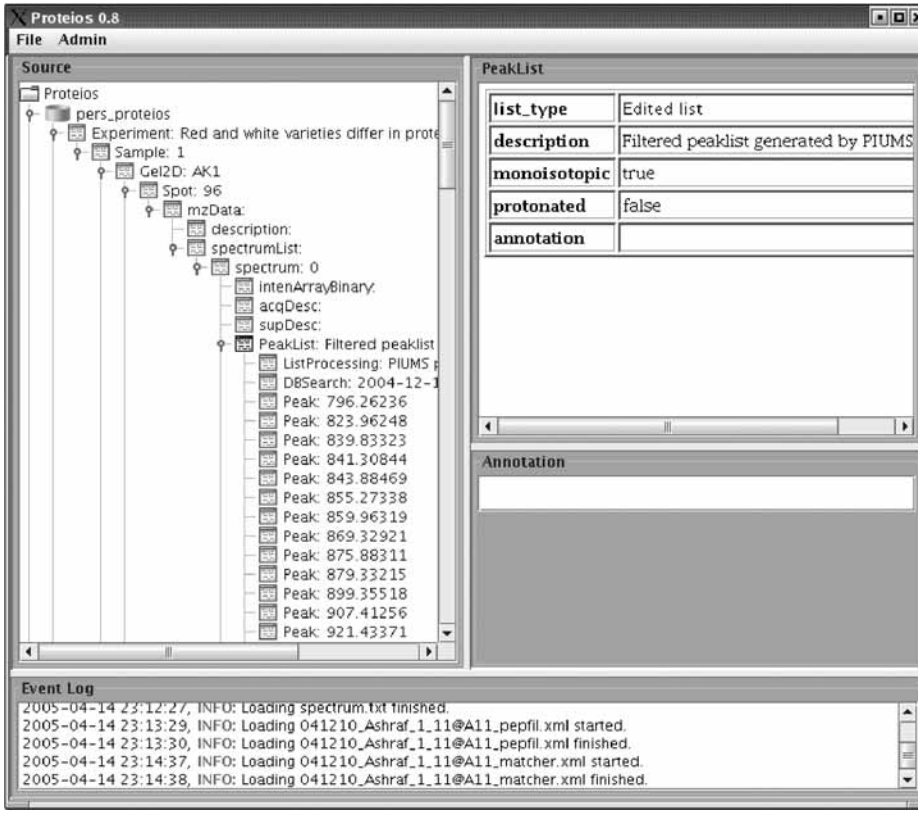


Fig. 1. The Proteios graphical user interface.

By running Proteios, a complete proteomics experiment can be set up by an individual researcher, combining manually created elements (from within the Proteios application itself) with imported files. Creating a complete experiment data set manually by inserting *all* data, e.g., all spots peaks and so on, is completely infeasible. On the other hand, not even in a perfect laboratory environment can all data be entered automatically. A realistic approach is semiautomatic, manually creating some data elements, but importing bulk data from files and then using the Proteios graphical user interface (GUI) for selecting and merging parts of the experiment together.

Proteios provides the GUI as its main client (*see* **Fig. 1**). Data is easily and intuitively viewed and handled as graphical objects. Data from different steps of an experiment workflow can be put together with intuitive drag and drop. The representation follows the Proteios data model, displaying a tree structure of elements—a tree view of the model graph, which can be rerooted to highlight items of interest. Rerooting, together with import and export, is very versatile;

any tree view can be exported. Export views can be given schemas to validate against a strictly defined tree in the graph. Together with extended style sheet transformations this provides a versatile system for basic report generation. Data can also be annotated and extended with, for instance, protein identifications from search engines like PIUMS ***(9)*** and Mascot ***(7)***.

As any element in the model graph can be used as the root of a *tree* view through the graph and XML files are tree structures, tree views can effectively be described by XML schemas, which enables standard XML tools to assist in interchanging data with Proteios. Although XML files sometimes can be very large, XML is a great advantage because it allows data to be *validated*. Validation is important because it prevents corrupt data from being accepted by Proteios. The XML schemas prescribe file formats, which can be imported to and exported from Proteios. Often, but not necessarily, the file format corresponds to a tree view of the Proteios data model. Viewing data rooted on Experiment is the default view of data. "Experiment" is also the root of an XML schema covering the whole Proteios data model. This schema can be used to import and export all experiment data with validation. Parts of an experiment can be imported separately with specific schemas validating import and export of these parts.

When storing data in databases, as well as in working memory, the graph structure of the model is retained. Any element can be directly accessed without having to navigate through a tree of elements. Viewing of data can be done rooted on any type of element, with "Experiment" being the default.

The following sections describe the Proteios data model in some detail. Diagrams are provided as *unified modeling language* class diagrams. Elements (classes) are shown as boxes with interconnecting lines to describe relationships between them. Relationships have multiplicity 1, 0..1, 1..n, * indicating that one single, one or zero, any number but at least one, or zero, or many elements respectively can be associated. Arrows indicate inheritance.

3.2. Sample Information

For a complete Proteios structure, sample generation is the starting point with an "Experiment" element as the basis (*see* **Fig. 2**). In a research situation, we start out with a hypothesis for our experiment and a sample to run our experiment on. This part of the model contains data which a researcher is expected to enter manually. The Sample element holds an identification code, the production date, and the name of the responsible person, facilitating sample tracking. As seen from the multiplicity in the diagram, an experiment may contain several samples, but any given sample only belongs to one single experiment.

"SampleOrigin" holds basic information, such as the specific biological material used, what tissue or subcellular fraction was studied (if appropriate),

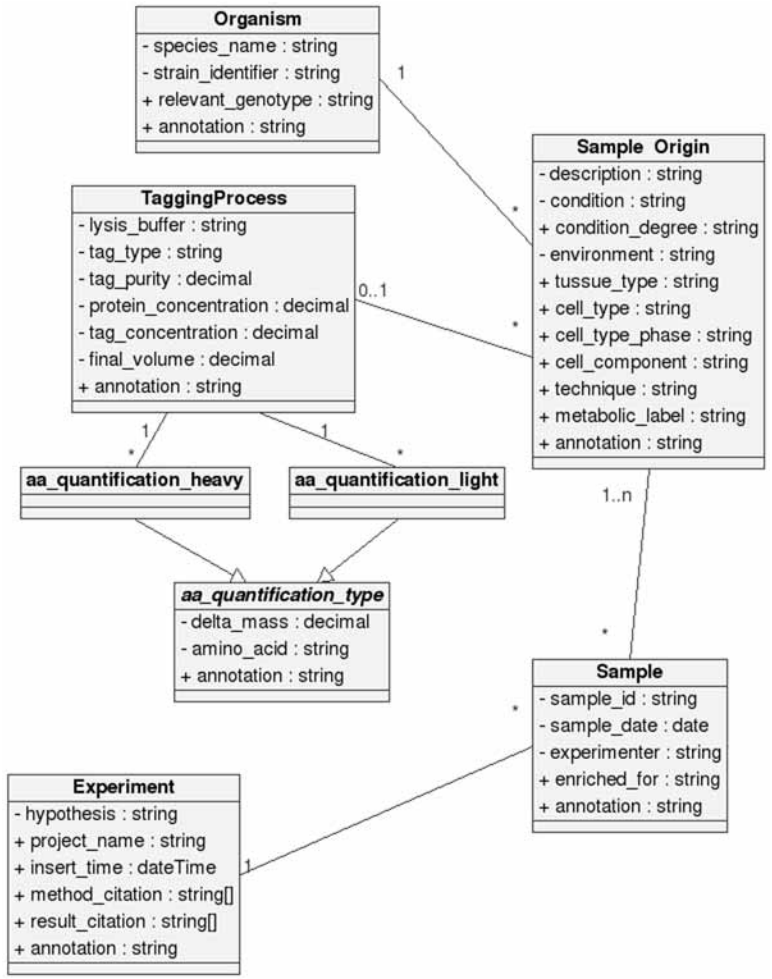


Fig. 2. Sample generation.

and the experimental conditions to which the organism was subjected. "TaggingProcess" describes the labeling of samples for quantitative studies, such as difference gel electrophoresis or isotope-coded affinity tag (ICAT) MS.

The "TaggingProcess" element allows for any quantification to be used. Typically, one would use stable isotopes like H, C, or N. The tagging in Proteios caters for any up- or downtagging. Just note the delta mass caused by the tagging isotope and which amino acid is being tagged. Regardless of tagging data here, there is an ICAT entity with the "DBSearchParameters" element (*see* **Subheading 3.5.**). For difference gel electrophoresis there is a corresponding entity in the gel element.

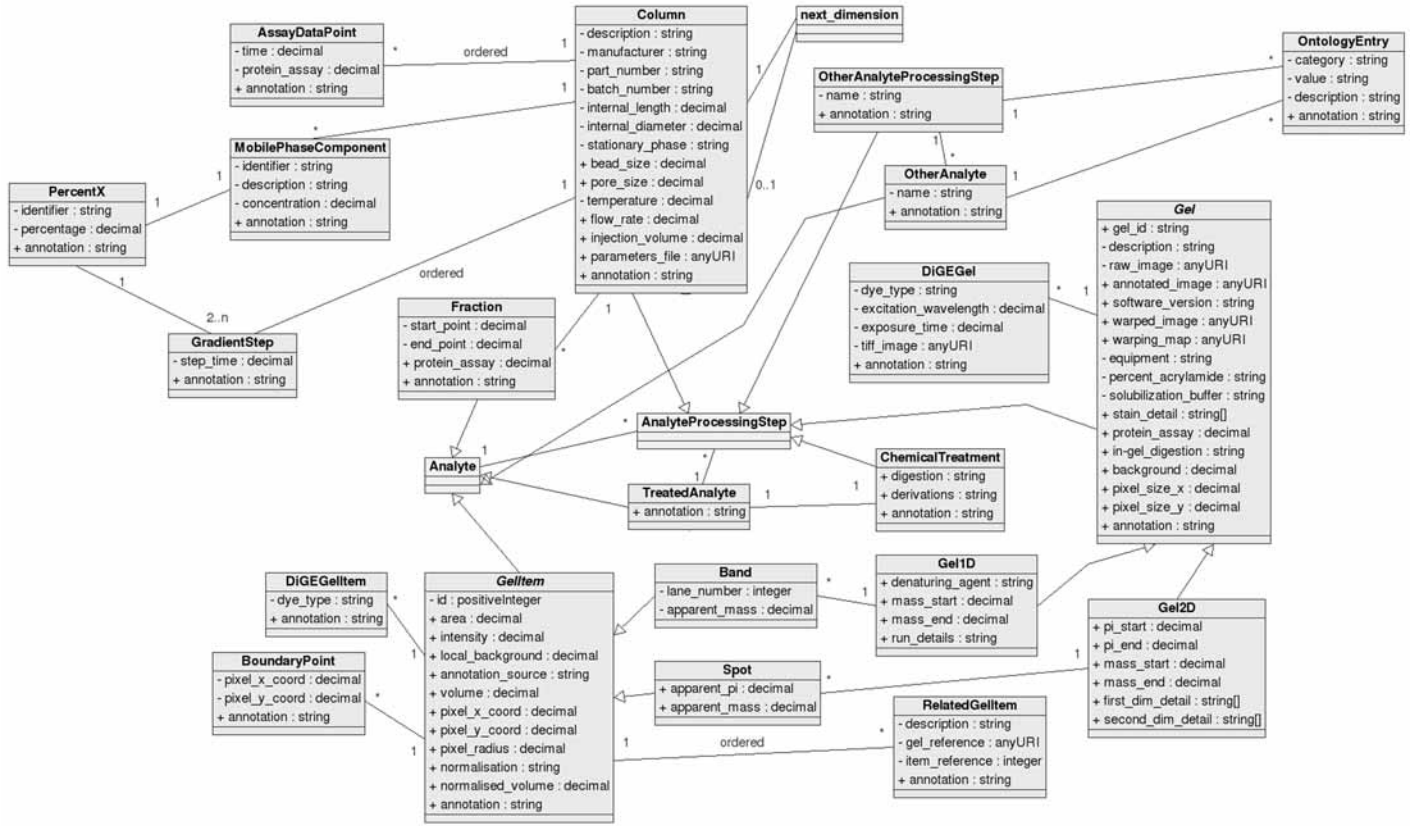


Fig. 3. Sample processing.

3.3. Sample Processing

The sample processing part of the data model (*see* **Fig. 3**) adds treatment of the sample under investigation. The Proteios data model caters for the usage of one-dimensional (1D) and 2D gels, as well as LC. Also there is the generic treatment "other" to handle any other kind of treatment.

A "Sample" in the model is one out of several possible instantiations of the abstract class "Analyte." It is needed in order for data to fill out the model. An analyte can directly be used as the source for MS, or it can be put through one or more analyte processing steps. The result of one analyte processing step can be fed back into the cycle as the next analyte. This cyclical design enables a complex series of concatenated processes to be easily described. Other analytes are "Fraction," "Band," "Spot," "TreatedAnalyte," or "OtherAnalyte." As a typical use case, a band from a 1D gel is run through 2D LC, before running MS using a particular fraction.

The Proteios data model allows five subclasses of the abstract superclass "AnalyteProcessingStep," four of which are "Gel1D," "Gel2D," "Chemical Treatment," and "Column." The fifth subclass, "OtherAnalyteProcessingStep," provides a mechanism to capture any other form of analyte processing by linking to a series of entries in an ontology (a controlled, structured vocabulary).

One "Gel" contains many "GelItems" (another abstract class of which "Spot" and "Band" are instances in the case of 2D and 1D gels, respectively). Typically, items in a gel are handled by picking robot. The experimenter will have to fill in gel-specific information the first time. The gel created can then be added to the sample set up in the previous section.

3.4. Mass Spectrometry

MS can be run on any analyte. Proteios uses mzData ***(5)*** to hold MS information. Data here is largely manual. Automatic data consist of "spectrumList" and "spectrum." The MS part of the model (*see* **Fig. 4**) differs from the rest of the Proteios model in that its elements are largely generic, whereas the rest of Proteios contains elements, which are very specific in its entities. The elements of main interest are "mzData," "spectrumList," "spectrum," and "sourceFile."

Proteios lets users manually add (and edit) the elements "mzData," "spectrumList," "description," "instrument," "detector," and "dataprocessing." Thus, the machine settings, together with the name of the experimenter, can be added manually.

Within mzData there can be spectrum raw data stored in binary form. Storing such huge binary data within XML only makes sense if one wishes to maintain *everything* all in one file. Proteios supports this. In mzData it is mandatory. In Proteios it is optional and the binary data will instead be stored in a referenced source file.

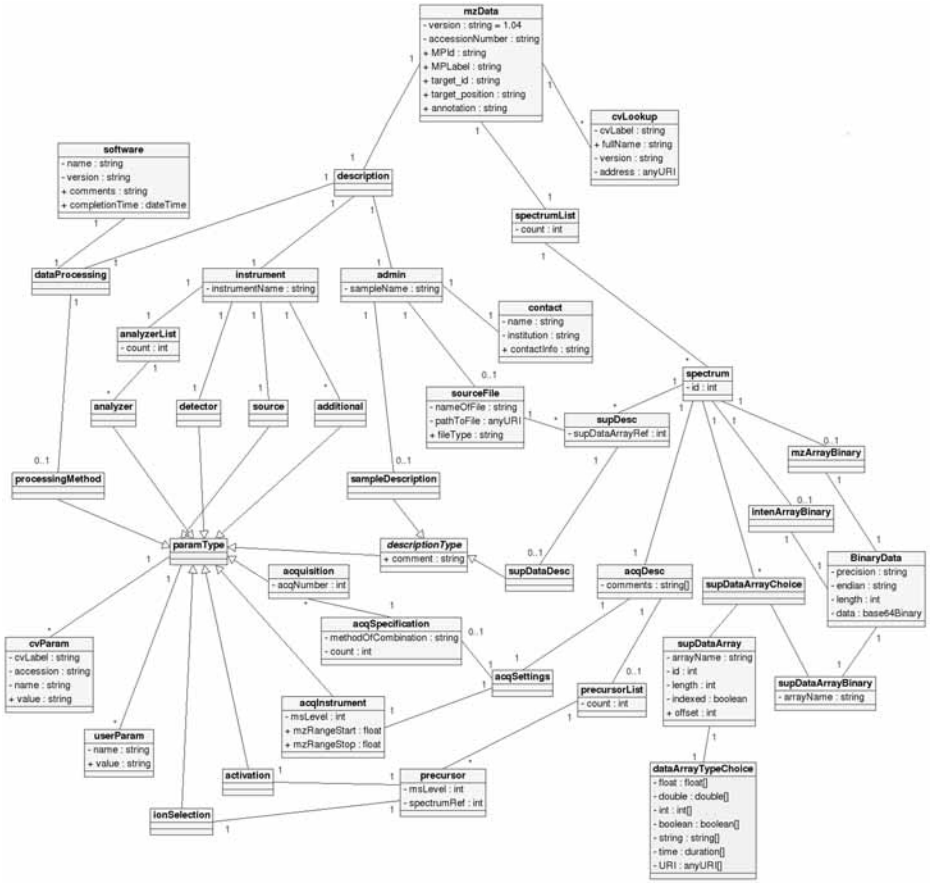


Fig. 4. Mass spectrometry.

3.5. MS Results Analysis

In analyzing MS data, Proteios handles one or several “PeakList” elements. Peak lists are processed spectra from a MS experiment. Each “PeakList” can have several “Peak” elements representing possible peptides. One will try to match this peak list with known protein spectra from a database. This is purely software work for which there is a wide range of third party tools. Not all masses will be fragments from a protein, some of them will be contaminants.

In principle, database search tools are independent of the type of mass spectrometer used and the database searched. This should be reflected in EBI’s upcoming mzIdent standard, which will be supported by Proteios. Proteios imports treated peak lists as defined by the PIUMS ***(9)*** search engine.

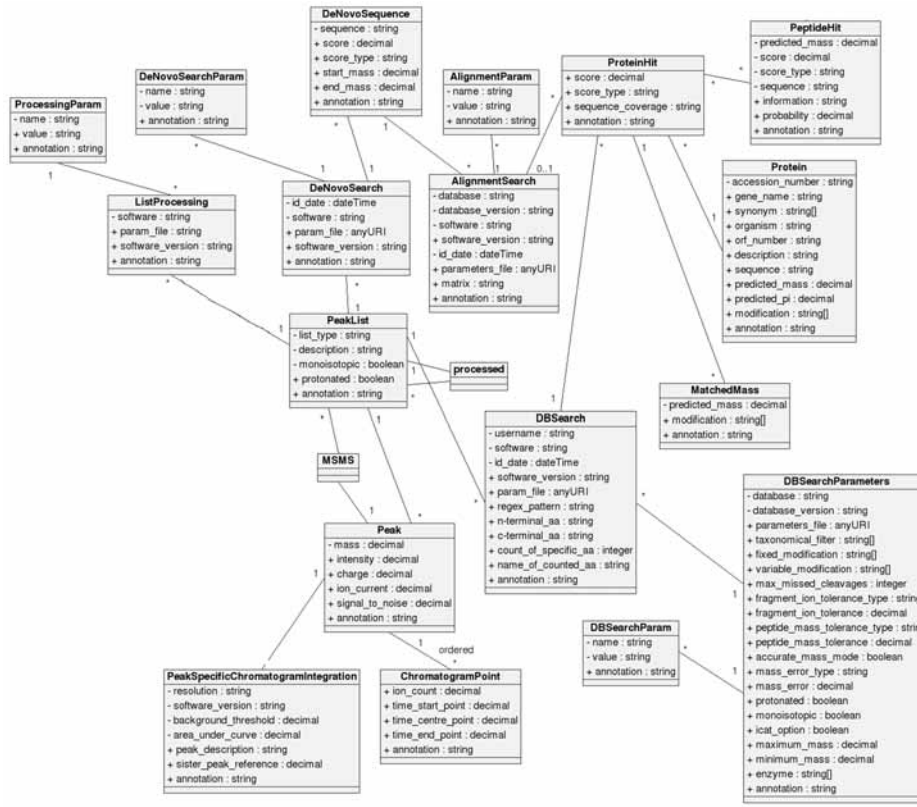


Fig. 5. Mass spectrometry result analysis.

Peak lists as defined by VEMS ***(10)*** can be imported, as VEMS uses Proteios as its data storage format.

The "PeakList" element (*see* **Fig. 5**) connects to a "spectrum" element in the MS part of the model. Based on any peak a derived peak list from the second stage of a tandem mass spectrometer run can be generated, so that this second peak list is a child of the precursor peak and may in its turn contain peaks. Through this cyclic construct Proteios can represent not only MS/MS, but MS to the degree.

The individual peaks in a list are described by mass, intensity, and charge. To perform a protein identification, a particular "PeakList" would be submitted to an identification tool, such as PIUMS ***(9)***, VEMS ***(10)***, Sequest ***(8)***, or Mascot ***(7)***.

The classes "DBSearch" and "DBSearchParameters" capture information about who did the identification, when they did it, what program they used, what database (of theoretical proteins from an *in silico* digest of an organism's predicted proteome) was used, what potential modifications were allowed on

proteins from the sample that generated the peak list, any additional information or another chemical analysis, and whether the ions carry ICAT labels (such that only cysteine-containing peptides were searched against).

Based on possible peptide hits, the identification process will (hopefully) lead to an identification of one or several proteins. Proteios represents this by having one or several protein hits connected to the "DBSearch" element. A protein hit can refer to a clearly identified protein, it can refer to any number of peptide hits or it can refer both at the same time. How can there be a protein hit without any protein? Having a protein hit element does not necessarily mean that a protein has been identified. A protein hit also serves as a container for peptide hits. So it might be the case that several peptides were identified, but there is not enough certainty as to what protein they adhere to.

4. Capturing an Experiment With the Proteios Data Model

4.1. Aggregating Experiment Data: A Case Study

This section describes how Proteios is used when working with 2D-gel electrophoresis and protein identification. Our sample in this case is a complex protein mixture from an extract. This sample is run on a 2D gel to separate proteins into distinct spots. The gel is scanned and passed through image analysis for the detection of spots, and a gel picking robot is set up to pick spots chosen for identification. The robot digests the proteins and spots the resulting peptides onto a matrix-assisted laser desorption/ionization (MALDI) plate, which is then analyzed in a MALDI–MS mass spectrometer. The resulting spectra are subjected to a database search to identify proteins. It is a good idea to save data at different points in the process. Proteios also provides "Restore" to reload the latest saved state.

To get started we need to manually create an "Experiment" element (*see* **Fig. 6**), the basic element of an experiment in Proteios. This is done by selecting one of our connected databases in the GUI and adding an "Experiment" element, which lets us enter information about our experiment hypothesis, (an optional) project name, insertion time and one or several method, and result citations. Thereafter we choose "add" on the "Experiment" we just created to add a "Sample" element. In the "Sample" element we type in information about the sample, like origin and organism. This sample is now an "Analyte," which can be subjected to one of several analyte processing steps, e.g., gel separation, liquid chromatogram, or even MS.

We use 2D-gel electrophoresis as an initial separation step (analyte processing step). A Gel2D element is created from within the Proteios GUI, to hold parameters used in gel creation—equipment, the percentage of acrylamide, and so on, as well as a reference to the resulting raw image (*see* **Fig. 7**). This information

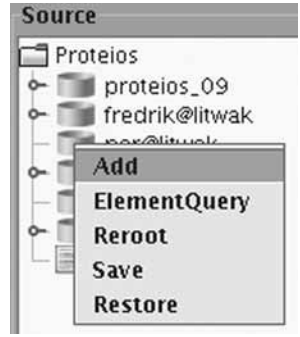


Fig. 6. Manually adding an experiment.

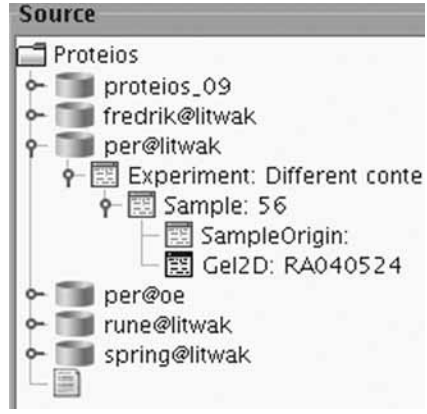


Fig. 7. Having created the structure Experiment-Sample-Gel2D.

must be entered manually. The gel image is analyzed to identify spots. Information on spots of interest is sent off to a picking robot, which treats the gel plugs with trypsin and washes out the peptides from the plug. This is an analyte processing step called "ChemicalTreatment." The robot applies the peptides from each spot onto a MALDI plate and produces an XML file containing information about the spots we selected, their location in the raw image, the enzymatic digestion, and where they were put on the MALDI plate. Proteios imports the picking robot file and presents the gel with all the spots beneath (*see* **Fig. 8**). We use drag and drop in the Proteios GUI to add the picking robot information and the spots to the gel we created manually (*see* **Fig. 9**). The whole gel handling step is thus semi-automatic with the most tedious task of gathering all the spots done automatically. Spots can be added one-by-one or, preferably, all in one go.

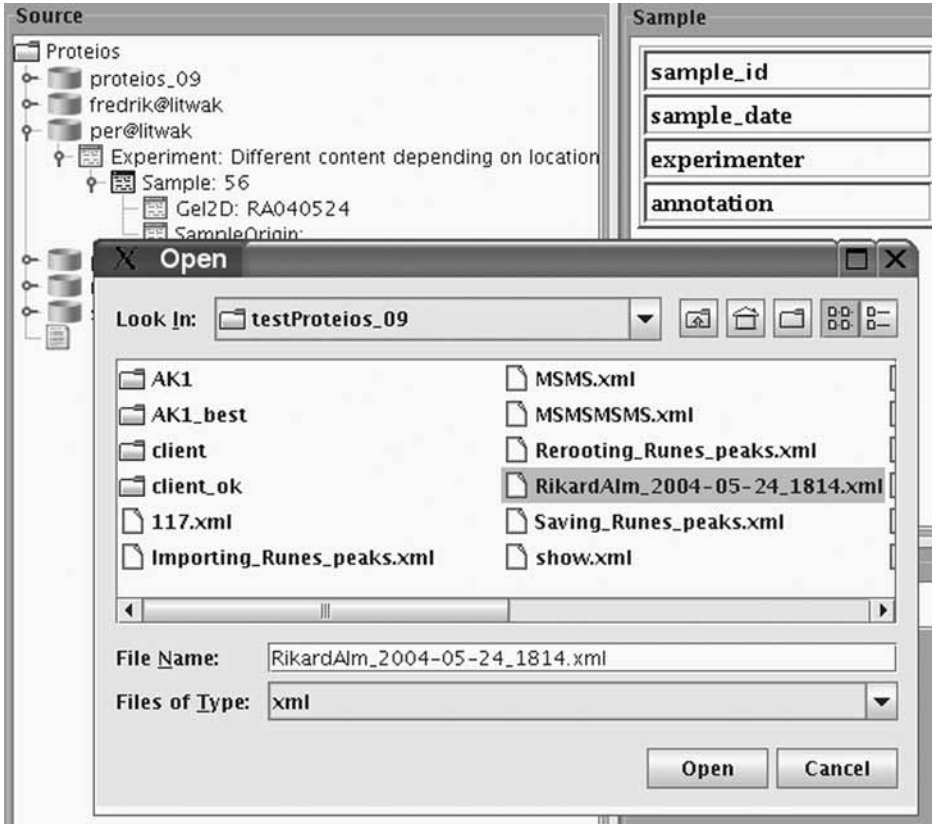


Fig. 8. Importing picking robot data.

Each spot from the 2D gel is an “Analyte” which can be subjected to further processing, in this case MALDI-MS. The mass spectrometer produces raw spectrum files that are saved externally and referred to from within Proteios with an universal resource identifier. After processing the spectra in a peak extraction and database search software, the resulting peak lists and protein identifications are imported into Proteios as files. Proteios is integrated with the PIUMS ***(9)*** database search tools, so all data, peak extraction, parameters, and database search results, can be imported in one go. The result of this import is mzData elements, which can be connected to the gel we are working on. Because Proteios is designed for keeping track of workflows, extra tracking can be added to facilitate the merging of the different parts. This makes Proteios comparable with other LIMS. In the end we have a complete experiment saved in our repository with minimal labor and as the basis for producing reports and to further annotate the data.

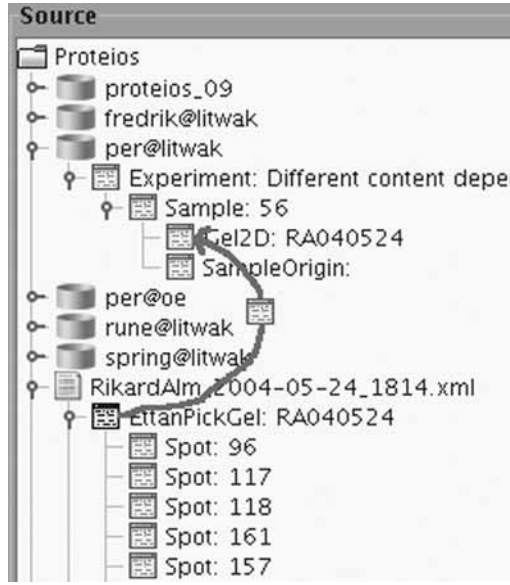


Fig. 9. Adding picking robot spots using drag and drop.

4.2. Usage Overview

The previous use case is intended to show how Proteios can be used in a specific proteomics experiment situation and at the same time give some indication of the strategy of the Proteios application. Describing all relevant use cases (e.g., ICAT, LC treatment of samples) goes beyond the scope of this text. However, it should be noted that typical use cases are semiautomatic. Proteios lets the user manually enter information where this is typically needed. Large data sets, which can be handled automatically, are imported as files. There is currently off-line support for both PIUMS ***(9)*** and VEMS ***(10)***, meaning that data from these applications can be imported. Proteios even shares its data model with VEMS, which means that data are interchangeable and Proteios can serve as a data repository for VEMS. Furthermore, Proteios can export MS information in the pkl format, which enables third party, e.g., Mascot ***(7)***, tools to use Proteios data. Search results can then be imported into Proteios.

4.3. Batch Handling

The same functionality as with the graphical user interface is also provided for batch processing. This is strongly related to automatic handling. One should be able to perform tasks without user interaction. Tasks may be repetitive and also one may wish to do something manually once to set up a pattern. A user

can define tasks completely independent of Proteios, save them for later usage, or extend them with further actions. The tasks are written in an XML file and can easily be reused or extended. Because task descriptions are validated before they are run it is difficult to go wrong. All commands that can be performed in the GUI can also be done through the batch handler.

As Proteios largely focuses on repository features and secondary experiment data analysis, one can do the tedious gathering of primary data as an overnight batch job. After such overnight work the experimenter refreshes his graphical view of Proteios and is all set to start working on secondary analysis.

4.4. Queries and Reports

Proteios allows users to browse data in the data repository in an intuitive way through the GUI. However, to make efficient use of the stored data, queries are supported. A user may be looking for specific data or needs to do a bulk export of data matching some criteria. Proteios currently has basic querying facilities that allow the user to retrieve data elements fulfilling specified criteria (e.g., protein hits from a certain gel and with a confidence level above a certain limit). In the future, Proteios will be extended with a plug-in interface enabling third party vendors to create analysis tools that interact with Proteios.

5. Outlook

5.1. A Proteios Server

In the very beginning Proteios was just a graphical tool to import some parts of a PEDRo file to a database. A graphical client has always existed. So far though, Proteios is a two-tier application. The server is one or many databases with the client being either the GUI or a batch program. When interacting with the Proteios core, both these clients have to do their own thread management, resource handling etc.

A Proteios server is a program running in the background for clients to interact with. Such a server will open up possibilities for external and third party programs to use Proteios online. The methods in the server's API will essentially be the same as the actions defined for the Proteios batch client. By developing a Proteios server we will step-by-step morph Proteios to a three-tier application. Or rather, Proteios will provide the necessary support for a third layer to interact with Proteios. Separating what current Proteios clients can do from what is the Proteios core will enforce the development of a more clearcut API. Eventually the existing GUI and the batch handling will use this server too.

It should be possible also for non-Java clients to interact with this server. A Proteios server will make it possible to build computational frameworks to integrate proteomic data with data from other types of experiments, such as

DNA sequencing efforts and DNA microarray studies. Applications such as BASE ***(14)*** can then easily interact with proteomics data.

5.2. Traceability: LIMS Functionality

Proteios is a free all-experiment encompassing data repository tool for proteomics, handling manually and automatically entered data. This makes Proteios a prime choice-independent LIMS system. There are already some LIMS-specific entities in Proteios. More will follow. Also, further connections can be created using XML namespaces and Proteios plug-ins.

5.3. Plug-Ins

Proteios has a very flexible data model and XML namespaces can further specialize its usage. The basic Proteios functionality though is given as it is compiled into the executable. It should be clear that functionality in this respect comes on two levels. The core Proteios functionality must be stable in order to provide a consistent application. On top of this core though, there are typical tasks which make sense to have user defined. Such tasks consist of imports, exports, addition (or withdrawal) of namespace restrictions, and analysis. Extra functionality should be possible to add dynamically, i.e., without recompilation, to a Proteios instance. Tasks in the form of small pieces of software are labeled plug-ins. The means for users to extend functionality is using Proteios plug-ins.

5.4. XML Namespaces

XML namespaces provide a means to specialize a generic model into a specific one. As Proteios is built on XML, namespaces are the perfect way to adapt to special needs. And, it can be done without any recompilation—just by letting data refer to a specific namespace. Namespaces can exist outside Proteios and completely independent of the application. Still, they are used for validation, so data will be checked on entering and leaving Proteios. Inside the application only much more generic basic types are used, e.g., “String” instead of something from an ontology. This makes it easy to set up Proteios to meet your specific needs.

5.5. Ontologies

An ontology is a hierarchic system of controlled vocabularies for data entities, with relationships between them. Each vocabulary is a catalog of the values that are assumed to exist and can be referred to by a data entity. A formal controlled vocabulary is specified by a collection of key-value pairs enabling the lookup of a value by its key. Referring to an ontology and supplying only the key enables the names themselves to be standardized and even to be changed, without altering the key itself. Ontologies are extremely important for data interchangeability and will be built into future versions of Proteios.

5.6. Interaction With External Tools

Often one will want to use the (Proteios) repository as input to external tools. This can be achieved in several different ways. Either one simply does it off-line, using data imports and exports. Or, one writes a tool-specific plug-in. If standard API:s for tools and equipment come up (we are not there yet), interaction can be readily done from within a standard Proteios installation. The most common cases are with protein and peptide matching database searches. However, in the longer perspective it is completely feasible to involve Proteios in the setting up of laboratory equipment. Proteios would be used not just to fetch data once equipment has been run, but already before equipment is run. One example is the ABI4700, which has an API to set it up. One could restrict (namespaces again!) the relevant elements in the MS part of Proteios to map nicely to the specific machine. The advantage lies at hand—settings are only written once and can directly be stored in the laboratory repository. Of course, one can also retrieve settings as well as experiment data once a machine has been run.

Acknowledgments

This work is in part supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium and by grants from FORMAS (22.6/2002-0042).

References

1. http://fuge.sourceforge.net/. Last accessed on 5/26/06.
2. Taylor, C., Paton, N. W., Garwood, K. L., et al. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.* **21,** 247–254. Application site: http://pedro.man.ac.uk/. Last accessed on 5/26/06.
3. Prince, J. T., Carlson, M. W., Wang, R., Lu, P., and Marcotte, E. M. (2004) The need for a public proteomics repository. *Nat. Biotechnol*. **22,** 471–472.
4. http://psidev.sourceforge.net/. Last accessed on 5/26/06.
5. PSI-MS: Mass Spectrometry Standards Working Group, http://psidev.sourceforge.net/ms/. Last accessed on 5/26/06.
6. http://sashimi.sourceforge.net/. Last accessed on 5/26/06.
7. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567. Application site: http://www.matrixscience.com/. Last accessed on 5/26/06.
8. http://fields.scripps.edu/sequest/index.html. Last accessed on 5/26/06.
9. http://www.hh.se/staff/bioinf/mass_spectro.html#PIUMS. Last accessed on 5/26/06.
10. Matthiesen, R., Lundsgaard, M., Welinder, K. G., and Bauw, G. (2003) Interpreting peptide mass spectra by VEMS. *Bioinformatics* **19,** 792–793. Application site: http://yass.sdu.dk. Last accessed on 5/26/06.

11. Bauer, C. and King, G. (2005) *Hibernate in Action*. Manning, Greenwich CT. Application site: http://www.hibernate.org. Last accessed on 5/26/06.
12. http://bioinformatics.icmb.utexas.edu/OPD/. Last accessed on 5/26/06.
13. Lundgren, D. H., Eng, J., Wright, M. E., and Han, D. K. (2003) An interactive bioinformatics tool for large-scale data exploration and knowledge discovery. *Mol. Cell. Proteomics* **2,** 1164–1176.
14. Saal, L. H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A., and Peterson, C. (2002) A platform for comprehensive management and analysis of microarray data. *Genome Biol.* **8,** SOFTWARE0003. Application site: http://base.thep.lu.se/. Last accessed on 5/26/06.

15

Proteomic Data Exchange and Storage

The Need for Common Standards and Public Repositories

Sandra Orchard, Philip Jones, Chris Taylor, Weimin Zhu, Randall K. Julian, Jr., Henning Hermjakob, and Rolf Apweiler

Summary

The ever increasing volumes of proteomic data now being produced by laboratories across the world have resulted in major issues in data storage and accessibility. The further demands of multilaboratory initiatives has highlighted issues when collaborators cannot import data generated within the same project but generated by different hardware types and processed by laboratory-specific work flows and analyses packages. There is an increasing need for common data standards that will allow the interchange of data between different instrumentation, search engines, and between laboratory databases. This could then lead to the establishment of data repositories from where benchmark datasets could be accessed and reanalyzed.

The Human Proteome Organization is currently supporting efforts to establish such standards. The work of the Proteomics Standards Initiative has lead to the development of the mzData XML interchange standard and is now broadening its scope to produce a spectral analysis output format, mzIdent. Accompanying controlled vocabularies allow the accurate, while systematic, representation of metadata throughout both schema.

Key Words: Proteomics; data standardization; protein interaction; mass spectrometry.

1. Introduction

The field of proteomics is increasingly broad and includes an expanding number of complex experimental techniques. Biological samples can be separated by several different methodologies, the best known being two-dimensional (2D)-gel electrophoresis and high-performance liquid chromatography. Component proteins can then be ionized and the resulting ions analyzed to give their mass. The ensuing spectral data is passed through an analytical pipeline, which should result in the identification of the proteins under investigation and may also yield

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

additional information as to their activation state, posttranslational modifications (PTMs), or interaction partners. Proteomic studies generate huge volumes of data, all of which poses an enormous problem of analysis, interpretation, and storage on the original researcher. High-throughput experimental techniques are rapidly becoming commonplace and the development of mass spectrometers is proceeding in parallel, such that the volume of data being generated by a single experiment will only increase. The ability of researchers to vary the type of instrumentation and analytical methodology used from experiment-to-experiment only adds to the problem. It is not unusual for a proteomics group to possess more than one type of mass spectrometer and vary their choice of machine with the type of experiment. This can create problems for data comparison even within a single laboratory because each instrument manufacturer typically provides its own proprietary program to analyze the output.

Publication of this data then presents a further challenge. Currently, a typical paper describing a proteomics dataset will consist of an experimental description, a summary data table, and an analysis and discussion of the data in that table. Additional supporting material is often supplied as a supplementary table, which is often little more than an expansion of the dataset published within the paper that may or may not be made public by the journal. Descriptions of the sample acquisition, preparation, and storage vary enormously, although these are crucial issues with direct impact on the final dataset and on the interpretation of the data by a third party. Even an issue as fundamental as protein identity can often be confusing, with author-derived nomenclature failing to coincide with that used by the public domain databases, often being misleading and causing confusion between unrelated proteins with similar, or identical, nonstandardized names.

The need for the development of common data representation standards in the field of proteomics was becoming apparent, with a requirement for a system that was both stable enough to provide a reliable platform for users, instrumentation manufacturers, and software developers to work from while flexible enough to keep pace with new developments in the field of both mass spectrometry (MS) and proteomics in general. This requirement became critical as large, publicly funded initiatives were being established in which the workload is spread across multiple laboratories that support different workflows, techniques, and instrumentation. The inability of the output from one piece of instrumentation to become the input to a second piece then becomes critical, hampering dataflow and resulting in an inevitable data loss. In order to address this problem, the Human Proteome Organization (HUPO) ***(1)*** set its internal bioinformatics committee the task of producing standards, initially to enable the transfer of information between laboratories participating in the HUPO tissue initiatives ***(2)***.

2. Standardizing Protein Identification and Description

One common issue across the entire field of proteomics is that of protein identification. Long-term data storage requires the use of stable protein identifiers. Protein names, gene names, and even sequences may change over time and unless these changes are tracked, data revisited after an elapsed period of time can prove to be relatively useless. One of the most significant developments with regard to protein sequence databases is the recent decision by the National Institutes of Health to award a grant to combine the Swiss-Prot, TrEMBL, and Protein Information Resource-Protein Sequence Database (PIR-PSD) databases into a single resource, UniProt (http://www.uniprot.org) ***(3)***. The Universal Protein Resource is a comprehensive catalog of data on protein sequence and function, maintained by a collaboration of the Swiss Institute of Bioinformatics (Geneva, Switzerland), the European Bioinformatics Institute (Cambridge, UK), and PIR, (Georgetown, US). UniProt is comprised of three components:

1. The expertly curated Knowledgebase (UniProtKB) which continues the work of Swiss-Prot ***(4)***, TrEMBL ***(4)***, and PIR ***(5)***.
2. The archive (UniParc), into which new and updated sequences are loaded on a daily basis.
3. The nonredundant databases (UniRef NREF) which facilitate sequence merging in UniProt and allow faster and more informative sequence similarity searches.

The UniProtKB is an automatically and manually annotated protein database drawn from translation of DDBJ/EMBL-Bank/GenBank coding sequences and directly sequenced proteins. Each sequence receives a unique, stable identifier allowing unambiguous identification of any protein across datasets. The UniProtKB also provides cross-references to external data collections, such as the underlying DNA sequence entries in the DDBJ/EMBL-Bank/GenBank nucleotide sequence databases, 2D polyacrylamide gel electrophoresis, and three-dimensional protein structure databases, various protein domain and family characterization databases, PTM databases, protein–protein interactions (IntAct) ***(6)***, species-specific data collections, variant databases, and disease databases. UniProtKB/TrEMBL contains a redundant sequence set, enriched by database cross-references and automatic annotation. Manual annotation of entries within UniProtKB/Swiss-Prot strives to augment each entry with as much information as is available, including the function of a protein, PTMs, domains and sites of importance, secondary and quaternary structure, similarities to other proteins, diseases associated with deficiencies in a protein, in which tissues the protein is found, pathways in which the protein is involved, and sequence conflicts and polymorphic variants. Sequences are merged within UniProtKB/Swiss-Prot to provide a single, nonredundant entry for a unique gene product from an individual organism.

The International Protein Index (IPI) ***(7)*** provides a top-level guide to the main databases that describe the human, mouse, and rat proteomes, namely UniProt ***(3)***, RefSeq ***(8)***, and Ensembl ***(9)***. IPI effectively maintains a database of cross-references between the primary data sources, providing minimally redundant yet maximally complete sets of human, mouse, and rat proteins (one sequence per transcript) and maintaining stable identifiers (with incremental versioning) to allow the tracking of sequences in IPI between IPI releases. This allows effective management of gene predictions, which vary with each release of both Ensembl and RefSeq. IPI thus provides a complete and nonredundant dataset for human, rat, mouse, zebrafish, *Arabidopsis* (with additional data from the *Arabidopsis* Information Resource ***[10]***), cow and chicken particularly suited to support protein identification in proteomics experiments.

3. Data Standardization and Retrieval

3.1. The Proteomics Standards Initiative

HUPO was formed in 2001 to consolidate national and regional proteome organizations into a single worldwide body. The Proteome Standards Initiative (PSI) was established by HUPO with the remit of standardizing data formats within the field of proteomics to the end that public domain databases can be established where all such data can be deposited, exchanged between such databases, or downloaded and utilized by laboratory workers ***(1)***. HUPO-PSI organized a series of meeting at which data producers, data users, instrumentation vendors, and analytical software producers gathered to discuss the problem. Because of the limited resources that could be dedicated to the effort, it was decided to concentrate initially on a few key areas, of which MS was of prime importance. Extensible markup language (XML) standards would be developed to assist in the transfer of data between different workstations and analytical platforms, and controlled vocabularies (CVs) or ontologies would be developed to give the user the flexibility to describe the specifics of a particular experiment within the framework supplied by the XML schema ***(11)***. Regular updates of the CVs will allow the terminology to remain current, as techniques advance and new methodologies are introduced, wheras any schema could remain relatively stable, allowing successful implementation by both software and hardware manufacturers.

3.2. The Proteomics Standards Initiative: Standards for MS

It was agreed early in the discussion process that the objectives of this group could best be achieved by aligning with the XML-based standard for analytical information exchange currently being developed by the American Society for Tests and Measures (ASTM) because both standards will have to describe

MS experiments and results. Standards for spectrometry data, such as the ASTM netCDF format (E1947-98 "Standard Specification for Analytical Data Interchange Protocol for Chromatographic Data," E2077-00 "Standard Specification for Analytical Data Interchange Protocol for Mass Spectrometric Data," ASTM International [*see* www.astm.org]) and the IUPAC JCAMP format (http://www.jcamp.org) have been successful because of broad vendor support and, being computer platform neutral, they have remained readable despite changes in computer technology. As useful as these standards are, it has proved difficult to keep them up to date owing to the very rapid changes in MS technology. XML was considered the best technology for allowing extensions to keep the standard up to date, while remaining computer platform neutral.

The PSI-MS schema is flexible enough to handle a diversity of experiments with a full range of experimental descriptors while still remaining compliant with the ASTM model. However, the encoding of peak list *m/z* and intensity values has been switched to Base64 in order to produce more compact files and work is currently in hand to broaden the specification to allow a full description of acquisition, to encompass both mass array and mass intensity. A stable version of the mzData format has now been released (http://psidev.sourceforge.net/ms) accompanied by the first release of the accompanying CVs. These are currently only available in list format but should be available in OWL format ***(12)*** from Autumn 2005. The base controlled vocabulary for mzData defines information items, which are shared across all experiments and instrument makers. All terms include a conserved accession number that is guaranteed to remain constant with the definition until deprecated allowing update of the associated text name without loss of information. The CV has been merged with that of mzXML developed by the ISB and distributed as mzXML.xsd. To ensure the ability to create PSI-MS submission documents from workflows using such schema, a unified controlled vocabulary was obtained by assigning HUPO accession numbers to all terms included in both the HUPO-PSI schema and the ISB XML schema. When complete, all fixed and variable attributes in mzXML will be mapped to corresponding HUPO-PSI ontology terms and both will be consistent with the IUPAC nomenclature for MS.

The published version of the PSI–MS XML data interchange format also gives access to tools that allow the user to both convert from MS text formats to PSI–MS XML format and view and browse stored data in PSI–MS XML format ***(13)***.

Acceptance of the mzData export format from the instrumentation manufacturers as a direct input to search engines has been good, and several vendors, for example, Matrix Science, Proteome Systems Ltd., and GeneBio already support the format, other companies such as Kratos Analytical Ltd., Thermo Electron, Waters Corp., and Bruker Daltronics will do so in their next releases. It has been proposed that the existing ASTM MS standard data dictionary be adopted, and

updated, for use as a controlled vocabulary within this model, with eventual ownership of this dictionary potentially passing to the American Society for Mass Spectrometry so that it could be used to support both the HUPO-PSI and the ASTM's raw data standardization efforts. mzData is now widely regarded as an acceptable format for representing MS data and will be valuable for both data exchange and reposition. In the long term, it is recommended that the ASTM standard be used for full raw data archiving when available (2005–2006); however mzData will remain the appropriate format for data exchange.

Currently in preparation is a spectral analysis output format, supporting a common syntax for peptide/protein identification and for protein modification description (analysisXML). The analysisXML standard is being designed to capture results from MS search engines and represent the input parameters for analysis algorithms, thus unifying results from different search engines. The requirements for mzIdent include the need to support the identification of both protein and peptides, by accession number or sequence, and must include the ability to describe modifications. Small molecule identification by either CML or SMILES must also be supported. The file format will include three major elements that are cross-referenced to each other—a molecular descriptor, an analysis results descriptor, and an analysis descriptor.

The molecule descriptor will describe the structural information of a molecule (descriptor, sequence, PTMs, structure), and references to the results. The relationship between molecules, for example, peptides as part of proteins will need to be clearly defined, allowing previous results to be used as source evidence. Analysis results could includes scores, spectrum annotation, reference to originating spectra/analysis, and to the matched molecule. The analysis descriptor may contain the name and version of the search engine and search parameters and can refer to a protocol to define a series of analyses. It will also cross-reference to the associated mzData file. Work is currently ongoing on the production of this standard and it is hoped that it will be published in draft format in 2005.

3.3. Posttranslational Modifications

The identification of PTMs, and an understanding of their functional significance is key to all areas of proteomics. Two existing databases already hold much of the required information in this field. The RESID Database of Protein Modifications ***(14)*** is a comprehensive collection of annotations and structures for protein modifications and cross-links including pre-, co-, and PTMs, concentrating on naturally occurring protein modifications. This database contains all the PTMs described within UniProtKB entries and is also cross-referenced in many other databases where an understanding of the state of a protein is or particular importance, for example, in the molecular interactions described within IntAct. However, MS has broader needs, in that, whereas a UniProt record for a protein

may contain all possible PTMs for any one protein, it cannot specify the particular state of a protein under a defined set of experimental conditions. Also, many of the modifications seen as a result of the process which a protein undergoes during the preparative and analytical procedures of MS are artifactual and not seen under natural conditions. UniMod is a database of protein modifications for use in MS applications and contains values for the mass differences introduced by both natural and artificial modifications ***(15)***.

A CV describing both natural and artificial modifications from MS processes will be required for the analysis and full exchange of proteomics data, and such a CV (P81-MOD) will be produced and maintained as part of the data standardization efforts.

3.4. Data Storage and Retrieval

With the development of these standards, the building of public repositories to hold such data is now possible. One such repository is Proteomics Identifications (PRIDE) ***(16)***. (www.ebi.ac.uk/pride) an open-source database that is available for the deposition and subsequent retrieval of proteomics data that is available in the public domain. This project implements an architecture that will run on any SQL-based database server, with configuration currently available for ORACLE and MySQL.

PRIDE will eventually hold the top-level data from the HUPO tissue initiatives, the plasma data being already available at the time of writing ***(17)***, as well as other data submitted by workers in the public domain, and is fully mzData compatible. Compatibility to analysisXML will be implemented as soon as this standard is published. The PRIDE database has been developed to provide the proteomics community with a public repository for protein and peptide identifications together with the evidence supporting these identifications. PRIDE holds details of PTMs including their relative positions on identified peptides. The data submitted to PRIDE can be submitted privately, for example prior to publication or as public data. Privately submitted data can be shared among collaborations of laboratories through the PRIDE collaboration system.

The PRIDE Java API now includes useful software tools to allow mzData files to be manipulated as Java objects and stored in a relational database. This software API can be used separately from PRIDE as required.

3.5. General Proteomics Standards

The context-sensitive nature of proteomic data necessitates the capture of a larger set of metadata than is normally required for genetic sequencing, where knowledge of the organism of origin will suffice. Not only is information of sample source, handling, stimulation, and eventual preparation for analysis required but also the detail of the analysis itself will need to be recorded.

For example, to compare images of 2D-gels knowledge of their mass and charge ranges are required, and this information will need to be retrieved by users wishing to perform meaningful analysis of this experiment. A standard representation of both the methods used and the data generated by proteomics are required to facilitate the analysis, exchange, and dissemination of proteomics data.

The Minimum Information About a Proteomic Experiment (MIAPE) guidelines are currently under development and are being designed to describe all the relevant data from any proteomics experiment, such as details of the experimenter, the sample source, the methods and equipment used, and all subsequent results and analyses ***(18)***. These could be, for example, data from a laboratory performing MALDI–MS on spots of interest from comparative 2D-gel electrophoresis or from a high-throughput screening facility using multidimensional liquid chromatography fed directly into a tandem mass spectrometer. A master document describes the metadata that needs to be collected to describe the overall process whereas a series of technology-specific documents give more detailed descriptions of particular areas, for example MS and MS Informatics. An XML format for data exchange will be derived from the General Proteomics Standards proteomics workflow/data object model, PSI-OM. This mark-up language (PSI-ML) is designed to become the standard format for exchanging data between researchers, and submission to repositories or journals. The object model must be flexible enough to cope with both rapidly evolving and completely novel technologies while fulfilling the immediate requirements of the scientists of today. As the name of the documents suggests, this process is analogous to the MIAME and MAGE-OM documentation and object model produced for the description of microarray data and much has been learned by close collaboration with workers in that field.

It is clearly desirable that public domain data such as that published in peer-reviewed journals should be accompanied by a defined set of information about the experiment and that all this information and the results obtained be deposited where it is available to all users. A MIAPE-compliant repository will contain sufficient information to allow users to, in principle, recreate any of the experiments stored within it and where possible, the information will be organized in a manner reflecting the structure of the experimental procedures that generated it.

It is intended that the object model would encompass, and utilize, exchange formats being developed by other groups sponsored by the PSI. For example, the mzData model for MS, and where possible common ontologies and vocabularies will be shared by all these varying domains and also developed in conjunction with the MGED consortium to describe common aspects of proteomic and microarray data. Together, the MIAPE guidelines, data model, ontologies, and various implementations will provide a sound base to describe proteomic experiments in their biological context.

4. Summary

Significant progress has already been made in improving the accessibility and utility of proteomic data, and to date, these efforts have been enthusiastically endorsed by the scientific community. Although these efforts are being coordinated by the HUPO-PSI, the work is being undertaken by a large body of scientists, representing the worlds of academia, industrial research, and instrumentation manufacture and it is to be hoped that they are laying the groundwork for common standards to be widely adopted throughout the entire user community. As these tools and models become more widely available it can be anticipated that they will play a major role in the direction that this important area of biology takes and the eventual utility of the data generated in increasingly high-throughput biology.

HUPO is also contributing to the establishment of large-scale datasets for a number of human tissues of particular interest in the pathogenesis of disease. The aim of these initiatives is to produce comprehensive lists of proteins present in normal plasma, liver, and brain, whereas identifying regional and racial variants within the population. All three project groups have acquired samples from large cohorts of donors, which are currently being analyzed in laboratories across the world by a standard set of procedures. The data generated will be deposited in the public domain using UniProt/IPI identifiers and HUPO-PSI data standards and will then be available to provide a reference dataset against which corresponding disease or drug-treated tissue can be compared ***(2)***.

The tools and standards required for the analysis of large-scale datasets generated by proteomic scientists working in the areas of drug discovery are either already available or will be released within the near future. Large reference datasets are being established and it is hoped that these will aid in the identification of a new generation of potential drug targets and assist in providing treatments for many of the life threatening or dehabilitating diseases that are common in man today. The development of standards that allows the transfer of data between different workstations and its eventual deposition into public repositories has been a major step forward in achieving this goal; the adoption of these standards by an ever-widening user community will ensure further distribution and sharing of data, and increase its importance as a research tool.

References

1. Orchard, S., Hermjakob, H., Binz, P. A., et al. (2005) Further steps towards data standardisation: the Proteomic Standards Initiative HUPO 3rd annual congress. October 25–27, 2004 Beijing, China. *Proteomics* **5,** 337–339.
2. Hanash, S. (2004) HUPO initiatives relevant to clinical proteomics. *Mol. Cell Proteomics* **3,** 298–301.

3. Bairoch, A., Apweiler, R., Wu, C. H., et al. (2005) The universal protein resource (UniProt). *Nucleic Acids Res.* **33,** 154–159.
4. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. **31,** 365–370.
5. Wu, C. H., Yeh, L. S., Huang, H., et al. (2003) The Protein Information Resource. *Nucleic Acids Res*. **31,** 345–347.
6. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., et al. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*. **32,** D452–D455.
7. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4,** 1985–1988.
8. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. **33,** 501–504.
9. Birney, E., Andrews, T. D., Bevan, P., et al. (2004) An overview of Ensembl. *Genome Res*. **14,** 925–928.
10. Rhee, S. Y., Beavis, W., Berardini, T. Z., et al. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.* **31,** 224–228.
11. Orchard, S., Montecchi-Palazzi, L., Hermjakob, H., and Apweiler, R. (2005) The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments. *Pac. Symp. Biocomput.* 186–196.
12. Aitken, J. S., Webber, B. L., and Bard, J. B. (2004) Part-of relations in anatomy ontologies: a proposal for RDFS and OWL formalisations. *Pac. Symp. Biocomput*. 166–177.
13. Pedrioli, P. G., Eng, J. K., Hubley, R., et al. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol*. **22,** 1459–1466.
14. Garavelli, J. S. (2004) The RESID Database of protein modifications as a resource and annotation tool. *Proteomics* **4,** 1527–1533.
15. Creasy, D. M. and Cottrell, J. S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics* **4,** 1534–1536.
16. Martens, L., Hermjakob, H., Jones, P., et al. (2005) PRIDE: The PRoteomics IDEntifications database. *Proteomics* **5,** 3537–3545.
17. Omenn, G. S., States, D. J., Adamski, M., et al. (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5,** 3226–3245.
18. Orchard, S., Taylor, C. F., Hermjakob, H., Weimin-Zhu, Julian, R. K. Jr., and Apweiler, R. (2004) Advances in the development of common interchange standards for proteomic data. *Proteomics* **4,** 2363–2365.

16

Organization of Proteomics Data With YassDB

Allan L. Thomsen, Kris Laukens, Rune Matthiesen, and Ole Nørregaard Jensen

Summary

In recent years the organization of mass spectrometry (MS) data obtained in large-scale proteomics projects became an important issue. This has catalyzed the development of a few different database schemes for storing MS data, as well as some dedicated user interfaces. However, many of these projects are still rather immature and often do not cover all needs. Because our needs were quite specific, it was necessary to build a database that accommodates all the major types of experiments generated in house and that could be easily extended by new modules made by collaborators or students. A database application named "YassDB" will be described in this chapter. The application is implemented in a "three-tier" application architecture, with a database layer, a middle layer consisting of web services and a client layer, containing the user interface. This offers high flexibility: it allows other applications, written in any language, to be written as clients to the database. The setup and use of the YassDB database application with two client programs "pProRep" and "VEMS" will be outlined.

Key Words: Database; organization; integration.

1. Introduction

With mass spectrometry (MS) as a downstream analytical technique, proteome analysis typically yields large and complex datasets. Proper organization and storage of the original data is important for further processing and essential to allow future re-evaluation or distribution as an electronic complement to a printed publication, in which case standardized data-formatting rules should be followed ***(1,2)***. Well organized data handling can serve as a first step toward building reliable biological knowledge databases, the next level in data integration ***(3)***.

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

Proteomic data can be stored in several ways. Data coming from different sources can be stored as separate files in the original format. This approach is straightforward and probably sufficient in small-scale proteome studies, but it does not retain the internal relationships between the data, and the organization of the data is entirely dependent on the person who produces it. A database system is more robust: the relational integrity of the data can be maintained, and standardized organization, analysis, searching, and management are possible.

A number of commercial systems to integrate, organize, and manage proteomics data in databases are available, generally labeled as "Laboratory Information Management Systems" or "LIMS." Besides their usually exuberant price tag, the code of these systems is often proprietary, limiting the potential to tweak them to certain needs. On the other hand, a number of open source projects are developing database schemes such as PEDRo ***(4)***, and database applications such as Proteios (*see* Chapter 14 ***[5]***) and PRIDE (*see* Chapter 15) to handle, organize, and distribute proteome data. Whether a database system is useful, however, depends on the client to access and manage it: a dedicated interface that reads original data files and transfers them to records in a series of database tables, and that is able to output data to a user readable format is needed.

This chapter presents an alternative approach to integrate data in a database system. Rather than configuring a dedicated client that directly interacts with the database, the interaction of a client with the database in this application is handled in a so-called "three-tier architecture" by web services: using a standard XML-based protocol that allows to query the PostgreSQL database in a simple way, from clients written in virtually any programming language from any machine connected to the internet. We will describe how this system, called "YassDB," can be set up and used in a typical proteomics environment, and show its particular advantages.

Its use with two client programs will be demonstrated. The first client is called "pProRep" ("php Proteome Repository"), and can be installed on a web server running "PHP" and offers a web-interface visualize proteome data. It was initially developed to interact directly with a relational database in MySQL, but a flexible database interaction layer now allows it to be configured as a client to other database types and web services, such as provided by YassDB. The second client is the "VEMS" program (*see* Chapter 7; ***6,7***), which can be run directly from the users computer and used for raw data processing, database searches, validation, and for quantify of peptide and proteins obtained from database searches.

2. Software

The software that was used for installing and building the database is listed next.

2.1. Required Software

For the YassDB database and web services:

1. Linux, Fedora Core 2 or newer. Download from: http://fedoralegacy.org/ or a newer version from http://fedora.redhat.com/.
2. Java 1.4 or newer, this might be included in the Linux distribution. Download JRE or JDK from: http://java.sun.com/.
3. Apache Jakarta Tomcat 5.5.7, download from: http://jakarta.apache.org/tomcat/index.html.
4. PostgreSQL 7.4, download from: http://www.postgresql.org/.
5. A JDBC driver to access the database. Download from http://jdbc.postgresql.org/. The driver should match the database version.
6. The web application (includes Axis). Download "axis.war" from: http: //yass.sdu.dk/yassdb/.
7. The database schema, download yassdb.backup, proteininfo.backup, and peptideinfo.backup from: http://yass.sdu.dk/yassdb/.
8. The webclient, download webclient.war from http://yass.sdu.dk/yassdb/.

2.2. For the pProRep Web Client

1. A web server running PHP5 (*see* **Note 1**). Nowadays PHP comes with most linux distributions, source code, and binaries for different platforms are available at www.php.net. The pProRep code, download the most recent version from http://www.ptools.ua.ac.be/pProRep/.
2. The pProRep—YassDB interaction layer and definitions. Download from http://www.ptools.ua.ac.be/pProRep/. Depending on which modules and extensions are used: the appropriate php libraries. Direct links to appropriate downloads can be found on the same website.

2.3. Optional Software

1. pgAdmin 3, download from: http://www.pgadmin.org/download.php.
2. Tomcat Admin Application. Download from http://jakarta.apache.org/tomcat/.
3. VEMS v3.0 (http://yass.sdu.dk).

2.4. Overall Design Considerations

The design and development of a database application involves a number of considerations. Its purpose and the available resource determine which technologies can be used. The YassDB database system was designed according to a number of requirements. The primary goal was to create a repository to store the diverse data from MS experiments, and organized these into projects. Functionality to import and export data from the protein identification search

engines used in the department was essential: VEMS (*see* Chapter 7) ***(7)*** and Mascot ***(8)*** are therefore supported.

2.5. Database Schema

The data schema ***(9)*** contains processed information from MS, LC–MS/MS, 2D-gel electrophoresis experiments, and metadata in the form of search result and quantitative information on the search result. The spectra are stored in the database as peak lists and only instrument settings and links to the raw data are stored in the database. The size of the raw data (typically 0.5–1.0 Gb) would quickly fill any affordable database server to the limited.

One of the goals when creating the database application was to avoid redundancy. Redundant data can easily clutter the search strategies by giving multiple search results, or require multiple searches, instead of only one. A redundant dataset complicates data validation. In addition, redundant data consumes more storage space and requires longer search times.

There are two strategies to avoid redundancy in a database. A control can be performed during data insertion by searching the database for existing data. Alternatively, the data can be just inserted and the redundancy check is done later, by a program that runs regularly and cleans the database from redundant information. For this application the first strategy appeared to be the best choice and was therefore chosen (*see* **Note 2**).

To avoid redundant data there are separate tables for peptide- and protein sequences (*see* **Note 3**). The central entity in the schema is the experiment. An experiment can contain multiple LC–MS/MS runs (*see* **Fig. 1**). The complex part of the schema consists of the connection of spectra, peptides, and proteins. There is a ternary relationship (*see* **Note 4**) from peptide data to processed data, peptide sequence, and proteins.

It is important that deleting experiments from the database occurs in a consistent way, without leaving any traces in the database. The database has been implemented in such a way that when an experiment is deleted from the experiment table, the rest of the database is updated accordingly. It should therefore never be necessary to delete information directly from other tables because of inconsistence. There can be other reasons to delete data, e.g., if a peptide assignment is found to be invalid then it must be possible to delete it separately. Additionally, the schema is defined in such a way that it does not allow deleting entries that are referenced in other tables.

2.6. Application Architecture

Flexibility was an important factor when this system was developed. This was achieved by implementing YassDB as a three-tier application (*see* **Fig. 2**). In a three-tier architecture, the clients access the database through a middle-layer, here implemented as web services ***(10)***. The middle layer controls access

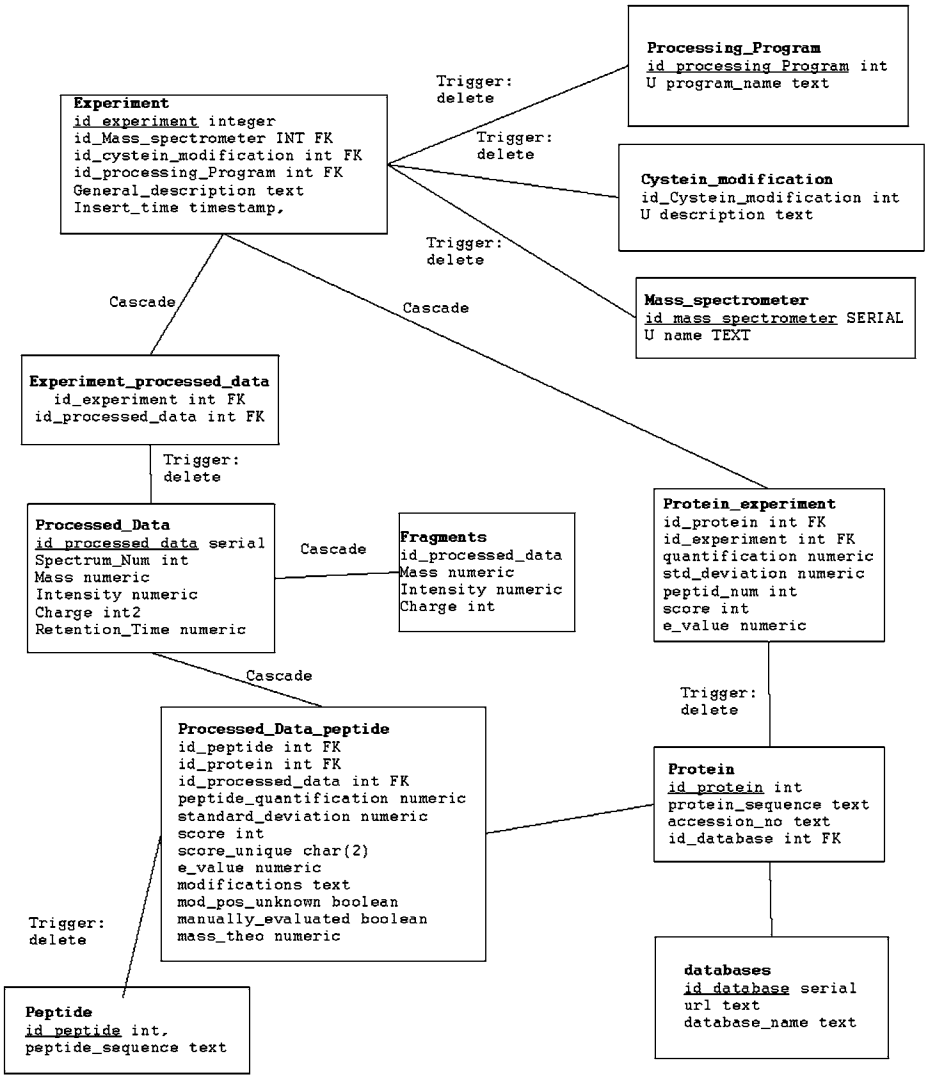


Fig. 1. The central part of the database schema. A complete view of the schema can be found here http://yass.sdu.dk/yassdb/downloads/databaseschema.pdf. There are two special notations: cascade and Trigger: delete. The meaning of "Cascade" is that when a row with a primary key is deleted this cascades to the table referencing the primary key with a foreign key. For example, when a row is deleted from Processed_data all rows in Fragments referencing id_processed_data is also deleted. The meaning of "Trigger: delete" is that when rows are deleted from a table there is a trigger in the database that delete from another table if there are no further references to some of the data. For example, when rows are deleted from Processed_data_peptide there is a trigger deleting any peptides, in the peptide table, that is no longer referenced. Together these two systems clean the database when an experiment or a part of an experiment is deleted.

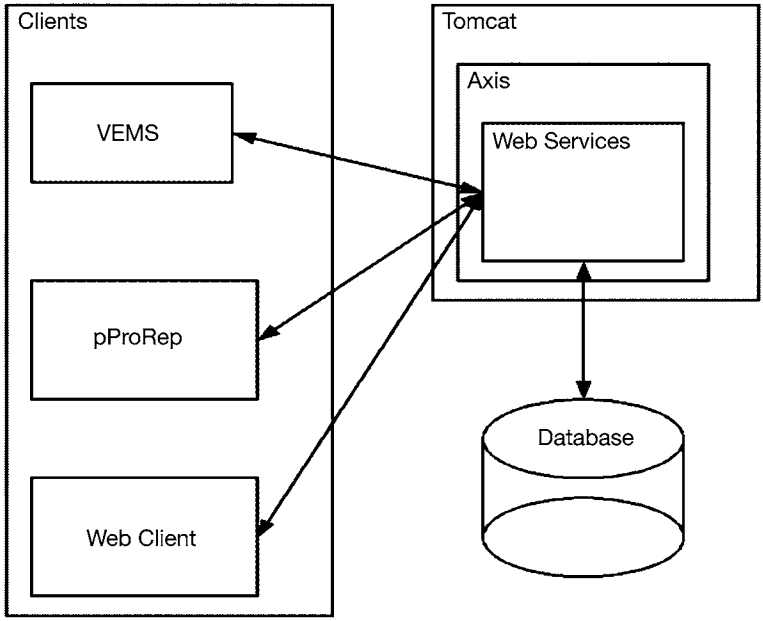


Fig. 2. An overview of the applications three-tier architecture, showing the communication links. The clients communicate only with the middle layer, which in turn communicates with the database. By keeping the layers separate it is easier to change the implementation of a layer, especially the client layer, without affecting the other layers.

and actions performed on the database. Web services are independent of platform and programming language. This gives the programmers the opportunity to easily create clients with new functionality (*see* **Note 5**).

2.7. Web Service Interface

The following methods have, at the time of writing, been implemented, and are accessible through the web services:

1. addProject(String name, String description): adds a project to the database assigned to the current user.
2. getProjects(): return all projects associated with the current user.
3. getExperimentIds(String project): returns all experiment identifiers associated with a given project.
4. getVemsExperiment(int id): returns a project in VEMS format as an attached file.
5. insertVemsExperiment(): used to insert experiments in VEMS-format, the experiment must be sent in an attached file. The method returns the identifier as soon as it is known, and then inserts the rest of the experiment.

6. experiment2Project(int id_experiment, int id_project): include an experiment in a project.
7. projectParticipant2Project(String participant, String project): include a user in a project.
8. getGelIds(): get identifiers for all gels.
9. getGels(int[] gel_id): returns all the gels associated with the identifiers in the request parameter, without images.
10. getGel(int id): returns a single gel, without image.
11. getGelImage(int id): returns the image of a certain gel as an attachment.
12. insertGel(Gel): inserts a gel, image as attachment, and returns the identifier, when inserted.
13. getSpots(int[] ids): returns data for the spots in the request parameter.
14. getSpotsForGel(): returns all spots for a given gel.
15. addSpots(): add a number a spots to a gel.
16. addSpot(): add a spot to a gel.
17. getFastaProteins(int experimentId): returns all proteins found in the given expriment, in Fastaformat.
18. getExperimentDescription(String projectName): returns all experiment descriptions for the named project.
19. getFragments(int experimentId, int spectrumNumber): returns all fragments from the given spectrum.
20. getVemsExperimentString: returns an experiment in VEMS text format embeded in the SOAP response.

It is basic functionality mainly associated with inserting and retrieving data from the database.

3. Installation

This section describes stepwise how to install and configure the different applications. For detailed installation instructions and command options see the appropriate documentation. The installation outline is based on a Fedora Core 2 system, but one should be able to use any other Linux system as long as the required software can be installed. Places where change to ip-addresses or passwords are necessary have been enclosed in square brackets [].

3.1. PostgreSQL

3.1.1. Step 1: Creating the Database

The PostgreSQL (*see* **Note 6**) database server must be installed, before continuing. First, the database should be installed by a user who is not root. In the following the username admin is used, if you use another please be sure to change all references accordingly.

Log into the system as admin and do the following:
Create a new database cluster with initdb (*see* **Note 7**):

```
mkdir data

initdb -D /home/admin/data   (see Note 8)
```

Create a directory for logging database events:

```
mkdir logs
```

3.1.2. Step 2: Configuring and Starting the Database

It is necessary to configure who is allowed to access the database and what authentication method should be used. This is done by editing pg_hba.conf file that is placed in /home/admin/data/. The following lines should be added:

```
local  all    all            trust
```

This gives access to all local users using socket connection without using authentication (*see* **Note 9**).

```
host   all    all   127.0.0.1      255.255.255.255 password
```

Giving access to all local users using TCP/IP, this is necessary for JDBC connections using password authentication.
The next line added is:

```
host   all    all   0.0.0.0        0.0.0.0       password
```

This gives access to all users **from** all ip-addresses with the use of password authentication. Start the database server:

```
postmaster -i -D /home/admin/data > /home/admin/logs/sessionlog

2>&1 & (see Note 10)
```

This command starts the database server, and sends all log messages and output to the file session log in the log directory created earlier.

3.1.3. Step 3: Installing the Database Schemas

Make sure that the database-server is started (*see* **Note 11**). Log in as admin and go to the directory where the three backup files are saved. Create a database:

```
createdb yassdb
```

Use pg_restore (*see* **Note 12**) to install the schema for YassDB, webclient and the userdb:

```
pg_restore -U admin -i -d yassdb -F c --no-owner yassdb.backup
```

pg_restore -U admin -i -d yassdb -F c --no-owner --data-only proteininfo.backup

pg_restore -U admin -i -d yassdb -F c --no-owner --data-only peptideinfo.backup

The last thing to do is to create two users, one YassDB user and one administrator for tomcat. The YassDB user is the one you use to login to the database and work with your experiments. He is created as follows, where any username and password can be used:

insert into users (user_name,user_pass) values('[username]','[password]');

insert into user_roles (user_name,role_name) values ('[username]','user');

insert into project_participants (db_user_name) values ('[usernname]');

The administrator is used to administer the webserver and must therefore be a member of both admin group and manager group (*see* **Note 13**). The admin user is created thus:

insert into users (user_name,user_pass) values('admin','[password]');

insert into user_roles (user_name,role_name) values ('admin','manager');

insert into user_roles (user_name,role_name) values ('admin','admin');

You also have to change the password for the "admin" user in the database:

ALTER USER admin PASSWORD '[password]';

It can be a bit of confusion regarding the admin users. There are three users who are all called admin, and each has his area of responsibility. There are one for the operating system, one for PostgreSQL and one for Tomcat. In our setup, they are in essence the same person, and therefore they are created with the same name and password.

3.2. Apache Jakarta Tomcat

Log in as root. Unpack Jakarta Tomcat (*see* **Note 14**) in the /usr/local directory ***(11)***. Then /usr/local/jakarta-tomcat-5.5.7 is the jakarta-home directory and all following file-paths will be relative to this directory, unless stated otherwise (*see* **Note 15**). As default Tomcat is using port 8080, but installed as a standalone server it should use port 80 (*see* **Note 16**).

3.2.1. Step 1: Configure Tomcat

Install the JDBC driver by placing the jar-file in /path/to/tomcat/common/lib. Configure the server by doing the following: open conf/server.xml and find

<Connector port="8080"

change it to

<Connector port="80"

Then you need to configure the authentication system. The system uses tables in YassDB to store names and passwords. To enable this you need to make the following changes to conf/server.xml:

Find and delete

<Realm className="org.apache.catalina.realm.UserDatabaseRealm" …… />

Add in the same place:

<Realm className="org.apache.catalina.realm.JDBCRealm"
debug ="99"

driverName="org.postgresql.Driver"

connectionURL="jdbc:postgresql://[server_name | ip]/yassdb"

connectionName="admin" connectionPassword = "[admin_password]"

userTable="users" userNameCol="user_name" userCredCol= "user_pass"

userRoleTable="user_roles" roleNameCol="role_name"

/>

At last the following lines should be added inside the <GlobalNaming Resources> tag to enable web service access from the webclient (*see* **Note 17**).

<Resource name="jdbc/users"

auth="container"

type="javax.sql.DataSource"

maxActive="50"

maxIdle="20"

maxWait="30000"

username="admin"

password="[admin password]"

driverClassName="org.postgresql.Driver"

url="jdbc:postgresql://[server_name | ip-address]/yassdb" />

3.2.2. Step 2: Starting Tomcat

Start tomcat by running the startup script (*see* **Note 18**):

/usr/local/jakarta-tomcat-<version>/bin/startup.sh

3.3. Web Service Configuration

To configure the application it is necessary to unpack the archive into am empty directory (*see* **Note 19**):

mkdir axis

cd axis

jar xf axis.war

This will unpack the web application into the current directory (*see* **Note 20**).

3.3.1. Step 1: Configuring Database Access

Open the file: META-INF/context.xml in an editor, you should see something like:

<Context reloadable="true">

<Resource

name="jdbc/postgresql" auth="container"

type="javax.sql.DataSource" maxActive="50"

maxIdle="20" maxWait="30000"

username="admin" password="[db_password]"

removeAbandoned="true" removeAbandonedTimeout="15"

driverClassName="org.postgresql.Driver"

url="jdbc:postgresql://localhost/yassdb" />

The url must be changed to point at your database, you must supply the password for admin to access the database and the docbase must be changed to reflect your tomcat version.

3.3.2. Step 2: Enable Logging

Create the path /home/share/logs, log in as root:

```
mkdir /home/share

mkdir /home/share/logs
```

3.3.3. Step 3: Deploy Web Services

Repack the application, if the current directory is the directory where it was unpacked, the web application can be repacked by

```
jar -cf axis.war *
```

Now the web services should be deployed. The easiest way to do this is to have the manager application installed. In a web-browser write:

```
http://server-name/manager/html
```

At the bottom of the page you find the deploy section which is divided into two. The last section titled “WAR file to deploy,” is the one to use. Click the browse button and find the newly repacked “axis.war” file. Click “deploy” and the web services should be running. If not there will be a “fail” message on top of the screen.

3.3.4. Step 4: Configure Webclient

Unpack the webclient into an empty directory

```
mkdir webclient

cd webclient

Jar xf webclient.war
```

Open META-INF/context.xml, and change the following references in the resource tag:

```
username=“admin”

password=“[password]”

url=“jdbc:postgresql://[your-url]/yassdb” />
```

Open WEB-INF/classes/log4j.properties and change the following line to fit into your system:

```
log4j.appender.R.File=/home/share/logs/webclient.log
```

Repack the webclient:

```
jar -cf webclient.war *
```

3.3.5. Step 5: Deploy Webclient

This is done in exactly the same way as deploying the webservices, using the webclient.war file.

After deploying the webclient you should be able to see a login page by entering the following link in a web-browser: http://localhost/webclient.

3.4. Configuring the pProRep Client Interface

3.4.1. Step 1: Prepare a Web Server Running PHP

PHP comes, together with the apache web server, in most typical Linux distributions (*see* **Note 20**). The correct parsing of php-pages can be quickly checked by uploading a text file, saved as "test.php," containing the following code, to the document folder of the web server:

```
<?

phpinfo();

?>
```

Upon pointing the browser to the corresponding url, an extensive description of the server PHP installation (including version info) is retrieved if the server parses PHP-files. If PHP5 (*see* **Note 1**) is not installed, you can download it from www.php.net.

3.4.2. Step 2: Install pProRep

1. Unpack the downloaded pProRep archive on your computer. Unpack the downloaded pProRep-YassDB interaction layer, and copy the content of the resulting folder to the unpacked pProRep folder.
2. Open install_YassDB.txt in the pProRep folder. This document contains the most recent and detailed instructions on how to configure pProRep with YassDB on a web server. Follow these instructions carefully. This consists of editing the relevant config-files in the "configuration" folder and adjusting the configuration settings. A few installation-dependent parameters need to be set correctly before pProRep is able to function: path-variables and the data-source need to be defined correctly. Other configuration settings are optional, but useful to adjust pProRep to certain requirements.
3. Upload the contents of the pProRep folder to the web directory on the web server.
4. Point a web browser to the corresponding url to verify the proper pProRep functioning.
5. The system can be further tweaked according to your requirements: needed modules can be activated, useful extensions can be added, and the appearance of the output can be adjusted (*see* **Note 22**). This is described in the pProRep documentation.

3.5. Starting YassDB Database Web Services From VEMS v3.0

VEMSv3.0 can interact with the YassDB through web services. Once VEMS and YassDB are installed and configured correctly all the Web services described in **Subheading 2.7.** can be used from the VEMS program. For installing VEMS (*see* Chapter 7). In the VEMS folder the file "Webservices.txt" contains the configuration file for web services. Open the file in notepad and edit the URL link so it references the directory containing the web services on the YassDB server. Now start VEMS and go to the "Web service" tab-page to start using the web services installed on YassDB.

4. Notes

1. Although PHP5 is recommended for running pProRep, it can be run as a YassDB-client with PHP starting from 4.2.2. In contrast to PHP 5.0, where native SOAP webservice functionality is provided, you need to install the nusoap library for older versions. You can download nusoap from http://sf.net/projects/nusoap/. Details can be found on the pProRep website. The same applies for the graphical extensions: GD2 is essential for dynamic image generation and manipulation in PHP: for example the output of gel-images by pProRep depends on it. It is available from www.boutell.com/gd/, but included in recent PHP versions. The output of the test.php script above gives information on whether gd2 is active. If not, check the instructions on the gd2 website and on www.php.net.
2. The disadvantage if this approach is that it takes longer to insert data. The alternative, where the tables are updated later, will lead to changes in the tables' primary keys, because some data, like peptide sequences, will be inserted independently of existing peptides and then later, when the same peptide sequence is found twice, one of the entries will be removed and all references updated. This can lead to confusion if a search is done between insert and update: a part of the search result will then be invalid because the primary key from some entries has changed.
3. Keeping the sequences separate also makes it easier later on to extent the information about a certain protein, for example with PubMed references. This information will then automatically be available to all researchers working on the same protein.
4. A ternary relationship is when a table has a relationship to three other tables. Here, we have the processed_data_peptide table that is related to the table's peptides, proteins, and processed_data.
5. A big advantage with web services is that the communication between clients and server is independent of the programming language used. The described system uses three different programming languages, and they all easily communicate with the web services. This offers the advantage that researchers writing programs that use the database can do that in the programming language they know, while taking advantage of the functionality already established in the web services.
6. The choice of database is actually a limited choice between MySQL and PostgreSQL, the two most popular free databases. They each have their advantages. MySQL offers a little better performance. On the other hand PostgreSQL

has implemented a larger part of the SQL standard, including nested queries, a good way to make some advanced queries, views and functions, which allow centralizing some standard procedures. The need for high performance, for example with many concurrent connections, did not outweigh the more advances SQL-implementation of PostgreSQL, thus a choice was made for the latter. However, the new MySQL v5.0, which is at the time of writing not yet fully released, will apparently implement a larger part of the SQL standard ***(12)***.

7. This is a database cluster in the sense that you can have multiple databases at this location, not in the sense that it consists of multiple machines.
8. If initdb is not found you have to add it to your path. If a standard installation of PostgreSQL has been made you can use: PATH=$PATH:/usr/local/pgsql/bin.
9. This is convenient, because it does not require a password to log into the database server from the local machine, and make changes to the database. It is obviously a security risk if the physical database server is easily accessible for many people.
10. The last part of this command “2>&1 &” is the Linux-way to redirect output-streams from a program. “2>” redirects the standard-error to standard-output (&1). The finishing “&” makes the program run in the background, so you get command-prompt back.
11. To check if the database is started write: ps -ef |grep “postmaster” at a console. If the database is started a line containing something like: “postmaster -i –D /usr/local/pgsql/data” should be seen.
12. pg_restore is a commandline tool for reestablishing databases and is included in the binary distribution of PostgreSQL.
13. The membership of these two groups allows the administrative user to access the admin and manager applications associated with tomcat. If you for some reason don’t install these then this user is not needed and this step can be omitted.
14. Tomcat is the official servlet reference implementation, and is therefore bound to have implemented all features in the servlet specification. It was chosen because it is easily integrated into the apache http server and extensible with the web service engine Axis.
15. In addition to the standard installation archive it is recommended to download the Admin-application as it is a convenient web-based administration tool for the tomcat installation.
16. Port 80 is the standard port for web servers, so using this makes it easier to access the pages.
17. There are three connection pools, in the setup procedure, all point to the same database, so it would be enough with one of them. This design however gives the flexibility to separate the databases at a later time, without compiling the source files, e.g., if you want to make a separate authentication system, or you already have one.
18. If you get errors about a missing environmental variable called JAVA_HOME you must create it. Assuming you use bash-shell, do the following add: export JAVA_HOME = /your/java/path/ to the .bashrc file in your home directory, and execute bash.
19. If the system cannot find the jar program try export PATH = $PATH;$JAVA_HOME/bin. This should add the jar program to your path.

20. There are two subjects that must be addressed when configuring the web-application that access a database. First, define how to access the database and second, define access to the application. In both cases, it is important that the system is flexible and independent of the code. Here, it is in both cases implemented as container managed security. In the case of access to the application, this means that the access is restricted by the server. It is not implemented as a fine grained control; the server gives access if a user exists in the system and is a member of any group. This means that any user who needs to be able to use the application must be defined in conf /tomcat-users.xml and must be a member of one or more groups. As for access to the database it was chosen that anyone who has access to the application has full access to the database, through the application.
21. The pProRep installation described is based on a standard Linux-Apache-php setup. Note that this server should not be necessarily the same machine as the one running the YassDB database and webservices, and they are assumed to be distinct servers in the description. Though running YassDB and pProRep simultaneously on the same server is possible, a different server setup is required, in which Tomcat is integrated with an Apache-server running PHP. Check the Apache (http://httpd.apache.org/) and Tomcat (http://jakarta.apache.org/tomcat/) webpages for more information.
22. Some of the more advanced modules are dependent on additional (external) PHP libraries. Because they evolve as well, and their usability is dependent on the PHP version, it is important to check the most recent pProRep module documentation, especially if certain features do not work as expected.

Acknowledgments

R. M. was supported by grants from EU TEMBLOR and by Carlsberg Foundation Fellowships. K. L. is a postdoctoral fellow of the Fund for Scientific Research–Flanders (Belgium) (F. W. O.-Vlaanderen). O. N. J. is a Lundbeck Foundation Research Professor and the recipient of a Young Investigator Award from the Danish Natural Science Research Council.

References

1. Orchard, S., Hermjakob, H., Julian, R. K., Jr., et al. (2004) Common interchange standards for proteomics data: public availability of tools and schema. *Proteomics* **4,** 490–491.
2. Garwood, K., McLaughlin, T., Garwood, C., et al. (2004) PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics* **5,** 68.
3. Patterson, S. D. and Aebersold, R. H. (2003) Proteomics: the first decade and beyond. *Nat. Genet.* **33,** 311–323.
4. Taylor, C. F., Paton, N. W., Garwood, K. L., et al. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotech.* **21,** 247–254.

5. Garden, P., Alm, R., and Hakkinen, J. (2005) PROTEIOS: an open source proteomics initiative. *Bioinformatics* **21,** 2085–2087.
6. Matthiesen, R., Lundsgaard, M., Welinder, K. G., and Bauw, G. (2003) Interpreting peptide mass spectra by VEMS. *Bioinformatics* **19,** 792–793.
7. Matthiesen, R., Bunkenborg, J., Stensballe, A., Jensen, O. N., Welinder, K. G., and Bauw, G. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* **4,** 2583–2593.
8. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **18,** 3551–3567.
9. Ramakrishnan, R. and Gehrke, J. (2003) *Database Management Systems*. McGraw-Hill, New York.
10. Deitel, H. M., Deitel, P. J., Gadzik, J. P., Lomelí, K., Santry, S. E., and Zhang, S. (2003) *Java Web Services for Experienced Programmers*. Prentice Hall, Upper Saddle River, NJ.
11. Chopra, V., Bakore, A., Eaves, J., Galbraith, B., Li, S., and Wiggers, C. (2004) *Professional Apache Tomcat 5*. Wrox, Hoboken, NJ.
12. Converse, T. and Park, J. (2004) *PHP5 and MySQL Bible.* Wiley, Hoboken, NJ.

17

Analysis of Carbohydrates by Mass Spectrometry

Kudzai E. Mutenda and Rune Matthiesen

Summary

The analysis method described in this chapter demonstrates the structural characterization of carbohydrates based on their molecular mass, as well as the mass of their respective fragment ions using mass spectrometry (MS). The carbohydrate molecules are first converted into gaseous ions, under vacuum, after which their mass-to-charge ratio is measured. The mass-to-charge ratio provides information on their preliminary identification, which is further elucidated by fragmenting the ions under a process of collision-induced dissociation. The masses obtained in the first stage of MS together with those obtained in the subsequent stages (MS^n) are combined into a mass list that is loaded into the program, Virtual Expert Mass Spectrometrist (VEMS) v3.0. The mass lists obtained are then used by VEMS to search a database of glycans to give the identity of the carbohydrate and the correct assignment of the fragment ions.

Key Words: Carbohydrates; electrospray ionization; ion-trap mass spectrometry; tandem mass spectrometry; saccharide fragmentation; VEMS.

1. Introduction

Carbohydrates play both a functional and a structural role in nature. Carbohydrates are involved in diverse roles such as molecular recognition, intra- and intercellular signaling, energy generation, protein conformation modification, as well as being structural components. In addition, carbohydrates are implicated in many disease states. Industrially, carbohydrates find application in the food, textile, paper, and pharmaceutical industries, just to mention a few. The analysis of carbohydrates is therefore essential in both understanding the role of these molecules in biology and to improve their industrial application.

The complexity of carbohydrates, in general, poses a big challenge in their analysis. Unlike other biomolecules, such as proteins and nucleic acids,

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

carbohydrates exhibit a higher structural complexity. The complexity arises from the occurrence of multiple isomers ***(1)*** and branching, which give rise to extreme heterogeneity. This is a consequence of there being more than one site available for linkage between the constituent monosaccharide units in a given carbohydrate. Analysis of carbohydrates by mass spectrometry (MS) is further challenging because of the isobaric nature of not only the parent molecules but also of the product ions on fragmentation. For example, it is difficult to distinguish between different hexoses that all have a monoisotopic mass of 162.05 Da. In addition, the unavailability of automated procedures for sequencing and synthesizing carbohydrates for further studies adds to the challenge of structurally analyzing carbohydrates.

The analysis of carbohydrates by MS provides information on molecular mass, constituent monosaccharides, sequence of the monosaccharides, linkage type, stereochemistry of the monosaccharide units, anomericity of the glycosidic bond, branching positions, type of branching, modifying groups, types of modifying groups, and the quantity. However, isomers are not readily distinguishable. Distinguishing isomers may require derivatization of the carbohydrate sample to obtain unique and diagnostic fragmentation patterns ***(2,3)***.

Tandem MS (MS^n) (multiple successive MS stages) is almost always obligatory to carbohydrate structural determination. Linkage type and the degree of branching can only be determined by MS^n. As described next, fragmentation of the carbohydrate molecule along the glycosidic bond provides sequence information, whereas cross-ring cleavages are essential for providing linkage and branching details. Ion-trap MS with the capacity for MS^n gives unambiguous assignment of fragments with similar masses from MS^{n-1}. Isobaric oligomers can be readily resolved by sequential trapping and fragmentation of isomer-specific ions ***(2,4)***.

The fragmentation of carbohydrates under MS/MS as described by Domon and Costello ***(5)*** gives rise to two types of cleavages, glycosidic bond cleavage and cross-ring cleavage. Four series of ions are generated through the cleavage of the glycosidic bond. B and C ions arise when the charge is retained on the nonreducing end. In B ions, cleavage occurs before the glycosidic oxygen and in C ions cleavage occurs after the glycosidic oxygen. Y and Z ions arise when the charge is retained on the reducing end. In Z ions, cleavage occurs before the glycosidic oxygen and in Y ions cleavage occurs after the glycosidic oxygen. In addition to the cleavage of the glycosidic bond, fragmentations within the saccharide ring unit also occur. Two series of ions are generated through cross-ring cleavage. A ions arise from cross-ring cleavage of the saccharide unit with charge retention on the nonreducing end, whereas in X ions the charge is retained on the reducing end. Superscripts are used to define the position of the ring cleavage (**Fig. 1**).

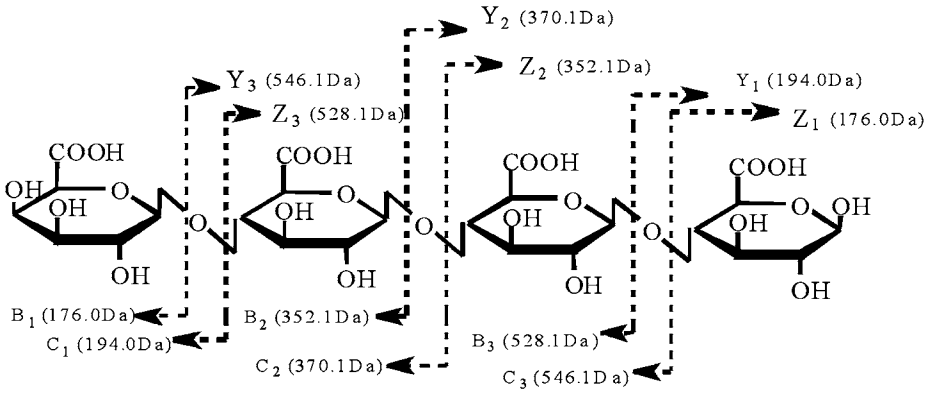


Fig. 1. Fragmentation pattern of glycans based on the Domon and Costello nomenclature ***(5)***. Note that in the example above, B and Z ions are isobaric, and C and Y ions are isobaric.

Fragmentation of carbohydrates during MS/MS is influenced by several factors including the ionization method used, ion analysis mode, adduct form, analyzer used, collision energy, and nature of derivatization (if derivatized).

1. The ionization method used. Depending on the ionization method one uses, the fragmentation pattern observed will differ. This mainly has to do with the "form" of the ionization method, whether it is a "soft" (ionization with virtually no unintended fragmentation) method or not. The electrospray ionization (ESI) ***(6)*** method is a "soft" ionization method and the examples reported in this chapter are from samples analyzed by ESI. In ESI, the sample is dissolved in an appropriate solvent and then placed in a needle. A high-voltage electric field is then applied to the tip of the needle resulting in an electrostatic spray of multiply charged droplets of the analyte. Following desolvation and the resulting charge concentration, gas-phase ions are produced. With ESI comes the possibility of MS^n with more than two stages. Structural information on carbohydrates is usually obtained by MS^3 or higher. This feature comes with ion-trap instruments. Theoretically, up to MS^{12} can be performed.
2. Ion analysis mode. B and Y ions are most abundant when the analysis is carried out in the positive ion mode. C and Z ions are most abundant in the negative ion mode.
3. Adduct form. MH^+ ions fragment more readily than MX^+ ions (X represents alkali metal cation). With MH^+ mainly glycosidic bond cleavages are observed. MX^+ ions produce more cross-ring fragments and hence more structural information. Structural information is obtained from the mass difference between the different ions of the same series. Branching positions, as well as modification sites, are deduced mainly from cross-ring cleavages.
4. Collision energy. The collision energy used in the fragmentation process influences the type of cleavage patterns observed. Low collision energies generally result in the cleavage of the glycosidic bond, whereas high collision energies lead

to cross-ring cleavages. A and X ions are mostly observed under high collision energies. In this regard, it means that low-collision energies give limited information on sequence, whereas more structural information can be obtained from cross-ring cleavage, a result of high-collision energies.

5. Derivatization. When carbohydrates are derivatized, the fragmentation pattern is different from the nonderivatized counterparts. Derivatization is not only useful for increasing the hydrophobicity of carbohydrates and thereby their ionization efficiency, but also for aiding in structural determination. Derivatized carbohydrates have a characteristic fragmentation pattern, giving diagnostic fragment ions ***(7,8)***.

2. Methods

The general approach for analyzing carbohydrates involves, first, the release of the carbohydrate moiety from the glycoconjugate, if it is conjugated. Two general methods are available for releasing carbohydrates from their conjugates: enzymatic methods and chemical methods. Enzymatic methods offer the advantage of preserving the conjugate, i.e., the conjugate is not degraded so it is available for analysis if required. Chemical release, which includes alkaline elimination and hydrazinolysis, preserves the carbohydrate moiety but degrades the conjugate. After release of the carbohydrate moiety, the next stage is separation of the carbohydrate moiety. This is not always necessary but it offers the advantage of reduced sample complexity. The analysis of carbohydrates by MS requires working with smaller units and if the released carbohydrate turns out to be too big, reduction to workable units can be achieved by either chemical or enzymatic digestion. In chemical digestion the glycosidic bond is hydrolyzed with acids. Depending on the choice of conditions, the hydrolysis will result in (1) complete cleavage whereby the constituent monosaccharide units are obtained, (2) partial, nonspecific cleavage in which random cleavage results in a mixture of oligosaccharides with varying numbers of monosaccharide units, or (3) partial, specific cleavage in which defined cleavage sites are targeted producing predictable oligosaccharides.

In enzymatic cleavage, specific enzymes are used to depolymerize the chain giving units that can be analyzed further. Two types of enzymes are available, exoglycosidases and endoglycosidases. Exoglycosidases cleave one monosaccharide unit at a time from the nonreducing end of the carbohydrate chain. Endoglycosidases cleave within the chain depending on their specificities. The structure of the carbohydrate sample, in terms of monosaccharide sequence and possible branching, can already be deduced from the enzymatic reaction.

Once the necessary size samples have been obtained, the carbohydrate is ready for structural characterization. The complete analysis of a carbohydrate sample involves the following:

1. Identification of the constituent monomers. The first stage involves the determination of the monomers making up the carbohydrate unit.

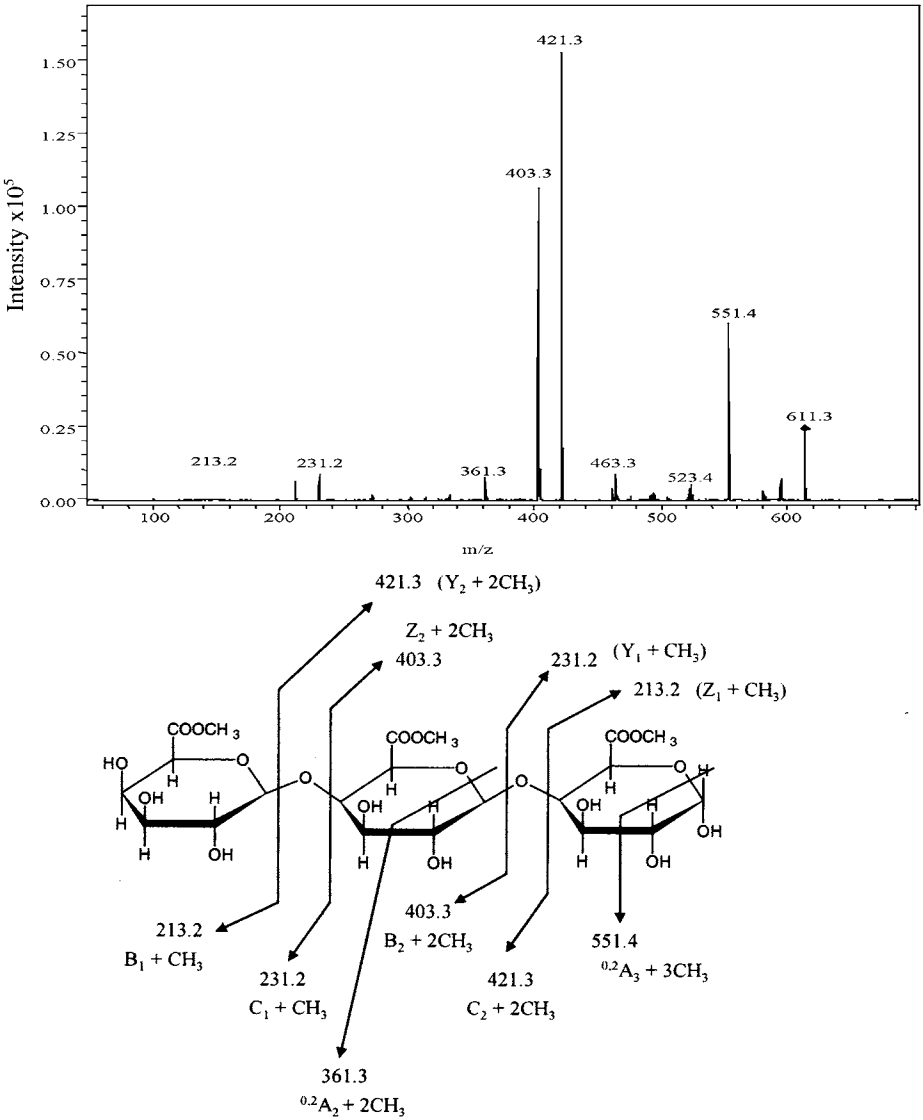


Fig. 2. Positive ion mode nanoESI MS^2 of a fully methyl esterified trimer and structural assignment of the trimer based on the fragmentation analysis with VEMS.

2. Sequence determination. The second stage involves the determination of the sequence of the monomers in the carbohydrate unit.
3. Linkage type. In the third stage, the task is to determine the type of linkage(s) in the carbohydrate unit.

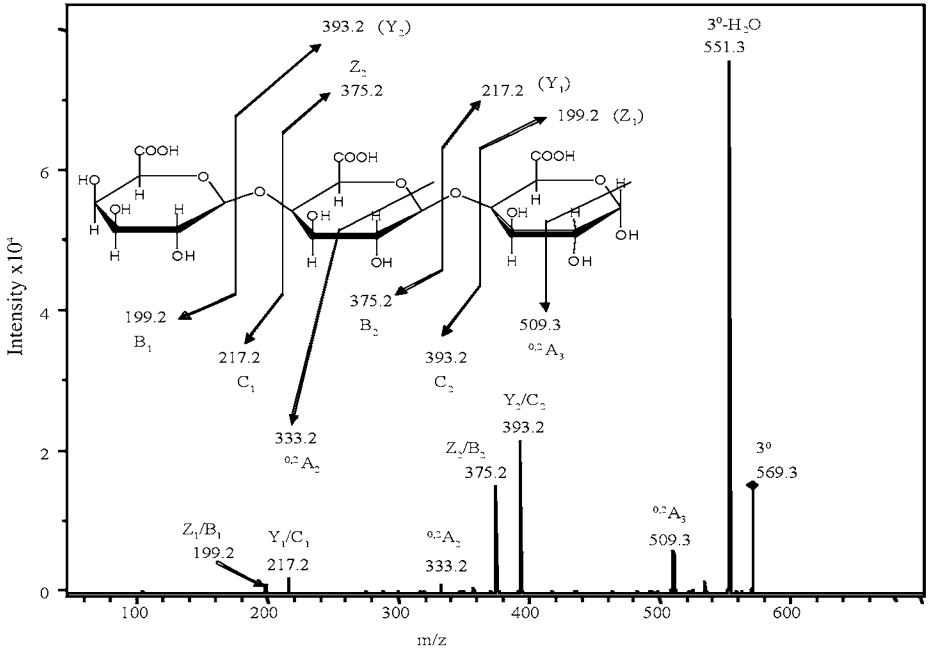


Fig. 3. Positive ion mode nanoESI MS^2 of a nonmethyl esterified trimer (the structural elucidation is shown in inset).

4. Stereochemistry of monosaccharide units. In addition to the linkage types, the stereochemistry of the monosaccharide units needs to be established.
5. Anomericity of the glycosidic bond. As well as the stereochemistry of the monosaccharide units, it is important to determine the anomericity of the glycosidic bond.
6. Modifying groups. Finally, if there are any modifying groups on the carbohydrate moiety, their identity should be revealed. For example, carbohydrate moieties can be modified by phosphate, sulfate, and other groups.

Figures 2–5 show structural elucidation of carbohydrate samples digested into smaller oligomers by enzyme treatment. The methyl esterified galacturonic acid samples were obtained by isolating the esterified oligogalacturonides after digesting a pectin sample with the enzyme pectin lyase A [E. C. 4.2.2.10]. The nonmethyl esterified samples were obtained by digesting polygalacturonic acid the enzyme polygalacturonase [E.C: 3.2.1.15].

The analysis of the mass spectra was as follows:

1. The peak list is load into Virtual Expert Mass Spectrometrist (VEMS) v3.0 (*see* Chapter 7 for installation of VEMS). Go to "File → Open data → Open multiple spectra." Choose the processed spectra file "7mer.pkl" and "6mer.pkl." Remember to click "Multiple," "Transfer," and close the window.

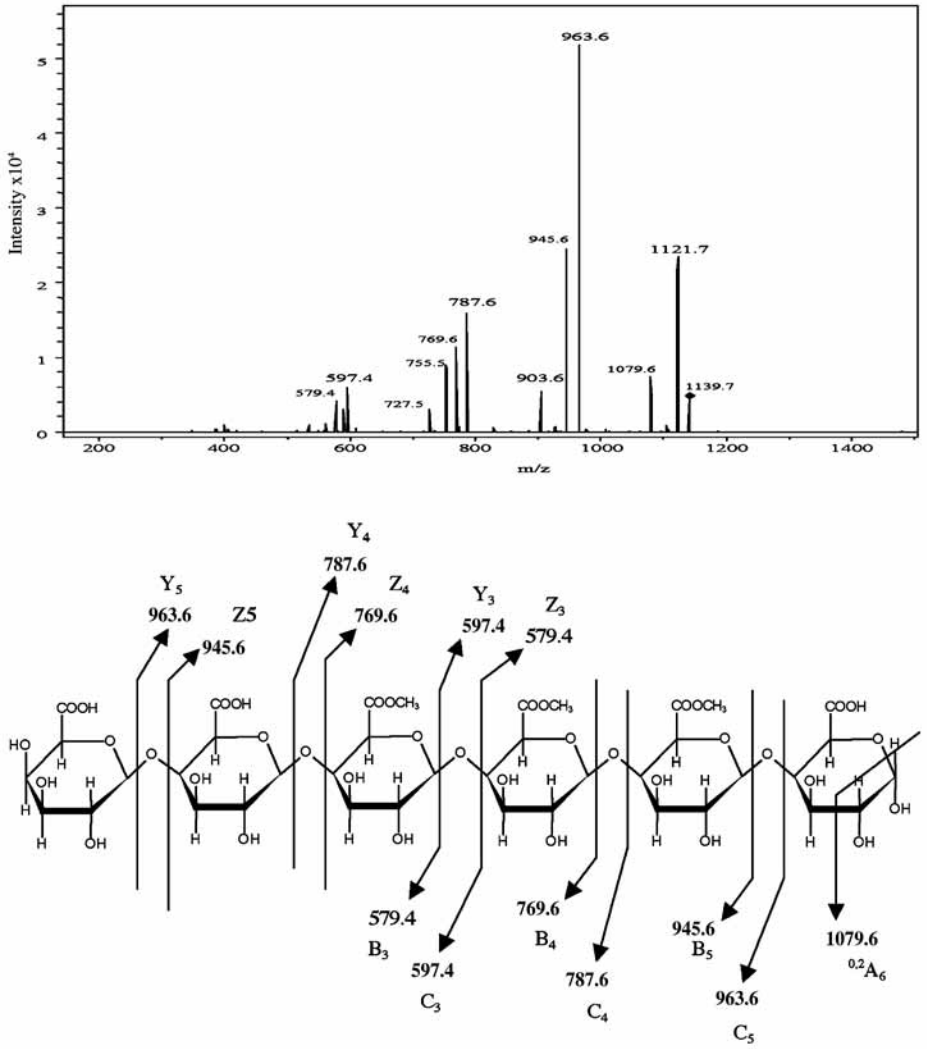


Fig. 4. Positive ion mode nanoESI MS^2 of a methyl esterified hexamer and structural assignment of the hexamer based on the fragmentation analysis with VEMS.

2. Go to the Tab-page “Tables.” Press the radiobutton “Glyco” to load a mass table for carbohydrate residues. Click on the button “Settings” and press the key “F2” to load variable modifications and default settings relevant for carbohydrates.
3. Go to the Tab-page “Databases” and specify a database of glycans. “Glycan.txt” (this file is by default installation located in the VEMS folder “Databases” and is therefore shown in listbox FASTA databases). Click on the file “Glycan.txt” and press the button “<<.” Click on the checkbox “Glycan” to specify that it is a glycan database.

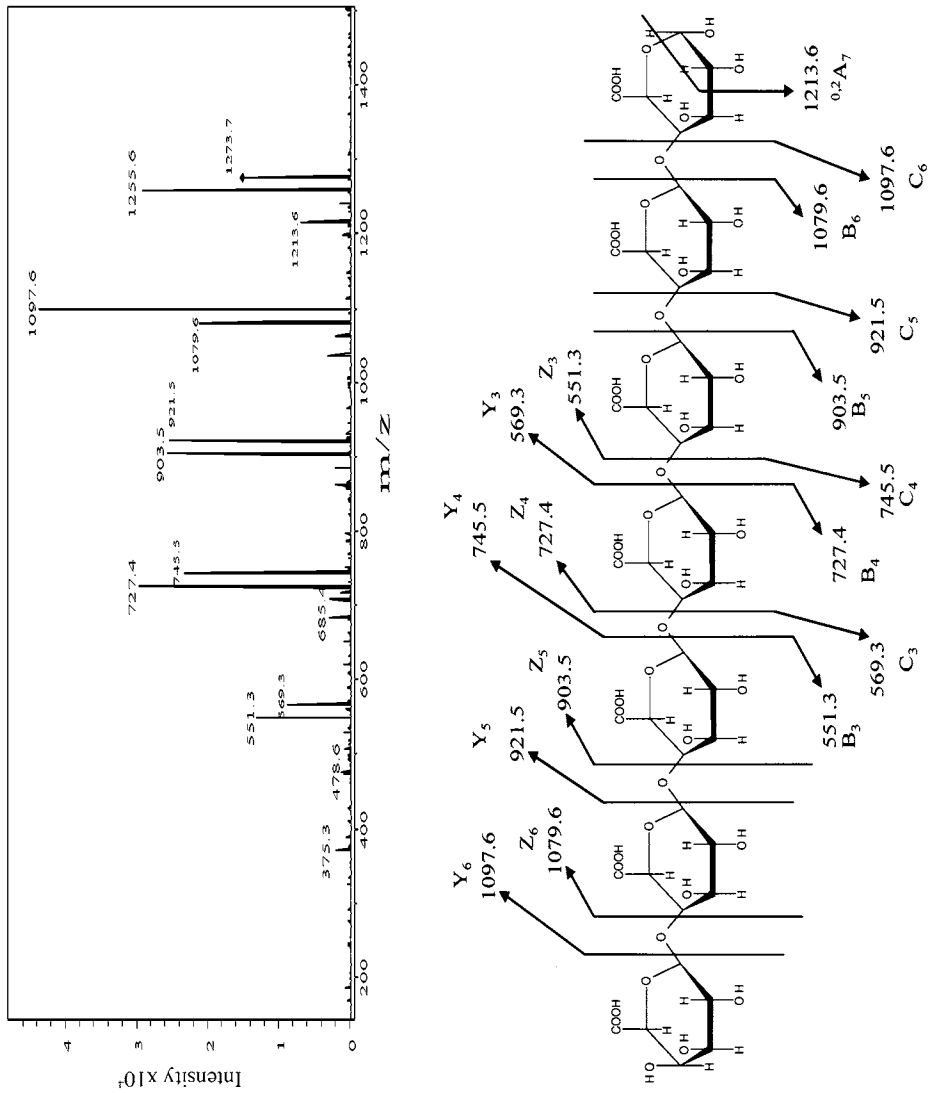


Fig. 5. Positive ion mode nanoESI MS^2 of a nonmethyl esterified heptamer and structural assignment of the heptamer based on the fragmentation analysis with VEMS.

4. Go to the Tab-page "Output" and click on "Start."
5. To visually validate the proposed structures click on a solution then right-click and choose "View spectrum." Click on the checkbox "a, b, and y" (*see* **Notes 1–3**) and right-click on the spectrum; this will give an automatic annotation of the peaks and an overlay of the theoretical masses (*see* **Note 4**).

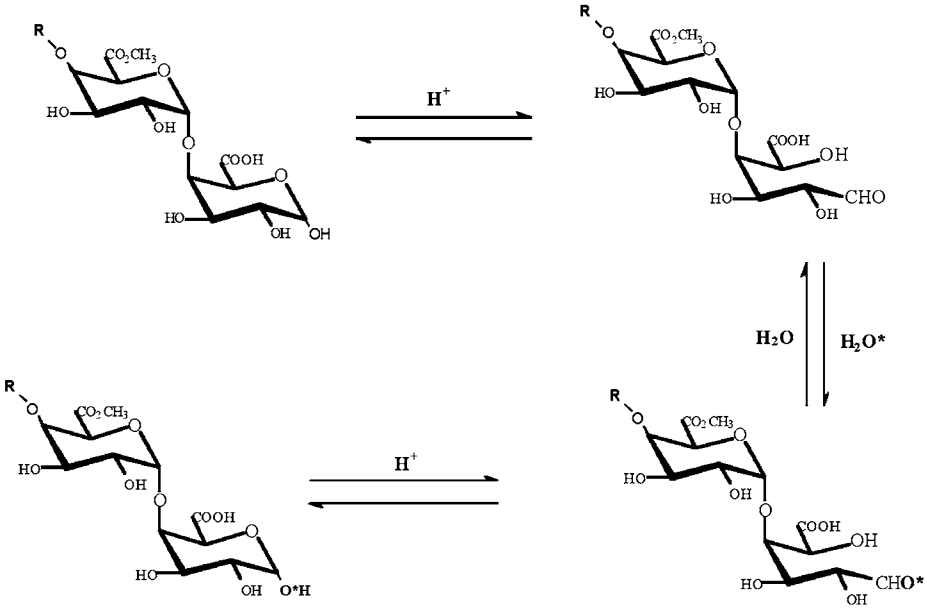


Fig. 6. The incorporation of ^{18}O into the reducing end during mutarotation.

6. On Tab-page “Settings” one can customize the view and annotation of the spectra. Make sure that the checkbox “Delete annotation on update” is unchecked. This will allow manual annotation of peaks. If the checkbox is checked then the manual changes will be deleted by the automatic annotation function in VEMS.
7. Go back to the Tab-page “Spectrum.” Press the space key until the radiobutton “Annotation” is checked or click on it. Now try to click on peaks in the spectrum. This brings up a window where manual annotation of the peak can be made.
8. The customized view and the spectrum annotation can be saved by right clicking in the bottom of the Spectrum window and choosing “Save spectrum.” For testing right-click again in the bottom of the page and choose “load spectrum.”

The characteristic fragmentation pattern of carbohydrates is shown in **Fig. 1**. It can be seen that isobaric fragment ions can be obtained, which will make distinguishing of the reducing end from the nonreducing end impossible. To overcome this we have used stable isotope labeling of the carbohydrate sample with $H_2{}^{18}O$ ***(9–13)***. This means that the ^{18}O is only incorporated into the reducing end during mutarotation (ring opening and closing) as shown in **Fig. 6** (*see* **Note 5**).

Under acidic conditions ring opening is accelerated. The aldehyde form, the open ring form, can form a hemiacetal, hence incorporating ^{18}O if the sample

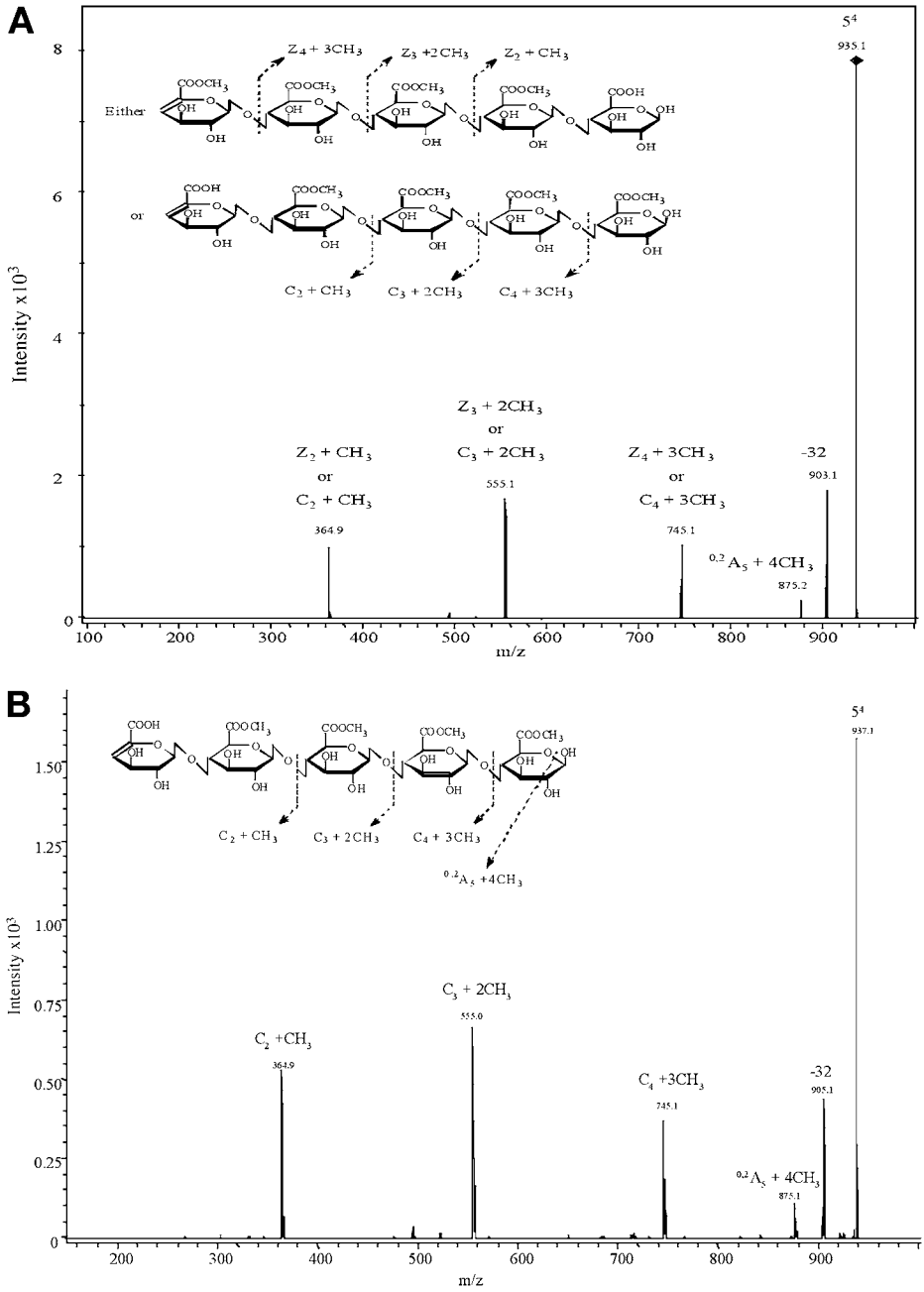


Fig. 7. **(A)** Negative ion mode MS^2 spectrum of a pentamer with four CH_3 groups (*m/z* 935.1). **(B)** Negative ion mode ESI MS^2 of an ^{18}O-labeled pentamer with four methyl groups. (**A**, before labeling and **B**, after labeling.)

is incubated with $H_2{}^{18}O$. Under MS analysis the ion with the reducing end residue has a 2-Da mass increase and no mass change is observed on the ion with the nonreducing end. Thus, any fragment ions with a mass of 2 Da more than the expected will, therefore, be carrying the reducing end residue. Again, using the steps outlined previously in VEMS, a 5^4mer was correctly assigned as shown in **Fig. 7**.

4. Notes

1. Cation adducts. When carbohydrates ionize, irrespective of which method is used, they do so not only by losing/gaining proton(s), but also by binding metal cations. In the positive ion mode they ionize by protonation as $(M + nH^+)^{n+}$ or as $(MX)^+$, (where X is an alkali metal). The relative affinities of X for M is Cs>K>Na>Li>H ***(7,14)***. It is therefore important when interpreting the spectra to take this into consideration. Usually one would look for either the ions $(M + nH^+)^{n+}$ or the ions $(M - nH^+)^{n-}$ in which case $(M + X^+)^+$ would not be considered. Further complication is seen especially with acidic carbohydrates where they lose/gain proton(s) and at the same time adduct metal cations as in $(M + X^+)^+$ and $(M + X^+ - 2H)^-$. This may pose problems when using programs that assume ionization to be only proton gain/loss. With VEMS this can be adjusted by choosing variable modifications.
2. Additional modifications. Care should be taken in interpreting spectra of carbohydrate samples that have the potential of carrying additional modifying groups. For example, oligosaccharides obtained from pectin, not only do they have methyl ester groups but also acetyl groups. This may give rise to false assignment. In cases like these more stages of MS will be necessary to clear the ambiguity.
3. Water loss dominance. Depending on the conditions used during collision-induced dissociation, fragment ions that have lost water are widely observed. These ions provide no structural information and they can be the major, most intense ones in a majority of the cases. One should select for the ions that are a result of water loss to be fragmented further, and this is a feature one has on an ion trap-type instrument.
4. The nomenclature used for the checkbox is “a, b, and y,” which is used for peptide ion fragments. However, the nomenclature for glycan fragments uses capital letters A, B and Y. In the VEMS spectrum viewer the checkbox was simply reused.
5. ^{18}O labeling. The labeling conditions may result in acid hydrolysis of, for example, methyl ester groups and thus a compromise is required between speed of ^{18}O exchange and loss of methyl ester groups. The higher the acid concentration the higher the exchange rate and the more the methyl ester hydrolyzes. Körner et al. ***(12)*** used 0.5% formic acid for 2 d at room temperature and obtained over 90% labeling and no significant hydrolysis. Labeling can be achieved by either (1) incubating a digested carbohydrate sample in acidic $H_2{}^{18}O$, or (2) performing the digestion in $H_2{}^{18}O$ medium. Hydrolyzing enzymes like polygalacturonase incorporate ^{18}O in the products when the reaction is carried out in $H_2{}^{18}O$-buffered medium. A combination of (1) and (2) can be used to distinguish products of polygalacturonase from those of a nonhydrolyzing enzyme like pectin lyase if combined digests are

analyzed. ^{18}O labeling is a prerequisite for sequencing. Labeling is essential for distinguishing Z and C ions in unsaturated oligomers and Y and C, and Z and B ions in saturated oligomers.

References

1. Laine, R. A. (1994) A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 X 10^{12} structures for a reducing hexasaccharide: the *Isomer Barier* to development of single-method saccharide sequencing or synthetic systems. *Glycobiology* **4,** 759–767.
2. Gaucher, S. P. and Leary, J. A. (1998) Steroechemical differentiation of mannose, glucose, galactose, and talose using zinc (II) diethylenetriamine and ESI-ion trap mass spectrometry. *Anal. Chem.* **70,** 3009–3014.
3. Gaucher, S. P. and Leary, J. A. (1999) Determining anomericity of the glycosidic bond in zinc (II)-diethylenetriamine-disaccharide complexes using MS^n in a quadrupole ion trap. *J. Am. Soc. Mass Spectrom*. **10,** 269–272.
4. Desaire, H. and Leary, J. A. (2000) Utilization of MS^3 spectra for the multicomponent quantification of diastereomeric N-acetylhexosamines. *J. Am. Soc. Mass Spectrom*. **11,** 1086–1094.
5. Domon, B. and Costello, C. E. (1988) A systematic nomeclature for carbohydarate fragmentations in FAB-MS/MS spectra of glyconjugates. *Glycoconjugate J.* **5,** 397–409.
6. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71.
7. Cancilla, M. T., Penn, S. G., Carroll, J. A., and Lebrilla, C. B. (1996) Coordination of alkali metals to oligosaccharides dictates fragmentation behavior in matrix-assisted laser desorption ionization/Fourier transform mass spectrometry. *J. Am. Chem. Soc*. **118,** 6736–6745.
8. Cancilla, M. T., Wong, A. W., Voss, L. R., and Lebrilla, C. B. (1999) Fragmentation reactions in the mass spectrometry analysis of neutral oligosaccharides. *Anal. Chem*. **71,** 3206–3218.
9. Hofmeister, G. E., Zhou, Z., and Leary, J. A. (1991) Linkage position determination in lithium cationized disacchaides: tandem mass spectrometry and semiempirical calculations. *J. Am. Chem. Soc*. **113,** 5964–5970.
10. Asam, M. R. and Glish, G. L. (1997) Tandem mass spectrometry of alkali cationized polysaccharides in a quadrupole ion trap. *J. Am. Soc. Mass Spectrom.* **8,** 987–995.
11. Viseux, N., de Hoffmann, E., and Domon, B. (1997) Structural analysis of permethylated oligosaccharides by electrospray tandem mass spectrometry. *Anal. Chem.* **69,** 3193–3198.
12. Körner, R., Limberg, G., Christensen, T. M. I. E., Mikkelsen, J. D., and Roepstorff, P. (1999) Sequence of partially methyl-esterified oligogalacturonates by tandem mass spectrometry and its use to determine pectinase specificities. *Anal. Chem.* **71,** 1421–1427.

13. Mutenda, K. E., Korner, R., Christensen, T. M., Mikkelsen, J., and Roepstorff, P. (2002) Application of mass spectrometry to determine the activity and specificity of pectin lyase A. *Carbohydr. Res.* **337,** 1213–1223.
14. Mohr, M. D., Bornsen, K. O., and Widmer, H. M. (1995) Matrix-assisted laser desorption/ionization mass spectrometry: improved matrix for oligosaccharides. *Rapid Commun. Mass Spectrom.* **9,** 809–814.

18

Useful Mass Spectrometry Programs Freely Available on the Internet

Rune Matthiesen

Summary

The intention with this chapter is to give an overview of a broad range of freely available programs on the internet which are useful for analyses of mass spectrometry data. The list is by no means a full list of free proteomics tools on the net and I apologize if there are other good tools on the internet that have been missed. The presented programs should cover the needs for the most general tasks in data analysis of proteomics data and have to a limited extend been tested. The links provided in this chapter will over time become invalid. In such cases it is worth while to try a World Wide Web search using the program packages names. This will often reveal the updated links. Some of the links presented in this chapter will also be maintained at http://yass.sdu.dk.

Integrated Proteomics Applications for Database-Dependent Search, Quantitation, and Data Storage

- Seattle Proteome Center (SPC) Proteomics tools (http://tools.proteomecenter.org/).
- VEMS v3.0 (http://yass.sdu.dk/) (Chapter 8).

Database-Dependent Search Programs

- Global Proteome machine (http://www.thegpm.org/).
- GutenTag (http://fields.scripps.edu/).
- Probid (http://projects.systemsbiology.net/probid/).
- Mascot (http://www.matrixscience.com/search_form_select.html).
- Pepsea (http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html).
- Sherpa (http://www.hairyfatguy.com/Sherpa/).
- OMSSA (http://pubchem.ncbi.nlm.nih.gov/omssa/index.htm).
- X!Tandem (http://www.thegpm.org/TANDEM/index.html).

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

- P3 (http://www.thegpm.org/).
- Inspect (http://peptide.ucsd.edu/).
- Phenyx (http://www.phenyx-ms.com/).
- PepHMM (http://msms.cmb.usc.edu/PepHMM/PepHMM.htm).

De Novo Sequencing

- Lutefisk (http://www.hairyfatguy.com/Lutefisk/).
- Inspect (http://peptide.ucsd.edu/inspect.py).
- PepNovo (http://www-cse.ucsd.edu/groups/bioinformatics/).
- OpenSea (http://medir.ohsu.edu/~geneview/).
- De Novo peaks (demo version) http://www.bioinformaticssolutions.com/products/PEAKSStudio/.
- http://proteome.sharcnet.ca:8080/help.htm#spider.
- http://dove.embl-heidelberg.de/Blast2/msblast.html.
- PepHMM (http://peptide.ucsd.edu/).

Programs for Quantitative Proteomics

- MSquant (http://www.pil.sdu.dk/silac_msquant.htm) (Chapter 15).
- SAM (http://www-stat.stanford.edu/~tibs/SAM/).

Intensity-Based Quantitation

- MSight (http://www.expasy.org/MSight/).
- MSGraph (http://homepage.sunrise.ch/mysunrise/joerg.hau/sci/).
- MapQuant (http://sourceforge.net/projects/mapquant/).

Glycomics

- http://au.expasy.org/tools/glycomod/glycanmass.html.
- VEMS (Chapter 16).

Other Useful Tools

- IsoPro (http://members.aol.com/msmssoft/).
- Isotopic distribution calculator (http://www.chemcalc.org/).
- VEMSiso (http://yass.sdu.dk) (Chapter 2).
- MAGTRAN (http://www.geocities.com/SiliconValley/Hills/2679/magtran.html).
- MassAnalyzer (http://www.geocities.com/SiliconValley/Hills/2679/magtran.html).
- http://prospector.ucsf.edu/.
- MassXpert (http://frl.lptc.u-bordeaux.fr/website-frl/massxpert/massxpert-main.html) tool for calculation of fragment masses.
- Peakerrazor (Chapter 4).
- OpenMS (http://open-ms.sourceforge.net/).

Data Storage and Exchange Formats

- Pride (http://www.ebi.ac.uk/pride/).
- Proteios (http://www.proteios.org/).

- Seattle Proteome Center (SPC) Proteomics tools (http://tools.proteomecenter.org/).
- CPAS (https://proteomics.fhcrc.org/CPAS/Project/home/home.view).

Useful Databases

- http://www.abrf.org/index.cfm/dm.home.
- http://www.unimod.org/.
- KEGG (http://www.genome.jp/kegg/).
- PathDB (http://www.ncgr.org/pathdb/).
- OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM).
- cMAP (http://cmap.nci.nih.gov/PW).
- GO (http://www.geneontology.org/).

Information on Mass Spectrometry

- http://www.ionsource.com/.
- http://www.i-mass.com/guide/movie.html.

Bioinformatics Tools

- http://www.r-project.org/.
- Taverna http://taverna.sourceforge.net/.
- Bioedit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).
- SATv1.0 (http://www.yass.sdu.dk/SAT/SATv1.0) (Chapter 10).
- http://www.systemsbiology.org/.
- www.bioexchange.com/tools/.
- http://bioinformatics.icmb.utexas.edu/OPD/.
- http://www.peptideatlas.org/.

Appendix

A. Abbreviations

BSA	Bovine serum albumin
BIRD	Blackbody infrared dissociation
CE	Capillary electrophoresis
CAD	Collision-activated dissociation
CID	Collision-induced dissociation
DE	Delayed ion extraction
DHB	2,5-Dihydroxybenzoic acid
ECD	Electron capture dissociation
ETD	Electron transfer dissociation
ESI	Electrospray ionization
ESSI	Electrosonic spray ionization
ESTs	Expressed sequence tags
FAB	Fast-atom bombardment
FTICR	Fourier transform ion cyclotron resonance
HCCA	α-Cyano-4-hydroxy cinnamic acid
HED	High-energy dynode detector
HPLC	High-performance liquid chromatography
IRMPD	Infrared multiphoton dissociation
IT	Ion trap
LDI	Laser desorption ionization
LC	Liquid chromatography
MALDI	Matrix-assisted laser desorption/ionization
MCP	Microchannel plate detector
MS	Mass spectrum/spectra
MS	Mass spectrometry
MudPIT	Multidimensional protein identification technology
PD	^{252}Cf plasma desorption
PMF	Peptide mass fingerprinting
PMM	Peptide mass mapping
PSD	Post source decay

From: *Methods in Molecular Biology, vol. 367: Mass Spectrometry Data Analysis in Proteomics*
Edited by: R. Matthiesen © Humana Press Inc., Totowa, NJ

RP	Reversed-phase
SA	Sinapinic acid
SELDI	Surface-enhanced laser desorption/ionization
SDS	Sodium dodecyl sulfate
SID	Surface-induced dissociation
TFA	Trifluoroacetic acid
THAP	2,4,6-Trihydroxyacetophenone
TOF	Time-of-flight
TPCK	*N*-tosyl-L-phenyl chloromethyl keton

B. Polynomial Calibration of Mass Spectra by VEMS v2.0

There exist three methods for calibrating mass spectra. The methods are termed external calibration, internal calibration, and recalibration.

External calibration uses standards with known masses to calibrate the mass spectrometer. The standards are run as separate experiments in addition to the analysis of the sample. External calibration is the least accurate calibration method. External calibration is especially inaccurate in use with MALDI. The inhomogeneous crystallization of matrix and sample is the main reason for this inaccuracy. Because the external calibration is run as a separate experiment, it is also affected by variation in the temperature of the flight tube.

Internal calibration also uses standards with known masses but here they are mixed with the sample and their mass spectra recorded together with the sample. For internal calibration, problems with overlapping mass peaks from the sample and the standards can occur. In addition, the intensity of the sample mass peaks can be lowered owing to ion suppression, if the standards and sample is mixed at an improper ratio.

Recalibration uses significantly identified sample mass peaks to calibrating spectra that have all ready been calibrated by either external or internal standards. The recalibration relies on correctly identified mass peak. Great care should be taken to avoid mass peaks that are not significantly identified.

All three calibration methods estimate a polynomial mass correction function. Given a set of N significantly identified mass peaks $\{(m_{obs,k}, m_{cal,k})\}_{k=1}^{N}$, where m_{obs} are the experimental and m_{cal} the calculated masses, then the coefficients off the least-parabola

$$m_{cal} = A * m_{obs}^2 + B * m_{obs} + C$$

can be found by minimizing the function

$$E(A, B, C) = \sum_{k=1}^{N} (Am_{obs,k}^2 + Bm_{obs,k} + C - m_{cal,k})^2$$

(1). The minimization can be done by setting the partial derivatives $\partial E/\partial A$, $\partial E/\partial B$, and $\partial E/\partial C$ equal to zero. Rearranging gives the following linear system

$$\left(\sum_{k=1}^{N} m_{obs,k}^{4}\right) * A + \left(\sum_{k=1}^{N} m_{obs,k}^{3}\right) * B + \left(\sum_{k=1}^{N} m_{obs,k}^{2}\right) * C = \sum_{k=1}^{N} m_{cal,k} * m_{obs,k}^{2}$$

$$\left(\sum_{k=1}^{N} m_{obs,k}^{3}\right) * A + \left(\sum_{k=1}^{N} m_{obs,k}^{2}\right) * B + \left(\sum_{k=1}^{N} m_{obs,k}\right) * C = \sum_{k=1}^{N} m_{cal,k} * m_{obs,k}$$

$$\left(\sum_{k=1}^{N} m_{obs,k}^{3}\right) * A + \left(\sum_{k=1}^{N} m_{obs,k}^{2}\right) * B + N * C = \sum_{k=1}^{N} m_{cal,k}$$

The masses of matrix clusters can, for example, be used as a standard for internal calibration. Possible masses of HCCA matrix clusters were calculated using **Eqs. 2–4**. The number of possible combination for a given *n* is given by $(n+2) * \frac{n+3}{2} - 1$.

Mass (Da)	n	x	y	z
212.03	1	0	0	1
228.01	1	0	1	0
234.01	1	1	0	2
249.99	1	1	1	1
265.96	1	1	2	0
401.07	2	0	0	1
417.05	2	0	1	0
423.06	2	1	0	2
439.03	2	1	1	1
445.04	2	2	0	3
455.00	2	1	2	0
461.01	2	2	1	2
476.99	2	2	2	1
492.96	2	2	3	0
590.12	3	0	0	1
606.09	3	0	1	0
612.10	3	1	0	2
628.07	3	1	1	1
634.08	3	2	0	3
644.05	3	1	2	0
650.06	3	2	1	2

(Continued)

Mass (Da)	n	x	y	z
656.06	3	3	0	4
666.03	3	2	2	1
672.04	3	3	1	3
682.00	3	2	3	0
688.01	3	3	2	2
703.99	3	3	3	1
719.96	3	3	4	0
779.16	4	0	0	1
795.13	4	0	1	0
801.14	4	1	0	2
817.12	4	1	1	1
823.12	4	2	0	3
833.09	4	1	2	0
839.10	4	2	1	2
845.11	4	3	0	4
855.07	4	2	2	1
861.08	4	3	1	3
867.09	4	4	0	5
871.05	4	2	3	0
877.05	4	3	2	2
883.06	4	4	1	4
893.03	4	3	3	1
899.04	4	4	2	3
909.00	4	3	4	0
915.01	4	4	3	2
930.98	4	4	4	1
946.96	4	4	5	0

C. Table of Amino Acid Residue Masses and Some Elemental Masses

Compound	Monoisotopic mass (Da)
E	0.000549
H^+	1.00728
H	1.00783
O	15.9949146
H_2O	18.01056
Glycine (G)	57.02146
Alanine (A)	71.03711
Serine (S)	87.03203
Proline (P)	97.05276
Valine (V)	99.06841

(Continued)

Compound	Monoisotopic mass (Da)
Homoserine lactone (HSL)	100.03985
Threonine (T)	101.04768
Cysteine (C)	103.00919
Isoleucine (I)	113.08406
Leucine (L)	113.08406
Asparagine (N)	114.04293
Aspartic acid (D)	115.02694
Glutamine (Q)	128.05858
Lysine (K)	128.09496
Glutamic acid (E)	129.04259
Methionine (M)	131.04049
Histidine (H)	137.05891
Methionine sulfoxide (MSO)	147.0354
Phenylalanine (F)	147.06841
Cysteic acid	150.993935
Arginine (R)	156.10111
Carboxyamidomethyl cysteine (Cys_CAM)	160.03065
Carboxymethyl cysteine (Cys_CM)	161.01466
Methionine sulfone	163.03032
Tyrosine (Y)	163.06333
Propionamide cysteine (Cys_PAM)	174.04631
Tryptophan (W)	186.07931
N-formylkynurenine	186.07931
Pyridyl-ethyl cysteine (Cys_PE)	208.067039

Adapted from **refs.** ***2–3*** and ExPASy http://www.expasy.org/tools/findmod/findmod_masses.html.

D. Tables of Immonium Ion and Neutral Loss Masses

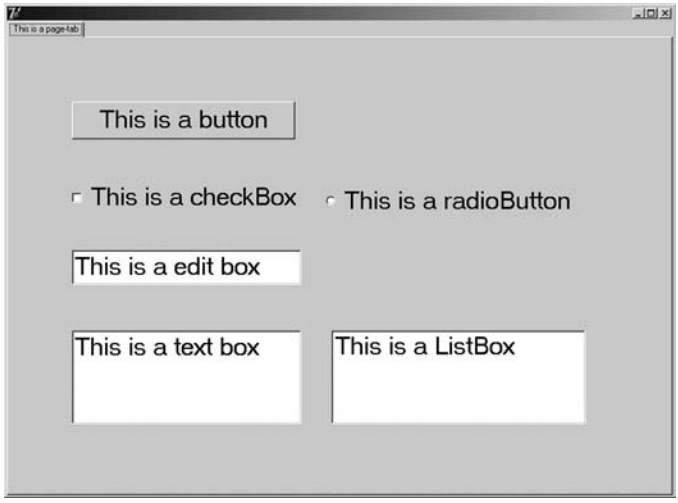
Immmonium ions	Monoisotopic mass (Da)	
Glycine (G)	30.032362	
Alanine (A)	44.052362	
Serine (S)	60.042362	–
Proline (P)	70.062362	+
Valine (V)	72.082362	+
Threonine (T)	74.062362	–
Cysteine (C)	76.022362	
Isoleucine (I)	86.092362	+
Leucine (L)	86.092362	+
Asparagine (N)	87.052362	
Aspartic acid (D)	88.042362	
Glutamine (Q)	101.072362	

(Continued)

Immmonium ions	Monoisotopic mass (Da)	
Lysine (K)	101.102362	–
Glutamic acid (E)	102.052362	
Methionine (M)	104.052362	
Histidine (H)	110.072362	+
Phenylalanine (F)	120.052362	+
Arginine (R)	129.112362	–
Tyrosine (Y)	136.072362	+
Cysteic acid	152.001211	
Tryptophan (W)	159.092362	+
Methionine sulfone	164.037596	
N-formylkynurenine	219.076416	

The masses were calculated with MassCal part of the VEMS v2.0 application. (+) and (–) indicates how well the amino acids are in forming the corresponding immonium ion.

E. Nomenclature Used for Referring to Interface



References

1. Mathews, J. H. (1992) *Numerical Methods for Mathematics, Science, and Engineering*. Prentice-Hall, Englewood Cliffs, NJ, pp. 276–278.
2. Coursey, J. S., Schwab, D. J., and Dragoset R. A. (2001) *Atomic Weights and Isotopic Compositions* (version 2.3.1), Available:http://physics.nist.gov/Comp. National Institute of Standards and Technology, Gaithersburg, MD.
3. Lide, D. R. (1993) CRC Press, London, UK. Handbook of Chemistry and Physics.

Index

N

O

P

Y

METHODS IN MOLECULAR BIOLOGY™ • 367
SERIES EDITOR: *John M. Walker*

Mass Spectrometry Data Analysis in Proteomics

Edited by

Rune Matthiesen
Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark

Mass Spectrometry Data Analysis in Proteomics is an in-depth guide to the theory and practice of analyzing raw mass spectrometry (MS) data in proteomics. As MS is a high throughput technique, proteomic researchers must attend carefully to the associated field of data analysis, and this volume outlines available bioinformatics programs, algorithms, and databases available for MS data analysis. General guidelines for data analysis using search engines such as Mascot, Xtandem, and VEMS are provided, with specific attention to identifying poor quality data and optimizing search parameters. Several different types of MS data are discussed, followed by a description of optimal methods for conversion of raw data into peak lists for input to search engines. Choosing the most accurate and complete databases is emphasized, and a report of available sequence databases is included. Methods for assembling expressed sequence tags (ESTs) into assembled nonredundant databases are provided, along with protocols for further processing the sequences into a format suitable for MS data. *Mass Spectrometry Data Analysis in Proteomics* describes publicly available applications whenever possible.

FEATURES

- **Essential reference for mass spectrometry (MS) in proteomics and glycomics**
- **Concise protocols for optimizing search parameters in MS search engines**
- **Practical guide to MS data analysis and bioinformatics**
- **Thorough review of available databases for proteomics researchers**

CONTENTS

Methods in Molecular Biology™ • 367
PYROSEQUENCING® PROTOCOLS
ISBN: 1-58829-563-X E-ISBN: 1-59745-275-0
ISBN13: 978-1-58829-563-7
ISSN: 1064–3745 humanapress.com

ISBN 1-58829-563-X
9 781588 295637 90000