Springer Protocols

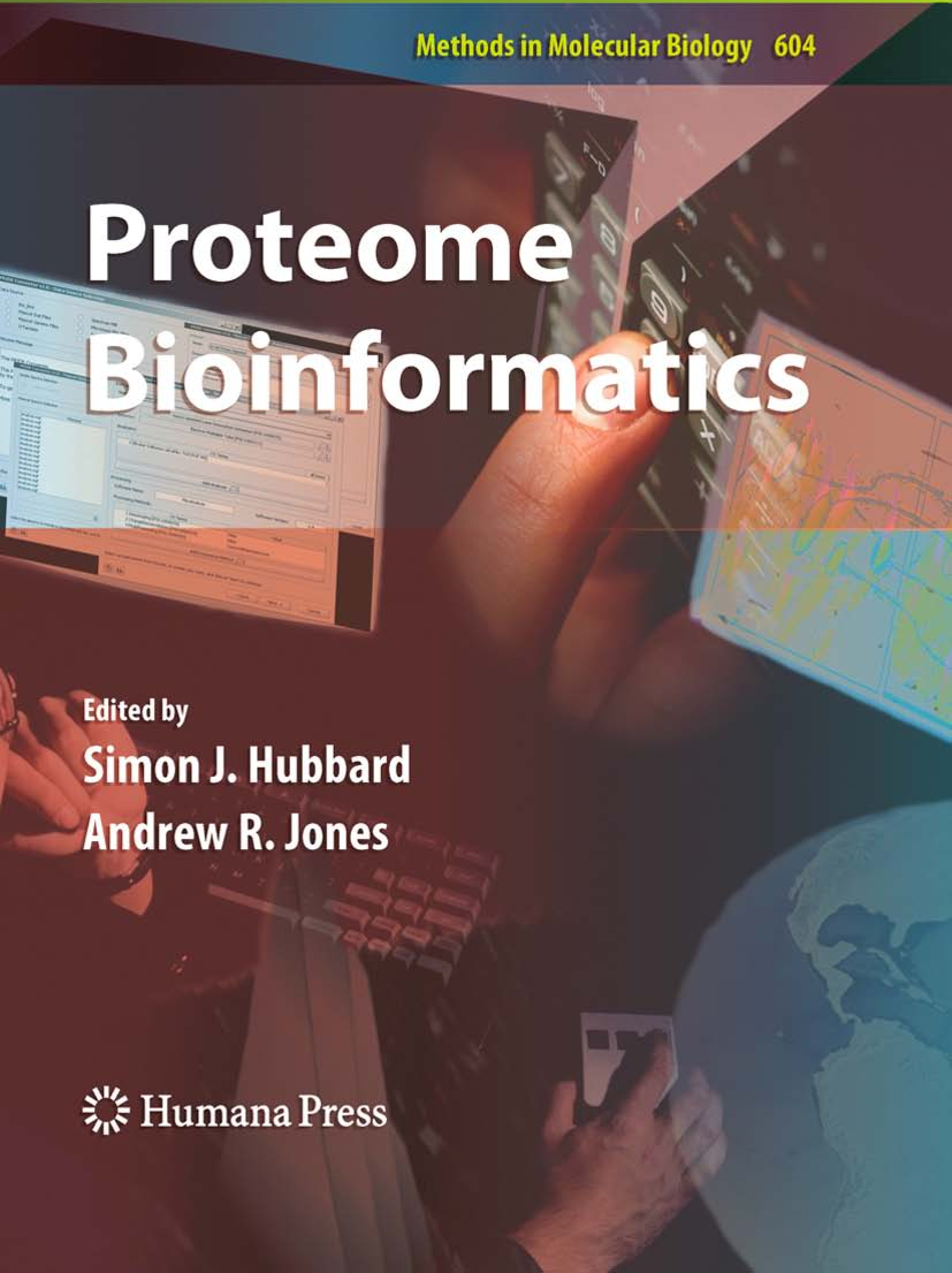
Methods in Molecular Biology 604

Proteome Bioinformatics

Edited by

Simon J. Hubbard

Andrew R. Jones



Humana Press

METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For other titles published in this series, go to
www.springer.com/series/7651

Proteome Bioinformatics

Edited by

Simon J. Hubbard

Faculty of Life Sciences, The University of Manchester, Manchester, UK

Andrew R. Jones

Department of Pre-clinical Veterinary Science, University of Liverpool, Liverpool, UK

Humana Press

Editor

Simon J. Hubbard
Faculty of Life Sciences
The University of Manchester
Manchester M13 9PT
UK
simon.hubbard@manchester.ac.uk

Andrew R. Jones
Department of Pre-clinical Veterinary Science
University of Liverpool
Liverpool L69 7ZJ
UK
andrew.jones@liv.ac.uk

ISSN 1064-3745 e-ISSN 1940-6029
ISBN 978-1-60761-443-2 e-ISBN 978-1-60761-444-9
DOI 10.1007/978-1-60761-444-9
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009941056

Printed on acid-free paper

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

Preface

The field of proteomics moves rapidly. New methods, techniques, applications, standards, models and software appear almost on a daily basis. Accompanying this are plenty of texts on the experimental side of the field and a few appearing on the informatic and data analysis side. This latterly includes one in the *Methods in Molecular Biology* series tackling the specific analysis of "Mass spectrometry data in proteomics" in MMB vol. 376. This current collection builds on this, but takes a broader view of proteome data analysis covering data analysis essentials, but also the databases and data models, as well as practical considerations for analysing database search results, annotating genomes, and speeding up searches. It also digs deeper into some topics, such as decoy database searching and aspects of signal processing in proteomic mass spectrometry. The aim of the volume is to provide the reader with a mix of reviews and methodology chapters, which build from the essentials of database searching in proteomics, on through specific data processing challenges to databases, data standards and data models.

The computational challenges facing proteomics are many and should not be underestimated. The direction in post-genome science is one of increasing complexity and larger, richer datasets. Proteomics is no exception and most active proteomics labs retain or work closely with bioinformaticians and computer scientists as they play an increasingly important role in the day-to-day running of the lab. This book covers all the essential topics that need to be considered and should help the novice and expert alike address their data analysis and management problems.

Manchester, UK — *Simon J Hubbard*
Liverpool, UK — *Andrew R Jones*

Contents

Contributors

RUEDI AEBERSOLD • *Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland; The Institute for Systems Biology, Seattle, WA, USA*

CONRAD BESSANT • *Bioinformatics Group, Cranfield University, Cranfield, Bedfordshire, UK*

ROBERT J. BEYNON • *Department Veterinary Preclinical Sciences, University of Liverpool, Crown Street, Liverpool, UK*

LUCA BIANCO • *Bioinformatics Group, Cranfield University, Cranfield, Bedfordshire, UK*

ISTVAN BOGDAN • *Department of Automatic Control & Systems Engineering, University of Sheffield, UK*

SUSAN M. BRIDGES • *Department of Computer Science and Engineering, the Institute for Digital Biology, Mississippi State University, Mississippi State, MS, USA*

MARKUS BROSCH • *The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*

SHANE C. BURGESS • *College of Veterinary Medicine, Mississippi State University, Mississippi State, MS, USA; Institute for Digital Biology, Mississippi State University, Mississippi State, MS, USA; Life Sciences and Biotechnology Institute, Mississippi State University, Mississippi State, MS, USA*

GERARD CAGNEY • *Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland*

LUONAN CHEN • *Department of Computer Science, University of Missouri, Columbia, MO, USA*

JYOTI CHOUDHARY • *The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom*

DANIEL COCA • *Department of Automatic Control & Systems Engineering, University of Sheffield, UK*

ERIC W. DEUTSCH • *Institute for Systems Biology, Seattle, WA, USA*

ERIC D. DODDS • *Department of Chemistry and Biochemistry, The University of Arizona, Tucson, AZ, USA*

ANDREW W. DOWSEY • *Institute of Biomedical Engineering, Imperial College London, UK*

MARTIN EISENACHER • *Medizinisches Proteom-Center (MPC), Ruhr-Universitaet Bochum, Bochum, Germany*

JOSHUA E. ELIAS • *Department of Cell Biology, Harvard Medical School, Boston, MA, USA*

ASHLEY C. GUCINSKI • *Department of Chemistry and Biochemistry, The University of Arizona, Tucson, AZ, USA*

HOWARD B. GUTSTEIN • *Department of Anesthesiology and Pain Management & Department of Biochemistry and Molecular Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA*

Steven P. Gygi • *Taplin Biological Mass Spectrometry Facility, Harvard Medical School, Boston, MA, USA*

Andreas Hildebrandt • *Center for Bioinformatics, Saarland University, Saarbrücken, Germany*

Wen-Lian Hsu • *Institute of Information Science, Academia Sinica, Taiwan*

Simon J. Hubbard • *Faculty of Life Sciences, University of Manchester, Michael Smith Building, Manchester, UK*

Christopher Hughes • *Department of Biochemistry, University of Western Ontario, London, ON, Canada*

Rene Hussong • *Center for Bioinformatics, Saarland University, Saarbrücken, Germany*

Andrew R. Jones • *Department of Pre-clinical Veterinary Science, Faculty of Veterinary Science, University of Liverpool, Liverpool, UK*

David Jones • *Department of Computer Science, University College London, Gower Street, London, UK*

Philip Jones • *EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*

Samuel Kerrien • *EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK*

Michael Kohl • *Medizinisches Proteom-Center (MPC), Ruhr-Universitaet Bochum, Bochum, Germany*

Oliver Kohlbacher • *Wilhelm Schickard Institute for Computer Science, Tübingen University, Tübingen, Germany*

Gilles A. Lajoie • *Department of Biochemistry, University of Western Ontario, London, ON, Canada*

Henry Lam • *Department of Chemical and Biomolecular Engineering, University of Science and Technology, Hong Kong*

Mark L. Lawrence • *College of Veterinary Medicine, Mississippi State University, Mississippi State, MS, USA; Institute for Digital Biology, Mississippi State University, Mississippi State, MS, USA; Life Sciences and Biotechnology Institute, Mississippi State University, Mississippi State, MS, USA*

Wenzhou Li • *Department of Chemistry and Biochemistry, The University of Arizona, Tucson, AZ, USA*

Allyson L. Lister • *CISBAN and School of Computing Science, Newcastle University, Newcastle upon Tyne, UK*

Bin Ma • *David Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada*

Lennart Martens • *EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*

Jennifer A. Mead • *Bioinformatics Group, Cranfield University, Cranfield, Bedfordshire, UK*

Helmut E. Meyer • *Medizinisches Proteom-Center (MPC), Ruhr-Universitaet Bochum, Bochum, Germany*

Jeffrey S. Morris • *Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA*

BINDU NANDURI • *College of Veterinary Medicine, Mississippi State University, Mississippi State, MS, USA; Institute for Digital Biology, Mississippi State University, Mississippi State, MS, USA*

SANDRA ORCHARD • *EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK*

PATRICK G. A PEDRIOLI • *Institute of Biochemistry, Swiss Federal Institute of Technology Zürich (ETHZ), Zürich, Switzerland*

MELISSA PENTONY • *Department of Computer Science, University College London, Gower Street, London, UK*

KNUT REINERT • *Department Computer Science and Mathematics, Free University of Berlin, Berlin, Germany*

ZHAO SONG • *Department of Computer Science, University of Missouri, Columbia, MO , USA*

CHRISTIAN STEPHAN • *Medizinisches Proteom-Center (MPC), Ruhr-Universitaet Bochum, Bochum, Germany*

TING-YI SUNG • *Institute of Information Science, Academia Sinica, Taiwan*

NAN WANG • *Department of Computer Science and Engineering, the Institute for Digital Biology, Mississippi State University, Mississippi State, MS, USA*

JONATHAN WARD • *Department of Computer Science, University College London, Gower Street, London, UK*

JASON W.H. WONG • *UNSW Cancer Research Centre, University of New South Wales, Sydney, NSW, Australia*

JAMES C. WRIGHT • *Department Veterinary Preclinical Sciences, University of Liverpool, Crown Street, Liverpool, UK*
Faculty of Life Sciences, University of Manchester, Manchester, UK

VICKI H. WYSOCKI • *Department of Chemistry and Biochemistry, The University of Arizona, Tucson, AZ, USA*

DONG XU • *Department of Computer Science, University of Missouri, Columbia, MO, Institute of Biomedical Engineering, Imperial College London, UK*

Chapter 1

An Introduction to Proteome Bioinformatics

Andrew R. Jones and Simon J. Hubbard

Abstract

This book is part of the Methods in Molecular Biology series, and provides a general overview of computational approaches used in proteome research. In this chapter, we give an overview of the scope of the book in terms of current proteomics experimental techniques and the reasons why computational approaches are needed. We then give a summary of each chapter, which together provide a picture of the state of the art in proteome bioinformatics research.

Key words: Proteomics, Bioinformatics, Mass spectrometry, Quantitation, Databases, Data standards

1. Introduction

The term "proteome" is broadly defined as the complete set of proteins that could be present in a sample or organism or the set of proteins that can be detected by some methodology. The related term "proteomics" was originally coined by Marc Wilkins and colleagues and is now widely recognised as one of the 'omics – the set of post-genomic technologies which seek to further understand the genome in terms of the molecules it encodes, their functions, interactions and related properties. Proteomics as a field of study is a broad one and can have slightly different connotations to different aspects of biomedical science. Here, we refer to the aspect of the field dominated by mass spectrometric-based characterisation of proteins rather than structural or other biophysical approaches. Many of the experimental techniques used in proteomics, such as gel electrophoresis, liquid chromatography and mass spectrometry (MS), have been around for several decades, yet it is only in the last 10–15 years that approaches have been developed allowing many proteins to be detected or

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_1, © Humana Press, a part of Springer Science+Business Media, LLC 2010

quantified simultaneously, coming closer to “global” methods of proteome analysis. The paradigm shift has been caused in part by technological developments in the design of mass spectrometers but more significantly by the availability of genome sequence data for many organisms that allow mass spectrometry data to be used to identify peptides and proteins on a large scale. Current methods are now capable of identifying thousands of proteins at relatively low cost. However, significant challenges remain in data handling, storage, dissemination and analysis. This book provides an introduction to the main computational challenges presented by proteome research, and the bioinformatics solutions being developed to allow maximum value to be derived from proteome data sets.

2. Protein Identification by Mass Spectrometry

Mass spectrometry plays a fundamental role in the majority of protein identification pipelines used in proteomics. Individual proteins can typically be “identified” either by a single stage of mass spectrometry, called peptide mass fingerprinting (PMF), or by two stages, called tandem MS or MS/MS. If a single stage is used, a protein is digested (for example with trypsin) and the pattern of peptide masses (PMF) is compared with a theoretical digest of database sequences to make an identification. Zhao Song, Luonan Chen and Dong Xu present an introduction to the bioinformatics challenges in PMF (Chapter 2). In tandem MS, individual peptides selected in the first stage of MS are subjected to a fragmentation process, and the products of fragmentation are analysed in the MS2 stage. Various databases search engines are available for comparing the mass spectrum produced with the complete set of spectra that would be expected from a theoretical digest of a protein sequence database. Simon Hubbard provides a general introduction to the computational approaches used (Chapter 3) and Markus Brosch and Jyoti Choudhary describe how statistical techniques are used to determine the reliability of peptide and protein identifications (Chapter 4). A statistical technique that is gaining prominence in proteomics involves the use of searching against a decoy database that comprises sequences known to be incorrect, to allow the false discovery rate to be estimated – described by Elias and Gygi in Chapter 5.

The information gained by collecting large numbers of high quality peptide spectra can potentially be exploited, and Wysocki and colleagues described studies demonstrating this principle, where specific amino acid compositions about a candidate bond can enhance or reduce the likelihood of fragmentation (Chapter 6).

An alternative approach for making peptide identifications involves the generation of libraries containing mass spectra and peptide identities that have previously been assigned to them empirically. Newly generated mass spectra are then compared with the library in order to make a putative peptide identification. These spectral library approaches are discussed by Henry Lam and Ruedi Aebersold in Chapter 7.

The approaches described above all rely upon using some type of database search to make peptide identifications. Methods have also been developed that can determine peptide sequences directly from the fragmentation pattern on MS2 spectra, as described in Chapter 8 by Hughes, Ma and Lajoie. These *de novo* peptide sequencing methods are particularly applicable for organisms with no sequenced genome, but they are also used for peptides that are difficult to identify by standard database searches. Outside of *de novo* techniques, a more general review of approaches developed to identify proteins across the species barrier where no complete genome is available is described by Wright, Beynon and Hubbard in Chapter 9. Proteome data has also been used for detecting, improving and confirming gene models in newly sequenced genomes, as outlined by Nanduri and colleagues in Chapter 10, in an area often referred to as Proteogenomics. Proteomics has untapped potential in this area since it also provides direct evidence that a gene is expressed at the protein (functional) level above and beyond transcriptional (EST, cDNA or array) evidence.

3. Data Processing

It is not a trivial process to convert the raw data collected directly from a mass spectrometer to the final result, such as peptide or protein identifications. First, the raw trace obtained must be converted to discrete peaks that can be used in a database search; these signal processing approaches are summarised by Rene Hussong and Andreas Hildebrandt (Chapter 11). A novel approach for improving the speed of spectral processing, using re-configurable hardware is presented by Coca, Bogdan and Beynon (Chapter 12).

Multiple reaction monitoring (MRM) is an MS-based experimental approach in which particular peptides are selected for analysis by the experimentalist, in contrast to the majority of proteome approaches where peptides are identified “at random” by the mass spectrometer. MRM approaches rely on prior knowledge of the fragmentation patterns of peptides, which can be derived by data mining, as described by Mead, Bianco and Bessant in Chapter 13.

There are several stages in processing from the raw trace to the final result, requiring a series of different file formats to be manipulated. The OpenMS architecture is designed to allow bioinformaticians to construct identification workflows, as described by Knut Reinert and Oliver Kohlbacher (Chapter 14). The TransProteomicPipeline (TPP) developed by the Institute for System Biology offers a series of packages for processing and analysing proteome data, described by Patrick Pedrioli in Chapter 15.

4. Protein Quantitation

Researchers are interested not only in the proteins present or absent in samples of interest but the relative or absolute abundance of each protein. In one commonly applied experimental approach, proteins are separated by two-dimensional gel electrophoresis before an MS-based identification process. The gels can be subjected to a detection agent allowing proteins to be visualised. Proteins can be quantified by analysis of gel images, and several of the computational issues are described by Andrew Dowsey and colleagues (Chapter 16).

Mass spectrometry in its native form is not generally thought of as a quantitative technique, since the abundance of a peptide ion detected correlates poorly with the quantity of the protein(s) from which it was derived in the source sample due to several confounding factors. Various approaches have been developed to alleviate these problems so that MS can be used for relative or absolute quantification. One of the most common approaches uses stable isotopes as labels, applied at various stages of sample preparation and which are subsequently mixed. Detected peptides from two samples appear in the mass spectra as adjacent, related peaks, separated by a fixed difference dependent on the given stable isotopes used. The relative peak height can be used to determine relative abundance of the protein between samples. Wen-Lian Hsu and Ting-Yi Sung describe generic analysis tools for handling such data (Chapter 17). Other methods are capable of determining relative or absolute protein abundance in the absence of isotopes labels, as presented by Cagney and Wong (Chapter 18).

5. Data Standards and Databases

A single proteome study can generate large volumes of data and metadata: describing sample processing, protein separation, raw and processed mass spectrometry data and the final results as protein identifications, potentially with quantitative values. It is now commonly accepted that there is significant value in the ability to re-assess data outside the producing laboratory. It is also widely acknowledged that differences in how samples are processed and which data analysis routines have been applied can affect the proteins that are detected and calculations of their abundance. Several databases exist for storing MS-based proteome data, including PeptideAtlas at the Institute of Systems Biology (Chapter 19, Eric Deutsch) and PRIDE at the European Bioinformatics Institute (Chapter 20, Philip Jones and Lennart Martens). Both databases have rich querying capabilities and allow published data sets to be analysed by other groups for deriving new biological findings and for improving analysis algorithms.

A major hindrance in proteome data processing has been the abundance of different file formats, produced by different MS instrument vendors and analysis software packages. Historically, it has been difficult to analyse data originating in a particular laboratory unless other laboratories have the same software packages. In recent years, several groups have developed data standards to facilitate data dissemination and exchange. These include: the Molecular Interaction format, which captures protein–protein interaction data (Chapter 21, Orchard and Kerrien), and the mzML format, which captures the output of mass spectrometers (Chapter 22, Deutsch), both developed under the auspices of the Proteomics Standards Initiative (PSI). The Functional Genomics Experiment (FuGE) Model is a general model used to facilitate the development of shared data standards across different omics techniques, which has been used by PSI and other groups for managing proteome data, as described by Andrew Jones and Allyson Lister in Chapter 23. The ProDac project is described in Chapter 24 by Christian Stephan and colleagues, which is a consortium of academic groups aiming to maintain a centralized proteomics data repository and support the development of data standards that enable efficient proteomics data sharing and dissemination.

Finally, protein disorder is discussed by Melissa Pentony, Jonathan Ward and David Jones in Chapter 25. This phenomenon is assuming increasing importance since disordered proteins are widespread in genomes and are correlated with post-translationally modified regions, and will be of interest to many proteomic practitioners.

Chapter 2

Bioinformatics Methods for Protein Identification Using Peptide Mass Fingerprinting

Zhao Song, Luonan Chen, and Dong Xu

Abstract

Protein identification by mass spectrometry (MS) is an important technique in proteomics. By searching an MS spectrum against a given protein database, the most matched proteins are sorted using a scoring function and the top one is often considered the correctly identified protein. Peptide mass fingerprinting (PMF) is one of the major methods for protein identification using MS technology. It is faster and cheaper than the other popular technique – Tandem Mass Spectrometry. Key bioinformatics issues in PMF analysis include designing a scoring function to quantitatively measure the degree of consistency between a PMF spectrum and a protein sequence and assessing the confidence of identified proteins. In this chapter, we will introduce several scoring functions that were developed by others and us. We will also provide a new statistic model to evaluate the confidence of the score and make an improvement for ranking proteins in protein identification. Our developments have been implemented in a software package "ProteinDecision," which is available at http://digbio.missouri.edu/ProteinDecision/.

Key words: Peptide mass fingerprinting, Protein identification, Confidence assessment, Scoring function, ProteinDecision

1. Introduction

1.1. Protein Identification Using Mass Spectrometry Technology

The basic unit of a protein is the amino acid, which has 20 types with specific molecular weights for each (see Note 1). The chain of amino acids forms the protein sequence and determines its structure and function. The main purpose of protein identification is to determine the composition of proteins in a sample of animal cells, bacterial or plant tissues, etc., often through mapping these proteins to known, better characterized ones. Major methods of protein identification apply mass spectrometry (MS) and a subsequent database search, in which heuristic algorithms (1) are designed to assign scores for all the candidate proteins in

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_2, © Humana Press, a part of Springer Science+Business Media, LLC 2010

a database. The general approach for MS protein identification is to match the features derived from the MS spectrum of a protein sample with the database that contains the sequence fragments of a protein digested by a specific enzyme (see Note 2). Several prevalent techniques are widely used currently. Peptide mass fingerprinting (PMF) protein identification compares the masses of peptides derived from the experimental spectral peaks with each of the possible peptides generated by computationally digesting proteins in the sequence database. The MS/MS method further breaks each digested peptide into smaller fragments, whose spectra provide effective signatures of individual amino acids in the peptide for protein identification. While the MS/MS methods provide more features in defining peptides, it is much more expensive and time-consuming than PMF. PMF is an economic alternative for protein identification, and it can serve as an effective filter to select appropriate proteins to conduct MS/MS analysis. In this chapter, we focus on the PMF protein identification, which is still of great value in proteomics.

1.2. Peptide Mass Fingerprinting

Protein identification is performed in two stages as shown in Fig. 1. At the experiment stage, protein samples are collected first. Then, with an extraction instrument, the proteins are separated from the samples and precipitated at different spots in a 2D gel page (2) according to their molecular weights and pI values. By selecting a specified spot in the gel, the corresponding protein(s) will be mixed with specific enzyme and digested into small pieces (see Note 3). Finally, the mass spectrometer, commonly a MALDI-TOF instrument, will generate the PMF spectrum (3, 4) for protein identification. At the computational stage, all the candidate proteins in the search database are theoretically digested using the same enzyme. The simulated spectrum for each candidate is created for comparison (see Note 4). The common PMF protein identification is carried out through two steps: (a) the experimental PMF spectral peaks are compared with simulated ones and (b) the proteins in the sequence database with best matches are considered the top candidates for proteins in the experimental sample.

Unlike 2D gels, PMF provides at least some sequence-level information for protein identification. The PMF of a protein is like a fingerprint, which is unique to the molecule or represents a small population of the proteins in the database. Following an enzyme digestion, a collection of peptides with the masses (or mass-to-charge ratios) identified from the PMF spectra will be mapped to known protein sequences (see Note 5). The use of the fingerprint to identify proteins relies on the ability to search sequences that are already present in databases. Hence, it is important that the organism has a complete, whole genome sequence available so

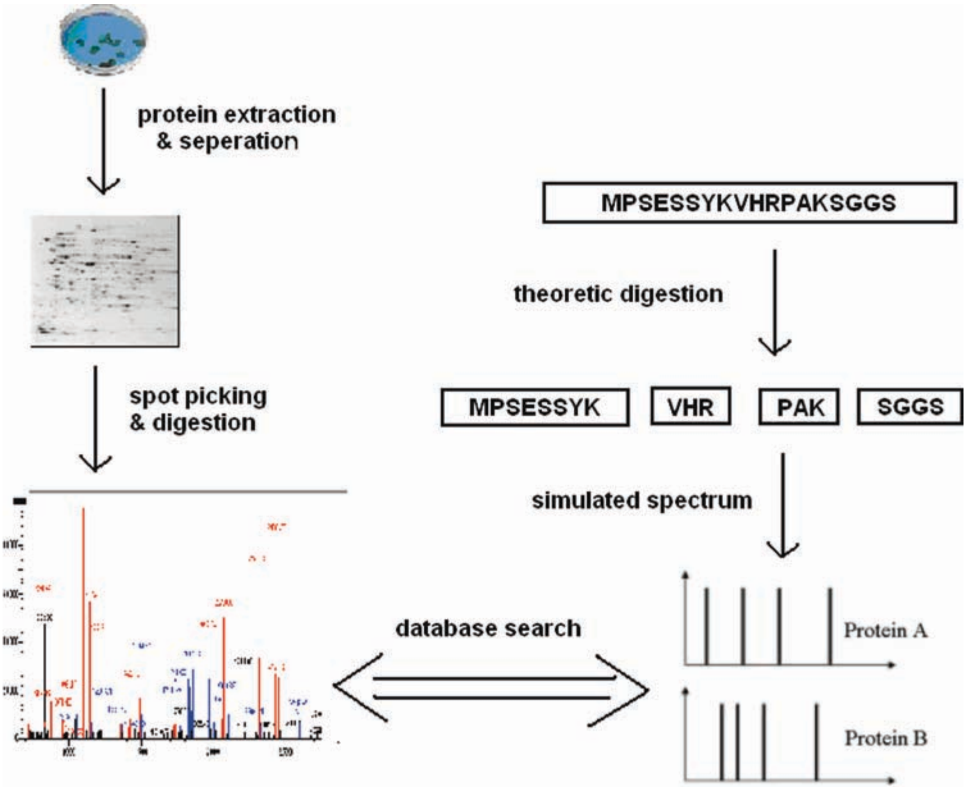


Fig. 1. Peptide mass fingerprinting protein identification is performed at two stages. At the experiment stage, proteins are digested into peptides, where the mass-charge ratios are shown in the spectra; at the computational stage, protein candidates are theoretically digested and simulated spectra are generated for comparison

that all the proteins can in principle be determined or predicted. When the whole genome sequence is unavailable, researchers often search an MS spectrum against a composite protein database containing all known proteins (such as UniProt or Swiss-Prot), trying to identify a protein in another species that is highly similar to the homolog in the native organism.

2. Materials

2.1. 12 Gel Spots Analyzed by MALDI-TOF MS

In order to provide benchmark data for the computational studies, we used seven protein standards, which yielded 12 gel spots.

1. In-gel trypsin (5), digests were performed for the Coomassie-stained 2D gel plugs. The digests were dried on a centrifugal evaporator, reconstituted, and desalted on C18 ZipTips.
2. The desalted digests were analyzed by MALDI-TOF MS with CHCA in the positive ion delayed extraction reflector mode.

3. After the sample spots were washed on target with diammonium citrate to reduce the interference from matrix ion clusters, they were reanalyzed by MALDI-TOF MS in positive ion delayed extraction reflector mode.

The 12 spots are from *Aspergillus niger*, Bovine, *Lens culinaris*, Horse and Soybean, respectively.

2.2. 40 Soybean Root Hair Samples Separated by 2-DE

The soybean MS data were provided by the lab of our collaborator Dr. Gary Stacey.

1. The proteins (600 mg) were extracted from soybean (cvWilliams 82) and root hair, which were separated by 2-DE (24 cm IPG strip, linear pH4–7). Four replicates were performed and gel pictures were analyzed using Phoretix (nonlinear dynamics, v2005).
2. Spots identified in at least three out of the four replicates were excised using a spot picker and their molecular weights and pIs were determined. The gel plugs were then digested using sequencing-grade modified trypsin (Promega, Madison, WI, USA).
3. Tryptic peptides were lyophilized, reconstituted in 10 mL of 700:290:10 by volume ACN/water/formic acid and 0.5 mL of the solution was mixed with the same volume of a-cyano-4-hydroxycinnamic acid (Fluka MS-grade, Sigma–Aldrich, St. Louis, USA) solution (5 mg/mL in 500:380:20:100 ACN/water/10% TFA/100 mM ammonium dihydrogen phosphate). The sample/matrix (0.3 mL) mix was deposited on a stainless-steel plate (ABI01-192-6-AB).
4. The tryptic peptides were analyzed on an Applied Biosystems Inc. 4700 MALDI TOF/TOF MS in positive ion delayed extraction reflector mode with a 355 nm (200 Hz) laser. The instrument was calibrated with ABI peptide standards (4700 Mass standards kit, 4333604).
5. Spectra were analyzed using the GPS Explorer software (v. 3.0) (Applied Biosystems) and the Matrix Science's MASCOT search engine (www.matrixscience.com) against the NCBI Viridiplantae protein database. Search parameters include a maximum of 150 ions per MS spectrum with an S/N 0.20, a mass error of 0.1 Da for the mono-isotopic precursor ions, a maximum of one allowed miscleavage by trypsin, and an exclusion of peptide masses corresponding to the autolysis of the trypsin, carbamidomethylation of cysteines and methionine oxidation, respectively, as fixed and variable modifications.

Forty proteins were identified confidently with the MS/MS mode, and we used their corresponding MS fingerprinting data as the inputs for our tests of scoring schemes. We assume that a test protein identification is correct if our search using the fingerprinting data matches the protein identified from the MS/MS data.

2.3. Search Database Preparation

The database used for protein identification is *sprot45* from UniProtKB/Swiss-Prot (last updated in January 2005), together with the 40 proteins from soybean (generated after January 2005) that we have identified but are not already included in the database. The database has 163,275 proteins in total, and it is formatted in a specific form, including eight fields for each entry (see Note 6): accession number, peptide number, peptide sequences, peptide masses, peptide lengths, protein sequence, protein name, and protein molecular weight. The molecular weight (see Note 7) of a peptide of N residues is calculated as

$$\sum_{i=1}^{n} \text{residue\_mass}_{\text{i}} + \text{mass}_{\text{water}} \qquad (1)$$

Equation 1 takes into account an amino-terminal hydrogen and a carboxy-terminal hydroxyl group, which sums up to 18.015.

In this study, we only consider complete trypsin digestion of a protein and peptide without including any missed cleavage. In addition, we assume that the charge state of all the peptides is 1 and no posttranslational modification exists in any peptide. We use only mono-isotopic peaks.

3. Scoring Function Methods

The scoring functions are quantitative measurements of protein identification. They evaluate the degree of matching between a collection of proteolytically derived peptides and an MS spectrum and map it to a comparable value using a mathematical formula. A good scoring function will consider different factors and balance them well which leads to the result that the correct protein(s) are always the top ranked candidates in the list of matches. The challenges in the process of developing a scoring function include: (a) make use of all relevant information that will affect the prediction, (b) design a good model to balance the effects and produce reasonable result.

In this section, we first describe briefly the widely used MOWSE scoring function and a successful commercial software "Mascot". Then we will illustrate the three novel scoring functions (Subheadings 3.3–3.5 in the following) that we developed.

3.1. MOWSE

MOWSE (6) is one of the earliest scoring schemes in protein identification using PMF data, which is still widely applied.

The scheme is based on the number of possible matches within a target protein and the occurrence of the molecular weight of each peptide. A frequency table is constructed for all peptide entries in the database. Each column in the frequency table represents the molecular weight of the protein and is divided into 10 kDa intervals. Rows represent the molecular weight of peptides and are divided into 100 Da intervals. Proteins found in the database are entered into the table based on their molecular weights and the weights of peptides found in each protein. Each cell thus comprises the occurrence of peptides within a specific molecular weight range in a protein of a given intact molecular weight. The frequency table is constructed by normalizing the value in each cell with the largest number found in each column. Specifically, the frequency f_{ij} in cell_{ij} is $f_{ij} = N_{ij} / N_{j\max}$, where $N_{j\max} = \max\{N_{1j}, N_{2j}, \ldots\}$ is the largest number in column-j. For protein identification, each protein in the target database is scored by multiplying together the frequency value f_{ij} of the matched peptides, the molecular weight of which differs from the experimental spectral peak within a cutoff value (typically 1 or 2 Da). This product is scaled with the protein molecular weight and then inverted. The final $\text{score} = 50{,}000 / (p_n \times w_p)$, where p_n is the product of matched distribution scores and w_p is the molecular weight of the candidate protein "hit" in the database. $p_n \propto \prod_{i=R(l),l\in H} f_{ij}$, where $R(l)$ represents the row number of the table for the l-th fragment of the mass spectra, and H is the set of the matched fragments of the mass spectra with the protein.

3.2. Mascot

Mascot (7) is an extension of the MOWSE algorithm that uses a probability score to rank the candidates. It is one of the most successful software for protein identification. However, as a commercial software tool, the algorithm details are not open to the public.

3.3. Normal Distribution Scoring Function

There are several other bioinformatics tools developed for protein prediction, such as ProteinProspector (8), ProFound (9), OLAVPMF (10), and Probity (11). To make use of the peak intensity and the quantitative difference between the experimental mass values of selected peaks and matched mass values in the protein database, we developed a number of other scoring functions (12, 13). One of them is a normal distribution scoring function (NDSF) (12) based on Eq. 2:

$$\text{Score} = \sum_i \left(\text{Int}_i \times \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(\text{mt}_i - \text{me}_i)^2}{2\sigma_i^2} \right) \right) \qquad (2)$$

In Eq. 2, mt_i is the mass of the theoretical peptide in the database, me_i is the mass of the matched experimental peak, Int_i is the corresponding intensity value for the experimental peptide, and $\sigma_i = \frac{1}{3}\mathrm{tolerance} \times \mathrm{me}_i$ (typically $\mathrm{tolerance} = 100$, $\mathrm{ppm} = 0.01\%$). NDSF assumes that all mass matches between theoretical peptides and experimental peptides follow a normal distribution with the mass of the experimental value as the mean.

3.4. Probability-Based Scoring Function

To make better use of the statistical properties and to handle these properties in PMF protein identification more systematically, we developed another new scoring scheme, probability-based scoring function (PBSF), based on the MOWSE occurrence table. In this case, when comparing a mass distribution of peptides (n fragment molecular weights in the spectrum) with the molecular weights in database entries (protein k in the column j), $R(l)$ represents the row number of the table for the l-th fragment of the mass spectra. When the difference between two peptide weights is within a tolerance value, it is a "hit" or match. Otherwise, it is non-matching. The probability for a match between a mass distribution of peptides and a protein k in the database is computed via

$$\Pr(P_k) = \prod_{i=R(l),l\in H_k} \left[1 - \left(1 - \frac{m_{ij}}{M_j}\right)^{n_{ij}^k}\right] \tag{3}$$

where Pr(Pk) represents a likelihood for protein k matching with the fragment peptides of the experimental mass spectrum. H_k is the set of the matched fragments of the mass spectrum with protein k, and n_{ij}^k is the number of peptides in cell_{ij} for protein k. Let m_{ij} represent the average number of occurrences of peptides in cell_{ij} for one protein in the database, and M_j is the total number of occurrence of peptides in the j-th column of the database, i.e., $M_j = \sum_{i=1}^{n_r} m_{ij}$, where n_r is the total number of rows in the table. Clearly, m_{ij} / M_j is the frequency in the cell_{ij} for the column j. Note that such a frequency is different from f_{ij} of MOWSE.

In mass spectra, high-intensity peaks are more likely to be peaks representing true peptides, whereas low-intensity peaks are more likely to be noise. To account for the peak intensity effect, we modify the Eq. 3 as

$$\Pr(P_k) = \prod_{i=R(l),l\in H_k} \left[\left(1 - \left(1 - \frac{m_{ij}}{M_{ij}}\right)^{n_{ij}^k}\right)(1 - I_l)\right] \tag{4}$$

where I_l is the normalized intensity ([0,1]) of the l-th spectrum, i.e.,

$$I_l = \frac{1}{1 + e^{-\sigma(\hat{I}_l - \bar{I})}} \tag{5}$$

In Eq. 5, $\hat{I}_l$ is the original intensity in the spectrum, $\bar{I}$ is the average intensity for all selected peaks, and α is a constant (see Note 8). To achieve good precision in computing, we adopt *−log Pr*(*Pk*) as the score of PBSF (13) for protein identification.

3.5. Modified Probability-Based Scoring Function

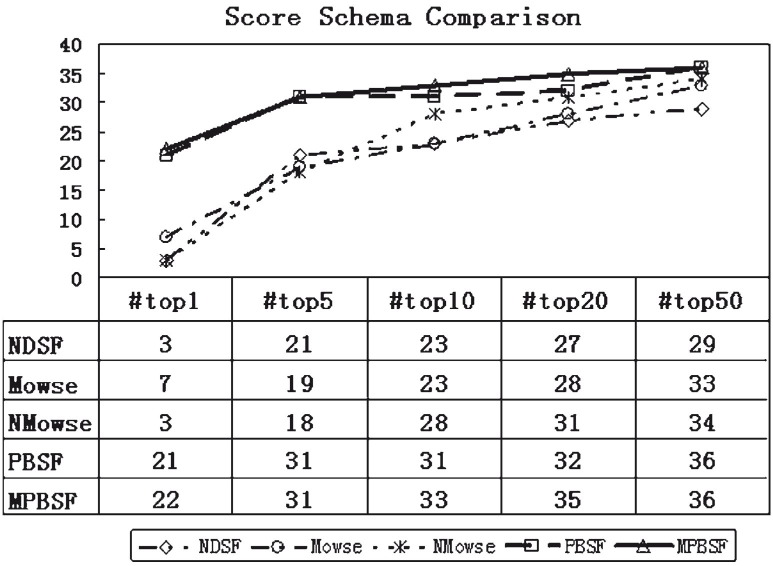
We further developed another scoring scheme by integrating the information for the neighboring matching peptides into a modified probability-based scoring function (MPBSF). The score, which utilizes the average distance of matched peptides, is defined as:

$$\text{ADMP} = \frac{\sum_{i=1}^{n_m - 1} \text{Dis}_{i,i+1}}{n_s / n_p} \tag{6}$$

where the numerator represents the sum of the distances between two adjacent matching peptides, while the denominator represents the total number of possible digested segments in the protein divided by the number of matching peptides. Specifically, n_m, n_s, and n_p are the number of the matching peptides in the spectra, the number of the digested segments, and the number of the matching peptides in the protein, respectively. In this score, when a number of matching peptides are clustered together on the protein sequence, the match is more significant. The final MPBSF score is calculated as −log *Pr*(*Pk*)−log(ADMP).

3.6. Scoring Functions Comparison

We compared the performance of PBSF and MPBSF with the other scoring functions using the same experimental dataset as described in Subheadings 2.1 and 2.2 (12 standards together with 40 soybean proteins). For each protein identification, we manually selected a set of peaks from a spectrum provided by the Proteomics Center, University of Missouri-Columbia. We determined that matched peptides should cover at least 25% of a protein sequence in order to be listed as a candidate of correct result. Figure 2 shows the comparative results of the five scoring functions (NDSF, MOWSE, NMOWSE, PBSF, and MPBSF) in terms of ranking correct proteins among the top hits. It indicates that PBSF and MPBSF performed significantly better than the other three methods (especially in the "#top 1" category). The performance results of PBSF and MPBSF are similar, while MPBSF slightly outperforms PBSF.



	#top1	#top5	#top10	#top20	#top50
NDSF	3	21	23	27	29
Mowse	7	19	23	28	33
NMowse	3	18	28	31	34
PBSF	21	31	31	32	36
MPBSF	22	31	33	35	36

Fig. 2. Scoring function comparison, where numbers in the cells show the occurrences in the categories

4. Confidence Assessment for PBSF

4.1. Why Confidence-Based Assessment?

In Subheading 3.4, our scoring model PBSF gives a ranking list of candidate proteins. However, the raw score (Eq. 3)-based ranking still has two problems:

1. The raw score is dependent on both query PMF and the database, and different PMFs searches will give different raw scores for top proteins. Hence, the raw score itself cannot tell the significance of the match.
2. Since PBSF may have bias, one can improve the confidence assessment based on the PBSF score distribution.

For these two reasons, we designed and implemented a statistical model for confidence assessment by testing the null hypothesis that the match between the MS spectrum and a specific protein given the PBSF score is by chance (14).

4.2. Confidence Assessment Model

The confidence assessment process is similar to the scoring function, except that the model is more rigorous statistically for hypothesis testing. The following steps are needed to conduct computational protein identification using PMF data:

Step-1. Take one fragment from a PMF spectrum, and compare it with all theoretically digested peptides in the search database. A score is assigned for each matching pair.

Step-2. Repeat Step-1 for all remaining fragments from a PMF spectrum one by one until all fragments of the spectra are compared with the protein. For each protein in the searched database, sum the scores of all the fragments from a PMF spectrum as the score for the protein.

Step-3. Rank all the proteins in the database according to their scores. The protein with the highest score in the database is taken as the best matched protein in the biological sample.

We follow the above three steps to analyze the statistical significance of the scores. In Step-1, assume we have the *l*-th fragment, which falls in the row-*i* of the frequency table (i.e., *i* = *R*(*l*)). When we compare this fragment with protein-*k* with the molecular weight in the column-*j*, the score of fragment-*l* is

$$s_l = \begin{cases} q_{ij} & \text{if } l \text{matches at least one peptide} \\ \bar{q}_{ij} & \text{otherwise} \end{cases} \tag{7}$$

As we have discussed in Subheading 3.4, Eq. 3 represents a statistical model scoring function while Eq. 4 additionally considers peak intensity, which makes an adjustment from Eq. 3. We adopt $-\log P(k|w)$ with the consideration of peak intensity as the score function in this study, and hence, $q_{ij} = -\log \{[1-(1-(m_{ij}/M_{ij})^{m_{ij}}]\ (1-\mathrm{I}_l)\}$ and $\bar{q}_{ij} = -\log\ [(1-\ (m_{ij}\ /\ M_{ij}))^{m_{ij}}\ \mathrm{I}_l]$ (penalty for mishit) or $\bar{q}_{ij} = 0$ (no penalty for mishit), where the variables have the same meaning as in Subheading 3.4. In either case of $\bar{q}_{ij}$, the assumption of $q_{ij} \geq \bar{q}_{ij}$ holds. Here, protein-*k* is assumed to be chosen randomly in the column-*j*, and has the average statistical property of all proteins in the column-*j*. In this case, n_{ij}^{k} in Eq. 3 is replaced by m_{ij}. Now, we examine the probability distribution of the score based on the occurrence in cell_{ij}.

$$P(s_l = s) = \begin{cases} 1-\left(1-\dfrac{m_{ij}}{M_j}\right)^{m_{ij}} & s = q_{ij} \\ \left(1-\dfrac{m_{ij}}{M_j}\right)^{m_{ij}} & s = \bar{q}_{ij} \end{cases} \tag{8}$$

and with the cumulative probability as

$$P(s_l >= s) = \begin{cases} 1\ \bar{q}_{ij} \geq s > -\infty;\ 1-\left(1-\dfrac{m_{ij}}{M_j}\right)^{m_{ij}} \\ q_{ij};\geq s > \bar{q}_{ij};0\ s > q_{ij} \end{cases} \tag{9}$$

where $1-(1-(m_{ij}\ /\ M_j)^{m_{ij}}$ is the probability of no match between fragment-*l* and any of the m_{ij} peptides in the cell_{ij} for protein-*k*. The average and standard deviation of score *s* are

$$\mu_l^k = \int_{-\infty}^{+\infty} sP(s_l = s)\mathrm{d}s \tag{10}$$

$$(\sigma_l^k)^2 = \int_{-\infty}^{+\infty} (s - \mu_l^k)^2 P(s_l = s)\mathrm{d}s \tag{11}$$

In Step-2, all fragments of the spectra are compared with the peptides for the protein-*k*. Let $S^K = \sum_{l=1}^{n} s_l$. Then, the probability distribution $P(S^K \geq s)$ is the convolution of *n* distributions of individual fragments. That is, $P(S^K = s) = P(s_1 = s) \otimes \cdots \otimes P(s_n = s)$, where $\otimes$ is the convolution operator. With the assumption of independent distribution of s_l, according to the central limit theorem and law of large numbers, the probability converges to the Gaussian distribution when the number *n* is sufficiently large. In this case, the approximation of the Gaussian distribution can be analytically expressed as

$$P(S^k \geq s) = \int_{s}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^k}} \mathrm{e}^{-\frac{(s-\mu^k)^2}{2\sigma^k}} \mathrm{d}s \tag{12}$$

where $\mu^k = \sum_{l=1}^{n} \mu_l^k$ and $(\sigma^k)^2 = \sum_{l=1}^{n} (\sigma_l^k)^2$

We apply Eq. 12 to derive the numerical solution for $P(S^K \geq s)$ (see Note 9). $P(S^K \geq s)$ can be interpreted as the *p*-value for the significance of the raw score *s*, i.e., the probability of achieving *s* or higher score by chance. The smaller the *p*-value, the less likely that the hypothesized distribution is correct and the more significant the score is in terms of protein identification.

4.3. Confidence Assessment Results

We used the protein identification of mitogen-activated protein kinase 2 [in *Glycine max*] (GI:33340593) as an example to show how our method works (see Table 1). Based on the raw scores, Q9CHU6 ranked number 1 and 33340593 (the correct protein) ranked number 2. By using the *p*-value, their orders were switched. From the table, the protein 33340593 has eight peptides matching 9 times in total with spectral peaks, while Q9CHU6 has seven peptides matching 11 times. Notice that a single peak may match multiple peptides. For example, peptide "TLEEFVGSLEKPR" of protein Q9CHU6 has three repeated matches for a single peak with a low intensity 0.05. This can explain why Q9CHU6 obtained a higher score, as the intensity for each of the three repeated matches is overestimated. This overestimation is corrected by our statistical model, which considers the overall distribution of possible scores.

By using Eq. 12, we obtained the ranks for the 52-protein dataset. We derived the ranks based on raw scores (SR) and *p*-values

Table 1
Protein identification for mitogen-activated protein kinase 2

	33340593		Q9CHU6	
Protein ID.	**Peptide**	**Int.**	**Peptide**	**Int.**
Peaks	QSFQEK	0.516	ALYFSK(2)	0.015
	QLPLYR(2)	0.064	YQEAVR	0.516
	KPLFPGR	0.166	TEEVYK(2)	0.052
	ICDFGLAR	0.026	YISTYK	0.011
	DHVHQLR	0.148	VLSGPAVNFSGDK	0.057
	DIVPPPQR	0.139	SENLPANLIQAQR	0.99
	YKPPIMPIGK	0.785	TLEEFVGSLEKPR(3)	0.05
	YIHSANVLHR	0.266		
MW(Da)	44,765		52,424	
Score	36.957681		37.233432	
p-value	5.07E–10		5.06E–09	
Rank change	2–1		1–2	

"Peptide" lists the matching peptides with the numbers in the bracket showing the occurrence if the peptide appears more than once in the protein. "Int." is the corresponding normalized intensity for the matching peptides

(PR) as well as the relative rank improvement, which is defined as (SR – PR)/(SR + PR). This "improvement" could be positive, zero, or negative. If it is positive, it shows that the statistic model improves the ranking of the correct protein. In total, 19 cases (38.5%) improved ranking for the correct proteins by the *p*-value, 32 cases (61.5%) have no change, and only 1 (1.9%) produced worse ranking. The average relative improvement for the 19 protein is 41.64%, which is significant. In particular, four out of five second ranked correct proteins were enhanced to the top by *p*-values. The unchanged cases often have the correct proteins on the top, indicating good ranking can be preserved by the statistic model. Overall, 98% of the samples obtain better or preserved high ranks using *p*-values. This shows that *p*-values significantly outperform raw scores in protein identifications.

5. Software

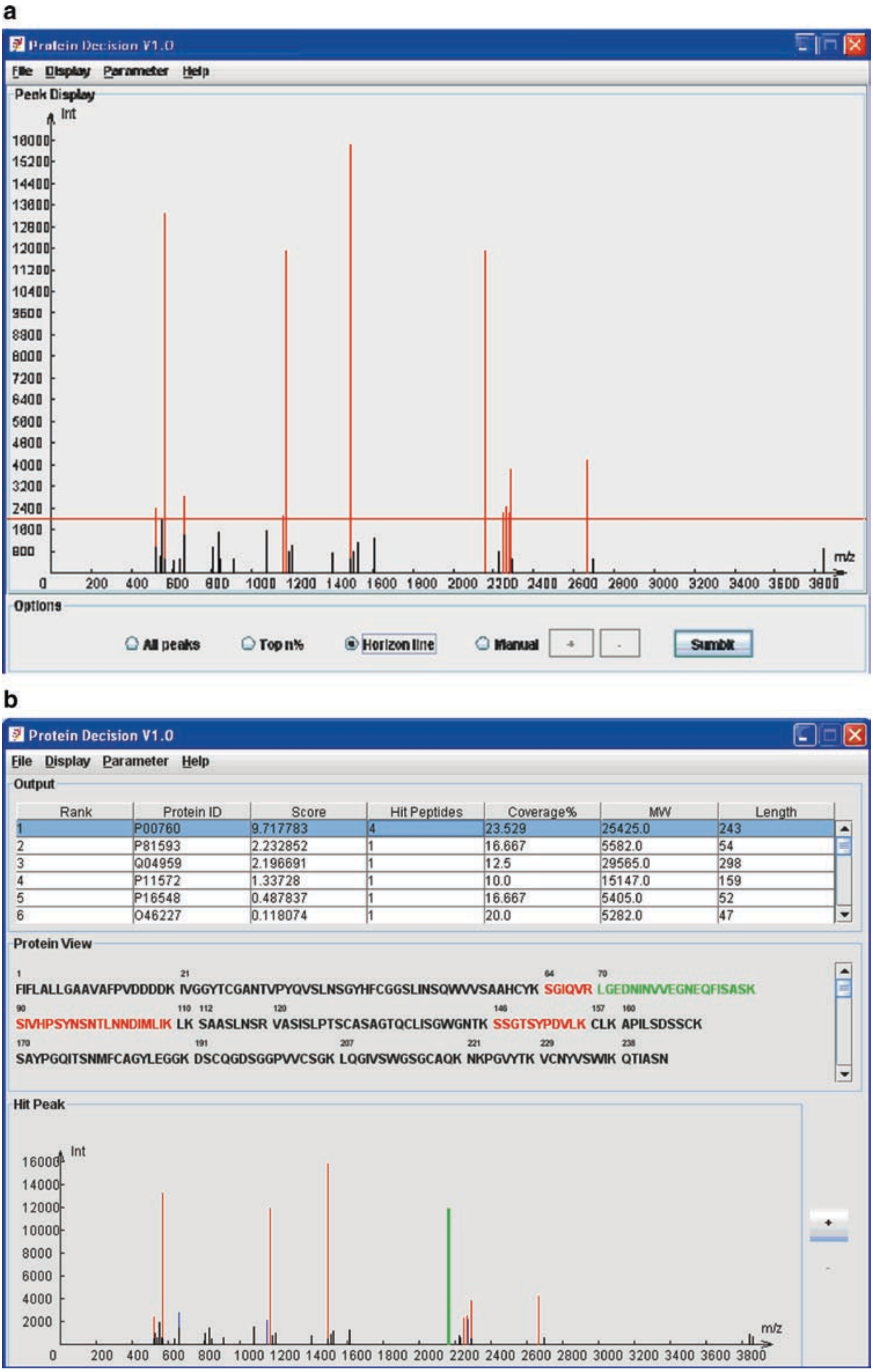
We implemented our methods in the software package "ProteinDecision" (see Note 10). The software supports the following functionalities:

1. Multiple ways for peak list selection. A spectrum peak list is required for the input file. A user-friendly interface is provided for the user to select peak list in different ways as shown in Fig. 3(a). Users can submit the entire spectrum, or select top *n*% as the input, or use a horizontal line to enclose the peaks above it, or manually select.
2. Support for the user's own sequence database in FASTA format. Users can choose and upload a search database of their own. If it is in FASTA format, the software can convert it and make it compatible for the program.
3. Flexibility in parameter selection. Users can set running parameters according to their own data's characteristics. These options include molecular weight boundaries, threshold for peptide matching, and the number of acceptable miscleavages, etc.

Results can be visualized in three subpanels as shown in Fig. 3(b). The top subpanel is a protein list ranked by their scores. A corresponding record includes rank, protein ID, score, *p*-value, the number of matched peptides, the matched peptide percentage, protein molecular weight, and the protein length. By single clicking of an entry, the sequence of that protein will be displayed in the middle subpanel and a double click will activate the internet browser, which leads to the detailed information at the ExPASy Proteomics Server. In the middle subpanel, the digestion boundaries of peptides are indicated by a space with the start and end positions labeled, and the matched peptides are marked in red. The bottom subpanel shows the spectrum. Peaks labeled in red are the selected peaks, peaks in blue represent the matched peptides, and the peak in green is the highlighted peptide.

6. Notes

1. Almost all the molecular weights of amino acids are unique. The exceptions are leucine and isoleucine, both of which have the same molecular weight but different topologies.
2. Trypsin is the most popular enzyme for protein digestion, as it cleaves very specifically at R–X and K–X bonds, except for X = P (the rule on X = P is that on some rare occasions there may be cleavage and therefore, it is not always used as a hard constraint).
3. The protein mixture can be complicated for three reasons: first, there might be more than one target protein in the spot sample; second, the enzyme itself has peptide segments, and these autolytic peptides can contaminate the target protein



Rank	Protein ID	Score	Hit Peptides	Coverage%	MW	Length
1	P00760	9.717783	4	23.529	25425.0	243
2	P81593	2.232852	1	16.667	5582.0	54
3	Q04959	2.196691	1	12.5	29565.0	298
4	P11572	1.33728	1	10.0	15147.0	159
5	P16548	0.487837	1	16.667	5405.0	52
6	O46227	0.118074	1	20.0	5282.0	47

1 FIFLALLGAAVAFPVDDDDK 21 IVGGYTCGANTVPYQVSLNSGYHFCGGSLINSQWVVSAAHCYK 64 SGIQVR 70 LGEDNINVVEGNEQFISASK

90 SIVHPSYNSNTLNNDIMLIK 110 LK 112 SAASLNSR 120 VASISLPTSCASAGTQCLISGWGNTK 146 SSGTSYPDVLK 157 CLK 160 APILSDSSCK

170 SAYPGQITSNMFCAGYLEGGK 191 DSCQGDSGGPVVCSGK 207 LQGIVSWGSGCAQK 221 NKPGVYTK 229 VCNYVSWIK 238 QTIASN

Fig. 3. Graphical user interface for ProteinDecision. (**a**) The multiways for peak selection; (**b**) the output panel for prediction result

spectrum; and third, miscleavages might happen during the digestion leading to unexpected peptides being present. In general, missed cleavages are quite common.

4. The simulated spectrum is generated on the peptide mass only (rather than m/z of higher charge states). The charge state is considered "+1" as the default. This assumption is valid for most cases in MALDI-TOF.
5. The collection of peptide masses from PMF is not absolutely unique due to the lack of explicit sequence information; however, PMF data may provide confident protein identification in many cases.
6. This is a self-defined database format for preprocessing. We provide a package to transform a FASTA sequence file into this format.
7. There are two ways to calculate the mass of amino acids and therefore peptides and proteins: average mass or mono-isotope. We use the latter one in this study.
8. Equation 5 is a sigmoid function, which is used to smooth the effect of peak intensities. The parameter α is set based on empirical estimate or by training.
9. Another solution for $P(S^K \geq s)$ is to use Gram-Charlier series with the higher-order moment information of the individual distributions. The Gram-Charlier series give an analytical expression of the eighth-order Gram-Charlier expansion for $P(S^K \geq s)$, which could be more accurate but time consuming.
10. "ProteinDecision" is a stand-alone software tool developed in Java. It incorporates PBSF scoring function and corresponding confidence assessment for protein prediction using PMF data. Currently, we have a beta version; in the long term, we will keep developing it and make it open source for free use.

Acknowledgments

This work has been supported by the MU-Monsanto Program. Zhao Song is also supported by the Shumaker Fellowship. The authors like to acknowledge the Proteomics Center at the University of Missouri-Columbia for MS services. We would like to thank Beverly DaGue, Brian Mooney, David Emerich, Gary Stacey, Jay Thelen, Chi-Ren Shyu, Guohui Lin, Chao Zhang, Yu Chen, Yang Guo, and Zhihai He for helpful discussions and/or technical assistance.

References

1. http://en.wikipedia.org/wiki/Heuristic_(computer_science).
2. Zhou, G., Li, H., DeCamp, D., Chen, S., Shu, H., Gong, Y., Flag, M., Gillespie, J., Hu, N., Taylor, P., Buck, M.E., Liotta, L.A., Petricoin III, E.C., and Zhao, Y. (2002) 2-D differential in-Gel electrophoresis for the identification of human esophageal squamous cell cancer specific protein markers. *Molecular & Cellular Proteomics* 1(2):117–124.
3. 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.
4. Harris, W.A., Reilly, P.T.A., and Whitten, W.B. (2005) Transportable real-time single-particle ion trap mass spectrometer. *Review of Scientific Instruments* 64:102–108.
5. Doucette, A. and Li, L. (2001) Investigation of the applicability of a sequential digestion protocol using trypsin and leucine aminopeptidase M for protein identification by matrix-assisted laser desorption/ionization – time of flight mass spectrometry. *Proteomics* 1:987–1000.
6. Pappin, D.J.C., Hojrup, P., and Bleasby, A.J. (1993) Rapid identification of proteins by peptide-mass fingerprinting Transportable. *Current Biology* 3:327–332.
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.R., and Burlingame, A.L. (1999) Role of accurate mass measurement (+/-10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry* 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. *Analytical Chemistry* 72:2482–2489.
10. Margnin, J., Masselot, A., Menzel, C., and Colinge, J. (2004) OLAV-PMF: A novel scoring scheme for high-throughput peptide mass fingerprinting. *Journal of Proteome Research* 3:55–60.
11. Eriksson, J. and Fenyo, D. (2004) Probity: A protein identification algorithm with accurate assignment of the statistical significance of the results. *Journal of Proteome Research* 3:32–36.
12. Ganapathy, A., Wan, X., Wan, J., Thelen, J., Emerich, D.W., Stacey, G., and Xu, D. (2004) Statistical assessment for mass-spec protein identification using peptide fingerprinting approach. *Proceedings of the 26th International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3051–3054. Sept. 1–5, 2004, San Francisco, CA.
13. Song, Z., Chen, L., Ganapathy, A., Wan, X., Brechenmacher, L., Tao, N., Emerich, D., Stacey, G., and Xu, D. (2007) Development and assessment of scoring functions for protein identification using peptide mass fingerprinting data. *Electrophoresis* 28:864–870.
14. Song, S., Chen, L., and Xu, D. (2009) Confidence assessment for protein identification by using peptide-mass fingerprinting data. *Proteomics* 9:3090–3099.

Chapter 3

Computational Approaches to Peptide Identification via Tandem MS

Simon J. Hubbard

Abstract

The peptide identification problem lies at the heart of modern proteomic methodology, from which the presence of a particular protein or proteins in a sample may be inferred. The challenge is to find the most likely amino acid sequence, which corresponds to each tandem mass spectrum that has been collected, and produce some kind of score and associated statistical measure that the putative identification is correct. This approach assumes that the peptide (and parent protein) sequence in question is known and is present in the database which is to be searched, as opposed to *de novo* methods, which seek to identify the peptide ab initio. This chapter will provide an overview of the methods that common, popular software tools employ to search protein sequence databases to provide the non-expert reader with sufficient background to appreciate the choices they can make. This will cover the approaches used to compare experimental and theoretical spectra and some of the methods used to validate and provide higher confidence in the assignments.

Key words: Bioinformatics, Peptide identification, Theoretical spectra, Data analysis, Proteomics

1. Introduction

Proteomics as a discipline has taken off in the last few years thanks to the numerous advances in mass spectrometry instrumentation, the growing numbers of genome sequences and the increasingly powerful bioinformatics tools available to analyse the data. Mass spectrometry allows peptides derived from their parent proteins to be characterised in detail in the gas phase, generating data that can be related to their amino acid sequence. This chapter aims to introduce the computational processes used by the various database search tools that attempt to achieve this, matching candidate peptide sequences to their experimental tandem mass spectra. This process underpins much of modern proteomics, whose real

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_3, © Humana Press, a part of Springer Science+Business Media, LLC 2010

goal is usually to identify the proteins themselves rather than their component peptides. However, protein identification is inferred from confident matches of mass spectra to their peptide sequences and much of the confidence and accuracy of proteomics is therefore dependent on this step.

The entire proteomics pipeline involves several stages, from preparation and separation of a complex protein sample, and digestion of the protein with a hydrolytic enzyme to produce peptides to further separation and the subsequent tandem mass spectrometry of these peptides. If high quality spectra can be obtained, then computational tools can be used to identify the amino acid sequence of the peptide from characteristic patterns in the mass spectrum. The computational challenges have been well reviewed by other authors (1–5). This chapter aims to focus specifically on the matching of experimental and theoretical spectra by the different software tools and to give a practical flavour of the parameters that the user needs to be aware of to ensure that they obtain good peptide identifications. Briefly, the background to the problem is first introduced, although the reader is referred to any of the above for more details.

As stated, the goal of most proteomics experiments is to characterise (identify and/or quantify) the component proteins in a sample. The proteins may have been extracted from a spot on a gel or a complex mixture extracted from a cell or tissue. The principal analytical technique used at present to characterise them is the mass spectrometer. Ideally, one would wish to simply analyse the protein directly and identify it from its mass spectrum. However, for a variety of reasons, protein molecules are too large and complex to analyse directly and there is no guarantee that they will ionise or correspond precisely in mass to the predicted mass calculated from their amino acid sequence. This latter point is important since modern protein sequences are usually inferred from DNA sequences and the starts/ends of the mature protein form are not necessarily known. Similarly, the post-translational state of the protein is not specified explicitly in the DNA sequence of the gene; mature proteins can be truncated, covalently linked to another chain, or more frequently modified by some other chemical group (e.g. acetylation, phosphorylation). Therefore, proteome scientists have found it much easier to break the problem down into manageable chunks, by digesting the protein into component peptides with a proteolytic enzyme (usually trypsin). Although some of the peptides will suffer from the same problems (will not ionise, have problematic modifications and do not correspond to the sequence in the database), the majority are perfectly amenable to mass spectrometric analysis and correspond exactly to the database sequence. Moreover, post-translational modifications and error-tolerant searches can be performed by most database search engines, which can take many modifications and uncertainties into account (usually at the expense of speed); such an undertaking would not be practicable for whole proteins.

Many proteomics experiments can be classed as "shotgun" proteomics, in which a proteome is digested in to component peptides and analysed, typically following one or two stages of chromatography (6). This shotgun approach is often referred to as "MudPIT" where multi-dimensional chromatography is first applied to the digested peptide mixture prior to the MS stage. Such a high-throughput proteomic technique can be compared to "shotgun" DNA sequencing. In both cases, larger molecules are fragmented into smaller components in order to piece together the solution. In the proteomic case, peptides are analysed in the mass spectrometer, often after separation in a liquid phase column by virtue of their hydrophobicity (using High Performance Liquid Chromatography – HPLC) or net charge (ion exchange chromatography). The separations can be directly coupled to the mass spectrometer, or samples may be prepared independently "offline". Once introduced in to the instrument via one of two common ionisation techniques, electrospray or Matrix Assisted Laser Desorption Ionisation (MALDI), peptide ions are analysed in the instrument. The first analysis detects the mass-to-charge (*m/z*) ratio of the peptide ions and can select particular ion species for subsequent further analysis in a second mass spectrometer, where the ion is first fragmented into smaller ions. Hence, this is where "tandem MS" is employed, where the second spectrometer analyses product ions. This step is necessary as peptides with different sequences can have very similar *m/z* values and the *m/z* value is usually not diagnostic for a single peptide (i.e. in the context of a whole genome or protein database). Fortunately, analytical chemists have been able to exploit the fact that peptides fragment in a reasonably predictable manner under certain circumstance. This dissociation is achieved using several fundamental processes such as Collisional Activation Dissociation in an inert gas (CAD) or by Electron Capture Dissociation (ECD) and Electron Transfer Dissociation (ETD). CAD is the "classical" technique and produces two predominant ion series from fragmentation at the peptide bond yielding *b* and *y* ions, whilst ECD and ETD also lead to *c* and *z* ions in the main. These ion series are shown in Fig. 1, with ions subtended at the N- and C-terminus of the peptide.

Since the masses of the atomic components are known to high precision, and the atomic composition of the amino acids are also known, it is therefore theoretically possible to work out the sequence of a peptide if a complete, noise-free, high quality product ion spectrum is obtained. Indeed, expert mass spectrometrists can often interpret spectra manually to sequence a peptide. However, as most spectra are not ideal, and high-throughput experiments typically produce many thousands of spectra, computational tools are required to automate this process.

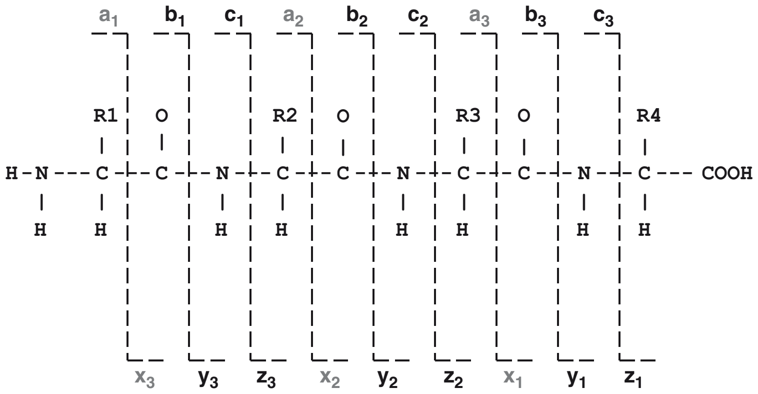


Fig. 1. Generalised product ion series generated from a peptide sequence upon fragmentation in the mass spectrometer. Collisionally activated dissociation (CAD) fragmentation generates mostly *b* and *y* ion series, which electron capture and electron transfer dissociation (ECD/ETD) instruments generated predominantly *c* and *z* ions

There are three principal approaches: (a) de novo sequencing, where the peptide sequence or a part thereof, is inferred directly from the spectra without using a database, (b), database-directed searching where the spectra is compared to candidate peptides in a protein sequence database, or (c), hybrids of (a) and (b) where tags and short inferred sequences are combined with database searching approaches. This chapter is not concerned with (a), and will concentrate instead on the approaches used in (b) as these are the most commonly employed in the majority of proteomics experiments.

2. Algorithm Overview

The overall spectrum matching problem is illustrated in Fig. 2. At first glance, this might appear to be a long and complicated process, however the aim of this figure is to illustrate the common features used in most identification pipelines and has been broken down into five stages which are described here, first in overview and then in more detail.

In the *first stage*, the spectra are collected from the mass spectrometer using the instrument settings and experimental protocol set up by the operator. Although this is not the primary focus of this chapter, the settings used will have a bearing on the quality of the spectra and the attendant success of the peptide identification protocol. Many experiments used a “data dependent acquisition” strategy, whereby the most abundant ions in the first round of MS are selected for fragmentation in a second round of MS. The complexity of the analyte will

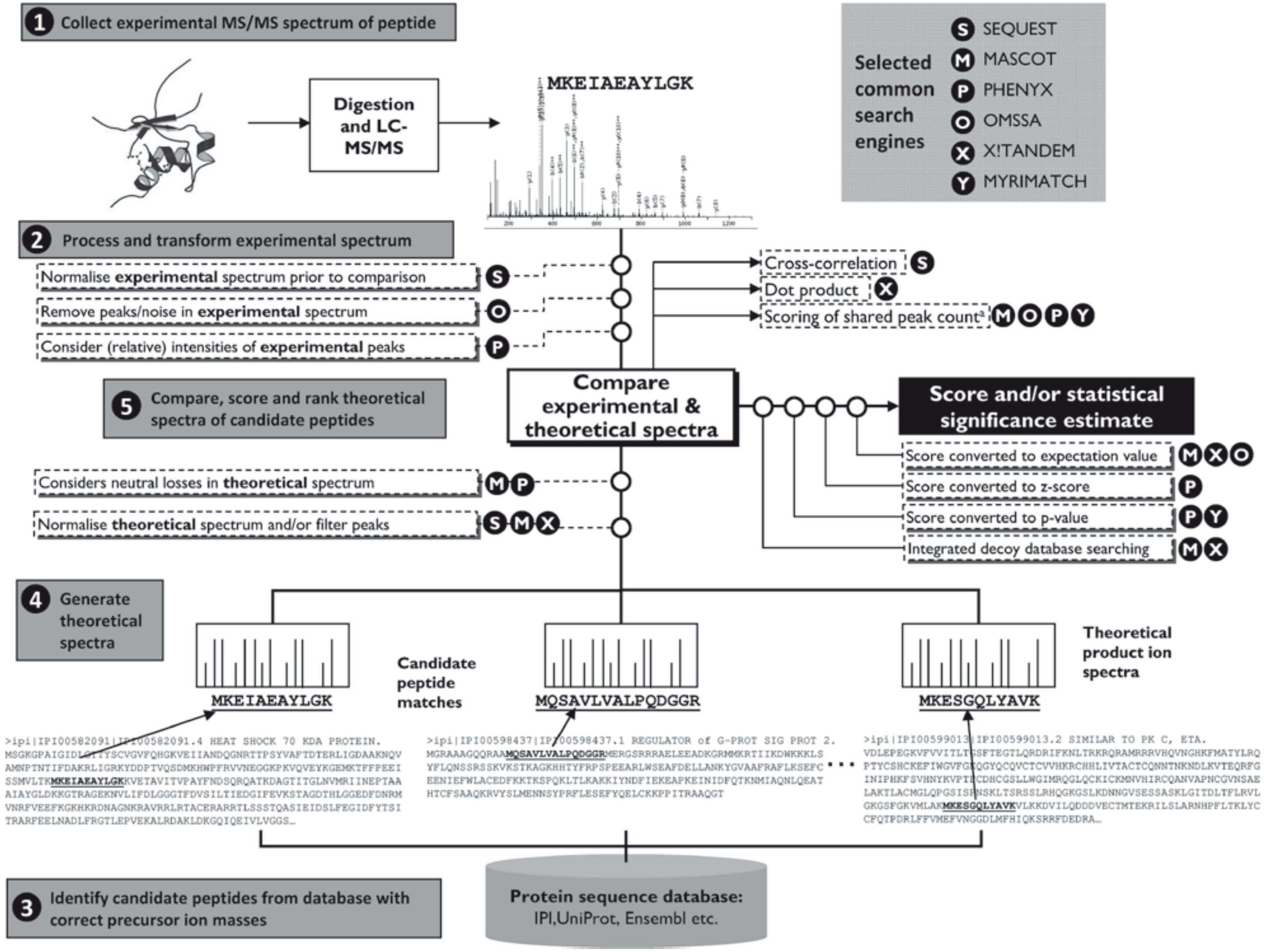

1 Collect experimental MS/MS spectrum of peptide
Digestion and LC-MS/MS
MKEIAEAYLGK
Selected common search engines
S SEQUEST
M MASCOT
P PHENYX
O OMSSA
X X!TANDEM
Y MYRIMATCH
2 Process and transform experimental spectrum
Normalise experimental spectrum prior to comparison S
Remove peaks/noise in experimental spectrum O
Consider (relative) intensities of experimental peaks P
Cross-correlation S
Dot product X
Scoring of shared peak count[a] M O P Y
5 Compare, score and rank theoretical spectra of candidate peptides
Compare experimental & theoretical spectra
Score and/or statistical significance estimate
Score converted to expectation value M X O
Score converted to z-score P
Score converted to p-value P Y
Integrated decoy database searching M X
Considers neutral losses in theoretical spectrum M P
Normalise theoretical spectrum and/or filter peaks S M X
4 Generate theoretical spectra
Candidate peptide matches
Theoretical product ion spectra
MKEIAEAYLGK
MQSAVLVALPQDGGR
MKESGQLYAVK
>ipi|IPI00582091|IPI00582091.4 HEAT SHOCK 70 KDA PROTEIN.
>ipi|IPI00598437|IPI00598437.1 REGULATOR of G-PROT SIG PROT 2.
>ipi|IPI00599013|IPI00599013.2 SIMILAR TO PK C, ETA.
3 Identify candidate peptides from database with correct precursor ion masses
Protein sequence database: IPI,UniProt, Ensembl etc.

Fig. 2. Schematic of overall process of peptide identification from experimental spectra to statistical significance of candidate peptide sequence matches (PSMs). Each stage is number and various protocols are indicated underneath throughout the figure. In the top *right*, the panel refers to the various search engines discussed here with a single letter in a circle signifying each engine. These one-letter codes are used to indicate which search engines carry out which process steps in the overall schema

therefore affect the volume and quality of the spectra acquired. Various peak detection and spectral processing pipelines may then be applied, often using software provided by instrument manufacturers. This includes noise reduction, background subtraction, centroiding and smoothing.

Once a spectrum has been acquired, the *second stage* involves further processing of the experimental spectrum prior to any comparison with any theoretical ones. This can involve normalisation, peak binning and removal of peaks to provide consistency in the spectra and peak intensities, which are subsequently used in matching algorithms. Once experimental spectra are processed, they can be compared to theoretical spectra generated from the database(s) selected for the search.

The *third stage* scans through the protein sequence databanks to be searched, usually selecting candidate peptides based on their precursor ion mass for comparison to the experimental spectrum. Typically, this process is also restricted to peptides generated from a theoretical digest of the peptide sequence databank by a named enzyme (e.g. trypsin), as well as to peptide masses that would generate a *m/z* within a fixed tolerance of the experimental precursor ion for a limited set of possible charge states. Some algorithms, such as Sequest, also employ a simple filtering step based on the number of peak matches to the theoretical spectra, although this usually improves only speed not performance. Indeed, this too may be unnecessary as Sequest search speed has been recently improved (7).

In the same way that the experimental spectra are processed, the theoretical spectra can be normalised and processed to include/exclude particular fragment ion peaks at the *fourth stage*.

In the *fifth stage*, a variety of mathematical functions can then be used to compare each experimental spectrum to the theoretical spectra generated for all candidate peptide sequence matches (PSM). This latter abbreviation is used quite frequently in the literature to describe candidate matches between spectra and peptide sequences and represents an important point. These algorithms generate a list of candidate matches, typically ranked on some scoring function, which may represent the actual amino sequence from which the experimental spectrum was generated. The term peptide “identification” is used frequently in the field, usually referring to the top-hit in this ranked list (which may or may not exceed some statistical measure of significance). However, it is possible for a variety of reasons that this could be a false positive and hence strictly we should refer to hits as “putative identifications”. Since this is rather an important point, the informatics field allied to proteomics fields has invested considerable time and effort developing statistical models and significance tests to assign *p*-values and expectation values to candidate PSMs.

A variety of techniques are applied in the *sixth stage*, some of which are in-built with search tools, which attempt to assign some such statistical or probabilistic measure of the likelihood of a candidate match being correct or being assigned by chance.

All the above are discussed in more detail in the following sections, with particular reference to the common search tools and how they handle the problem. Although there are many different search tools, with new programs appearing all the time, this review will concentrate on a subset of the most widely used, namely: Sequest (8), Mascot (9), OMSSA (10), X!Tandem (11), MyriMatch (12) and Phenyx (13). Other notable tools not covered here in detail include VEMS, InsPecT, ProbID, Crux and Paragon (14–18).

3. Experimental Spectral Processing

Different software tools take different approaches here, depending on the nature of the comparison technique they use with theoretical spectra. For example, Sequest (8) pre-processes both the experimental and theoretical spectrum. The signal intensities are normalised to local maxima, after first assigning the peak intensities to bins across the *m/z* scale. This step effectively ensures that the experimental spectrum is comparable with the theoretical one. In most MS/MS spectra, the *b* and *y* ions dominate over *a* ions and other species. Although there is no prior assignment of ion types to the spectral peaks in the experimental spectra, this is done for the theoretical spectra and the normalisation step ensures that the most abundant ion intensities in the experimental spectrum are comparable to the theoretical ones. The standard Sequest algorithm does not explicitly consider neutral losses (of water or ammonia) and identifications are therefore essentially based on *a*, *b* and *y* ions only. OMSSA (10) does not explicitly normalise spectra but does remove peaks it considers to be noise, attempting to maximise the signal:noise before any comparison with theoretical peaks, deleting peaks below 2.5% of the maximal signal intensity (although the 2.5% value can be changed by the user as well as dynamically by the algorithm). Additionally, OMSSA attempts to "de-isotope" the spectrum by removing peaks which appear not to be monoisotopic. The authors claim that this improves the performance of their algorithm substantially since peaks which are non-monoisotopic will complicate spectral comparisons as most algorithms do not model them in the theoretical spectra.

MyriMatch follows an alternative route, by filtering the experimental spectrum by a percentage of the total ion current, ensuring that a fixed percentage of ion signal is retained rather than removing peaks below a fixed cutoff (12). The spectral peaks are then binned in to a small number of user-definable intensity classes (in a default ratio of 2:1 for adjacent classes). These classes are subsequently used in the scoring function to upweight candidate peptides which match more high-intensity binned peaks.

Other peptide identification tools take a more sophisticated approach to model the experimental spectrum. Phenyx uses a probabilistic model for experimental spectra which includes consideration of relative intensities of given ion types, presence of ion series, missed cleavage propensities, and post-translational modification likelihoods (13, 19). This model can be trained further in a user's laboratory to improve performance for a local instrument/protocol set up.

Most algorithms also filter or remove ions from around the region of the spectrum containing the precursor ion *m/z*. This is done to prevent any theoretical product ions matching to the precursor ion and leading to false positive identifications. The necessity to perform extensive data processing on the experimental spectra clearly depends on the instruments software and the identification tool provided. Tools such as Mascot and X!Tandem do not use extensive data processing steps to the experimental spectra but work well regardless. The quality of the identifications obtained may still be improved, however, if sensible steps are taken to reduce noise prior to database searching as has been reported (20). Other approaches also address this by removing redundant or poor quality spectra using clustering or machine learning techniques (21–26).

4. Selection of Database Search Parameters

Very poor results can be obtained from high quality data if users do not take care to select search parameters appropriately. Even the smallest mis-setting of a parameter can reduce the number of confident identifications to a small fraction of those that are possible. Some of the key parameters are outlined subsequently.

1. Mass type. Although most modern mass spectrometers have sufficient resolution to be able to routinely determine monoisotopic peptide and fragment *m/z* values from spectra, large molecules and older instruments may not always lead to such good resolution. Monoisotopic ions contain only the ^{12}C (as well as the ground isotopic states of all other constituent elements), while average mass determinations are called from ^{13}C and other higher state isotopes and will always be larger than the monoisotopic values. Most search engines ask the user to specify the mass type they wish to search for.

This might seem trivial, but setting a search engine to look for monisotopic masses when average masses have been determined (and vice versa) will seriously degrade search performance and can catch out novice users.

2. Fragment ions types. As discussed earlier, the types of theoretical ion generated by the search engines can be limited by the user with some tools. Most tools also offer sensible protocols/default settings. For example, Mascot restricts ion types it expects to see based on the instrument type selected on the front page of the web-front end to the search engine. Some users pay little attention to this box, but it is prudent to select the closest instrument type to the one in use, to improve search results. It is also possible to customise the ions types to be searched if so desired with many search engines if the user has access to them.

3. Precursor ion charge state. Another important feature to consider is whether to trust the charge state inferred by the instrument software on the precursor ion. This charge will determine the possible fragment ions that can be produced and will have a marked bearing on the ions series produced (27). Chapter 6 also examines fragmentation propensities at length and relates this to peptide identification and search engine strategies. Most search engines support multiple charge state searches for multiply charged ions, considering both +2 and +3. OMMSA, for example, determines charge state based on the fraction of peaks which are greater than the precursor *m/z*. If 95% of the peaks are above this value, it assumes it is a +1 charged ion, otherwise it assumes it is multiply charged and searches first +2 and then +3. Mascot's default behaviour is to consider +2 and +3 for unknown charge states from electrospray experiments. MyriMatch also treats 3+ charged ions as special cases and models the theoretical product ion spectra differently depending on the amino acid content.

4. Database selection. This is an important choice that can radically change the numbers of matches returned, and is discussed in detail in Subheading 5.

5. Enzyme and properties. Although trypsin is the enzyme of choice in many instances, other proteolytic enzymes (such as endoproteinase Lys-C) have been used in some high profile experiments, e.g. (28). Similarly, chemical modification of the proteome can block some groups and cause the protease to behave effectively like another. For example, in a recent protocol aimed at N-terminal peptide study, McDonald and colleagues acetylate all free alpha-amino groups (including those on lysines) and a tryptic digestion effectively becomes a

"Arg-C" digest (29, 30). Although most enzymes are highly specific, proteome peptides can frequently arise which appear to be the results of non-specific cleavage. Trypsin cleaves after lysine and arginine (K and R in the amino acid one letter code) but apparently not when they are followed by a proline (P). This has raised some controversy in the field as some groups consider such cleavages to occur more frequently than might be expected (31), and others that they are predominantly the results of instrumental artefacts (ie. not enzyme digestion) (32). Regardless, most search engines support the option to consider KP and RP cleavages. Similarly, some will also allow the user to search for "semi-specific" cleavage, where only one end of the peptide conforms to the enzyme specificity rules. Although these searches will be much slower, they have the advantage of being able to detect processed peptides such as those produced from some internal cleavage or processing event e.g. after removal of signal peptides or pre-/pro-peptides.

Finally, the user needs to consider whether "missed cleavages" might have occurred where the enzyme has not digested the protein to all its limit peptides. These missed cleavages are common and search engines allow users to consider up to some fixed number. Since they are common, it is prudent to consider up to 1 or 2 missed cleavages if it is important to maximise the number of peptide identifications, but again it increases the search space at cost of reduced speed (33).

6. Chemical and post-translational modifications. All search engines can deal with both fixed and variable modifications which can be introduced either during the processing of the protein sample (chemical modifications) or by the organism after translation (post-translational modifications). The search engines usually consider these as either "fixed" or "variable." Fixed modifications are applied to all peptides containing the appropriate group, while variable modifications are applied to all possible combinations. For example, if there are three modifiable groups with the addition of +17 Da then the m/z, m/z+17, m/z+34, m/z+51 values will all be considered matches for that ion.

5. Selection of Search Database

The protein sequence database that is searched should be selected appropriately and with the aim of the experiment in mind. Although a full description of the various protein sequence databanks is beyond the scope of this chapter, it is worth making a few basic points. Protein sequence databanks are derived from nucleotide sequencing of the organism's genes and genome and hence it should be recognised that the protein sequences being searched do not necessarily correspond 100% to the actual protein sequence expressed in the cell or tissue under study. The reasons for a mismatch include post-translational modifications, proteolytic processing (signal and pre-/pro-peptide removal), polymorphisms, sequencing errors, and indeed absence from the available genome sequenced to date – few genomes are completely sequenced and annotated down to the last base pair and sometimes large blocks of sequence are missing. The reasons for failing to find your protein in a proteomics experiment are many, and it is incorrect to assume that because a given peptide is absent from an experiment that the parent protein is not present in a sample. Equally, most vertebrate proteins are expressed in multiple isoforms (34) and hence it can often be difficult to differentiate precisely which (if any) of the particular isoforms is being expressed. These types of considerations need to be taken into account when selecting a protein databank for searching. By default, most search tools use large databases concatenated from several member databases, such as NCBI's *nr* and Mascot's MSDB. Although comprehensive, these databases may not necessarily contain all the most up-to-date candidate sequences from a particular recent genome sequencing experiment (if it has not been publically released) and they are slower to search because of their size. Equally, they contain considerable redundancy and can lead to type I errors (false positives) because of their size. The converse problem can be produced with small, organism focussed databases which typically contain 10,000-30,000 genes (compared to the one million or so in MSDB). The size of these databases can lead to an increase in false negatives since some isoforms/variants may be missing from the databank. Equally, smaller databases can also lead to increased false positives (depending on the statistical model used to calculate scores/significance). For example, when using decoy database (*see* Subheading 8) searches, a small database may produce a poor decoy model and over-estimate the significance of peptide matches. Mascot's threshold uses the number of candidate peptides whose *m/z* is within the error tolerance of the experimental precursor ion to calculate its scores and redundancy can influence this.

The IPI databases (35) curated by the EBI (http://www.ebi.ac.uk/IPI) represent a good compromise, storing most known isoforms, providing cross-references to member databases, and a high level of annotation. A database for each individual proteome relating to a specific organism is provided by the IPI curators and is therefore a good choice if the species of the sample is known and sequenced. However, if not, then one of the more general databases might be more appropriate. Several search tools support the restriction of the search to known taxonomies or species,

Table 1
Mascot database search statistics for collections of spectra from example proteins

Protein	Species	Database	Taxonomic filter	Mascot threshold	True positives[a]	False positives[b]	Spectra
Molecular chaperone dnaK	*E. coli*	MSDB	None	48	11	15	87
		MSDB	*E. coli*	28	22	0	
		E. coli K12	None	19	24	1	
Ubiquitin protein ligase	Human	MSDB	None	49	8	7	129
		MSDB	*Homo sapiens*	35	28	1	
		IPI Human	None	33	32	1	

[a] Peptide-spectrum matches where the Mascot ion score exceeds the threshold and the peptide belongs to the protein listed in column 1

[b] Peptide-spectrum matches where the Mascot ion score exceeds the threshold but the peptide does not apparently belong to the protein listed in column 1. It is possible that a fraction of these are true positives and the protein sequence and annotation was not sufficiently clear to resolve its identity

or this can be achieved by combining individual sequence databases to generate more complex datasets. Of course, since it is possible to download and run the open source search tools in your own laboratory, one can therefore search any custom database.

By way of a simple example to illustrate how database size can influence search results, Table 1 shows some search results for a large number of candidate spectra of variable quality searched against different databases for a bacterial and human example. The Mascot search engine is shown as an exemplar as it calculates its scoring thresholds (see Subheading 7) based on the number of matching precursor ions in the database. In a large unfiltered database, a higher threshold is obtained and fewer true positive spectra exceed the nominal significance threshold that Mascot calculates. Restricting searches to species specific subsets or using a species-specific database improves the number of true positive identifications. It should be noted that it is not always straightforward to determine whether a PSM is a true/false positive when annotations are missing and hence the false positive columns may be overestimates. Nevertheless, the choice of database clearly has a marked effect on the search outcome.

6. Processing of Theoretical Spectra

Most search tools also perform some processing on the theoretical spectra. This can take the form of inclusion/exclusion of ion types. Usually, *b* and *y* type ions from CID experiments are included, but most tools also allow the user to consider other ion types (such as *a* ions and immonium ions). Others such as Mascot also consider neutral losses, depending on the amino acid content of the fragment ion concerned. If all the possible fragment ion peaks are generated, the theoretical spectra would be too complex and would match too many experimental spectra; search engines therefore take a more simplified view, summarised in Table 2. Sequest normalises the theoretical spectra like it does the experimental one, constraining *b, y* ions to have higher intensity than other ion types. OMSSA only considers *b* and *y* ions and counts the matches in sorted mass ladders between theoretical and experiment spectra. MyriMatch adopts a similar principle since its scoring is based on whether peaks are matched or not, and does not consider ion intensity explicitly. At the other extreme, Phenyx uses a sophisticated probabilistic model which includes the consideration of fragment ion intensities as well as a full range of ion types, as already mentioned.

Other authors have analysed the properties of high quality peptide identifications, examining properties such as enzymatic cleavage preferences (33), amino acid preferences at fragmentation sites (36–38), and indeed the intensities of product ions (39–41). However, as yet, these results have not really made an impact on the popular search engines and predicted fragmentation patterns and ion intensities have not been fully exploited.

7. Spectral Comparison Methods and Scoring

The actual comparison of the processed spectra forms the core of the peptide identification pipeline and different search engines approach this in different ways. Sequest was the first algorithm to be published which could make assignments to uninterpreted spectra and uses a cross-correlation model which essentially measures the similarity between the theoretical and experimental spectrum factoring small shifts between the two (7, 8, 42). This cross-correlation score, popularly known as the *Xcorr* score, is accompanied by a related score, ΔCn, which measures the difference between the top scoring match and the next highest scoring PSM after normalising the top *Xcorr* score to 1. This method does not explicitly take into account ion intensities but does do this implicitly since spectra are normalised and major ion series (*b* and *y*) weighted more highly. With the exception of Phenyx, most other tools do not make an attempt to model signal intensity explicitly in their comparisons. OMMSA and X!Tandem focus on the number of matching peaks between the two series. OMMSA does not consider peak intensities at all once the spectra have been noise filtered and pre-processed and bases the score on the

Table 2
Summary of properties of popular search engines

Algorithm	Spectral comparison	Spectral normalisation	Ions considered	Intensities considered	Accounts for noise	Semi-tryptic	Scoring
SEQUEST	Cross-correlation	Yes	*b*, *y* and *a* (neutral losses in commercial version)	Implicitly	Implicit	Yes	Xcorr and ΔC_n
MASCOT	Probabilistic	?	All (instrument specific definitions can be applied)	No	Presumed, implicit	Yes	Ion Score and E-value
X!Tandem	Dot product	Yes	*b* and *y* only	No	No	Yes	Hyperscore and E-value
OMSSA	Number of matched product ion peaks	No	*b* and *y* (1 forward, 1 reverse) ECD/ETD version also now available	No	Explicit, removing low intensity peaks, and those around principal ion peaks	Yes	E-value
Phenyx	Log-odds likelihood score	No	All	Explicitly	No	Yes	Logs score, *p*-value and Z-score
MyriMatch	Probabilistic	Yes, into intensity bins	B and y	Implicitly via binning	Explicit, using % of total ion current threshold	Yes	Probabilistic *p*-value of random match

number of matching peaks cast against an expected Poisson distribution of scores (10). OMMSA is also able to deal with *c* and *z* ions in spectra generated from ETD/ECD MS. X!Tandem defines a hyperscore based on the dot product of the matched intensities which is modified by the number of matched *b* and *y* ions. Hence, other ion types are not formally considered but there is implicit modelling of ion intensities; indeed, in part, the score is simply a sum of matched ion intensities of the *b* and *y* ions.

Mascot's scoring algorithm is not published, but it is reported to be probabilistic and clearly takes into account more than just *b* and *y* ions. It is also able to cope with ETD/ECD spectra. For each candidate PSM, an ion score is reported reflecting the likelihood that a match is true. It also calculates two thresholds which users may compare their ion scores to, reflect a nominal "identity" and slightly lower "homology" threshold. These are derived from the number of precursor ions passing the MS tolerance filter. The "identity" threshold is estimated to give a 1 in 20 chance of a false positive PSM. Phenyx's scoring algorithm is centred on the likelihood of observing a given PSM based on the observed peak matches, ion series, relative peak intensities, post-translational modifications, peptide/production mass errors, number of missed cleavages and amino acid composition. This score therefore represents the likelihood of observing any given match by chance. It is subsequently transformed into a Z-value. This is the number of standard deviations from the mean score of random peptides and therefore assumes that the log likelihood scores are normally distributed.

Finally, MyriMatch uses a multi-modal model of spectral peak intensities from which to estimate the probability *p* of a random match. This *p*-value is transformed into a score by taking the negative natural logarithm. As mentioned earlier, peak intensities are implicitly considered by the binning of peaks into intensity class bins which are considered in the calculation of the probability. This differentiates MyriMatch from algorithms such as OMSSA which only consider ladders of matched product ions and ignore signal intensity.

8. Evaluating Significance

Although all the algorithms generate some fundamental score, none of the core metrics give a statistical description of whether a match is significant or not. Most search engines offer some way to do this now, or their output can be used in third-party software which does this for you; e.g. PeptideProphet delivered as part of the TPP suite (43, 44). Mascot, OMMSA, X!Tandem and Phenyx all report expectation values (or similar statistical scores) which effectively report the number of matches that would be found by chance with a given score in the selected database. This popular formulism is well known to most biologists from the sequence searching in programs such as BLAST (45). However, it is clear that the different tools produce very different E-value estimates that are not, as would be the intention with such a statistic, directly comparable (46). However, it is possible to re-scale and combine such statistics and obtain a greater overall number of high confidence peptide identifications (47). Additionally, the output from multiple search engines can be combined in a Bayesian framework to generate a greater volume of high confidence peptide identifications in the SCAFFOLD package (48). A more detailed discussion of the statistics of search engines is available in a companion volume to this one (49).

Another recent advance in determining the significance of peptide hits exploits a decoy database search strategy to estimate false discovery rates (FDRs). FDRs have become popular in proteomic search strategies and have generated much controversy and coverage in the literature (50–57). The process uses a reversed or randomised database (the "decoy" database) in parallel with the standard database and all hits to the decoy are assumed to be false. From the results, FDRs can be estimated for a given score cut-off. An entire chapter is devoted to the subject in this volume (see chapter 5). Despite the controversy, search engines including Mascot and X!Tandem have integrated decoy search strategies into their tools.

Finally, it is worth commenting that all peptide identifications are usually also considered at the protein level. This is problematic due to peptide redundancy in proteins; multiple protein isoforms can contain the same peptide sequence and short sequences can recur in several unrelated proteins in a given proteome. This problem is usually referred to as the protein inference problem. A detailed review is beyond the scope of this article and interested readers are referred to a recent excellent publication (58). Most search engines provide a fairly rudimentary approach to resolving the problem although some of the commercial tools available with instruments address the problem well. Mascot has its own "MudPIT" scoring which attempts to down-weight proteins with large numbers of poor quality PSMs. There are also generic approaches; the Average Peptide Scoring (APS) method, used in several recent studies (59–61) can evaluate FDRs at the protein level after a decoy database search.

9. Summary

Inevitably as a researcher in this field, one must address the question: which search engine should I use? It is clear that they all have their advantages and disadvantages and it is not possible to recommend a single search engine above all others. Moreover, as discussed, many labs have discovered that improved performance can be obtained by combining the outputs from several together (46–48). There is an increasing pressure on authors to publish their data to minimum reporting standards (62, 63), to carry out stringent statistical quality control on their peptide identifications (50) and also to lodge their data in public repositories (64). These topics are also covered in detail in other chapters in this volume (Chapters 19–24). By following the best practice and doing more than just pressing "search" on the website, these goals can be met and high quality proteomics data can become widely available in repositories in the same way that we take high quality sequence data for granted.

Acknowledgments

The author would like to thank Jenny Siepen, Julian Selley and Andrew Jones for useful comments on the manuscript, and BBSRC for support from various research grants (BB/E024912/1, BB/F004605/1).as well as the EU ProDaC grant (European Commission project, 6th framework programme, project number LSHG-CT-2006-036814).

References

1. Colinge, J., and Bennett, K. L. (2007) Introduction to computational proteomics. *Plos Computational Biology* **3,** 1151–60.
2. Hernandez, P., Muller, M., and Appel, R. D. (2006) Automated protein identification by tandem mass spectrometry: Issues and strategies. *Mass Spectrometry Reviews* **25,** 235–54.
3. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology* **5,** 699–711.
4. Veltri, P. (2008) Algorithms and tools for analysis and management of mass spectrometry data. *Briefings in Bioinformatics* **9,** 144–55.
5. Webb-Robertson, B. J. M., and Cannon, W. R. (2007) Current trends in computational inference from mass spectrometry-based proteomics. *Briefings in Bioinformatics* **8,** 304–17.
6. 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,** 242–7.
7. Eng, J. K., Fischer, B., Grossmann, J., and MacCoss, M. J. (2008) A fast SEQUEST cross correlation algorithm. *Journal of Proteome Research* **7,** 4598–602.
8. 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. *Journal of the American Society for Mass Spectrometry* **5,** 976–89.

9. 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,** 3551–67.
10. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X. Y., Shi, W. Y., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *Journal of Proteome Research* **3,** 958–64.
11. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–67.
12. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research* **6,** 654–61.
13. Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3,** 1454–63.
14. Park, C. Y., Klammer, A. A., Kall, L., MacCoss, M. J., and Noble, W. S. (2008) Rapid and accurate peptide identification from tandem mass spectra. *Journal of Proteome Research* **7,** 3022–27.
15. Shilov, I. V., Seymour, S. L., Patel, A. A., Loboda, A., Tang, W. H., Keating, S. P., Hunter, C. L., Nuwaysir, L. M., and Schaeffer, D. A. (2007) The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Molecular & Cellular Proteomics* **6,** 1638–55.
16. Tanner, S., Shu, H. J., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: Identification of posttransiationally modified peptides from tandem mass spectra. *Analytical Chemistry* **77,** 4626–39.
17. Zhang, N., Aebersold, R., and Schwilkowski, B. (2002) ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2,** 1406–12.
18. Matthiesen, R., Trelle, M. B., Hojrup, 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. *Journal of Proteome Research* **4,** 2338–47.
19. Colinge, J., Masselot, A., Cusin, I., Mahe, E., Niknejad, A., Argoud-Puy, G., Reffas, S., Bederr, N., Gleizes, A., Rey, P. A., and Bougueleret, L. (2004) High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. *Proteomics* **4,** 1977–84.
20. Samuelsson, J., Dalevi, D., Levander, F., and Rognvaldsson, T. (2004) Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics* **20,** 3628–35.
21. Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., and Pevzner, P. A. (2008) Clustering millions of tandem mass spectra. *J Proteome Res* **7,** 113–22.
22. Salmi, J., Moulder, R., Filen, J. J., Nevalainen, O. S., Nyman, T. A., Lahesmaa, R., and Aittokallio, T. (2006) Quality classification of tandem mass spectrometry data. *Bioinformatics* **22,** 400–6.
23. Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., and Yates, J. R., 3rd (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* **75,** 2470–7.
24. Tabb, D. L., Thompson, M. R., Khalsa-Moyers, G., VerBerkmoes, N. C., and McDonald, W. H. (2005) MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J Am Soc Mass Spectrom* **16,** 1250–61.
25. Wong, J. W., Sullivan, M. J., Cartwright, H. M., and Cagney, G. (2007) msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. *BMC Bioinformatics* **8,** 51.
26. Beer, I., Barnea, E., Ziv, T., and Admon, A. (2004) Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **4,** 950–60.
27. Huang, Y. Y., Triscari, J. M., Tseng, G. C., Pasa-Tolic, L., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2005) Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Analytical Chemistry* **77,** 5800–13.
28. de Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Frohlich, F., Walther, T. C., and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251–U60.
29. McDonald, L., and Beynon, R. J. (2006) Positional proteomics: preparation of amino-terminal peptides as a strategy for proteome simplification and characterization. *Nature Protocols* **1,** 1790–98.

30. McDonald, L., Robertson, D. H. L., Hurst, J. L., and Beynon, R. J. (2005) Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nature Methods* **2,** 955–57.

31. Rodriguez, J., Gupta, N., Smith, R. D., and Pevzner, P. A. (2008) Does trypsin cut before proline? *Journal of Proteome Research* **7,** 300–05.

32. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics* **3,** 608–14.

33. Siepen, J. A., Keevil, E. J., Knight, D., and Hubbard, S. J. (2006) Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *Molecular & Cellular Proteomics* **5,** 1350.

34. Modrek, B., and Lee, C. J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics* **34,** 177–80.

35. 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,** 1985–88.

36. Breci, L. A., Tabb, D. L., Yates, J. R., 3rd, and Wysocki, V. H. (2003) Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal Chem* **75,** 1963–71.

37. Tabb, D. L., Huang, Y., Wysocki, V. H., and Yates, J. R., 3rd (2004) Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal Chem* **76,** 1243–8.

38. Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H., Lin, D., and Yates, J. R., 3rd (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem* **75,** 1155–63.

39. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., and Gygi, S. P. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat Biotechnol* **22,** 214–9.

40. Gehrke, A., Sun, S., Kurgan, L., Ahn, N., Resing, K., Kafadar, K., and Cios, K. (2008) Improved machine learning method for analysis of gas phase chemistry of peptides. *BMC Bioinformatics* **9,** 515.

41. Zhou, C., Bowler, L. D., and Feng, J. (2008) A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC Bioinformatics* **9,** 325.

42. MacCoss, M. J., Wu, C. C., and Yates, J. R., 3rd (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal Chem* **74,** 5593–9.

43. 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,** 5383–92.

44. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75,** 4646–58.

45. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25,** 3389–402.

46. Balgley, B. M., Laudeman, T., Yang, L., Song, T., and Lee, C. S. (2007) Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol Cell Proteomics* **6,** 1599–608.

47. Jones, A. R., Siepen, J.A., Hubbard, S.J., Paton, N.W. (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* **9,** 1220–9.

48. Searle, B. C., Turner, M., and Nesvizhskii, A. I. (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J Proteome Res* **7,** 245–53.

49. Nesvizhskii, A. I. (2007) Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol Biol* **367,** 87–119.

50. Choi, H., and Nesvizhskii, A. I. (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J Proteome Res* **7,** 47–50.

51. Kall, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res* **7,** 29–34.

52. Kall, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res* **7,** 40–4.

53. Kim, S., Gupta, N., and Pevzner, P. A. (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res* **7,** 3354–63.

54. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods* **4,** 787–97.

55. 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.

56. Tabb, D. L. (2008) What's driving false discovery rates? *J Proteome Res* **7,** 45–6.

57. Wang, G., Wu, W. W., Zhang, Z., Masilamani, S., and Shen, R. F. (2009) Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Anal Chem* **81,** 146–59.

58. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4,** 1419–40.

59. Chepanoske, C. L., Richardson, B. E., von Rechenberg, M., and Peltier, J. M. (2005) Average peptide score: a useful parameter for identification of proteins derived from database searches of liquid chromatography/ tandem mass spectrometry data. *Rapid Commun Mass Spectrom* **19,** 9–14.

60. Shadforth, I., Dunkley, T., Lilley, K., Crowther, D., and Bessant, C. (2005) Confident protein identification using the average peptide score method coupled with search-specific, ab initio thresholds. *Rapid Commun Mass Spectrom* **19,** 3363–8.

61. Wright, J. C., Sugden, D., Francis-McIntyre, S., Riba-Garcia, I., Gaskell, S. J., Grigoriev, I. V., Baker, S. E., Beynon, R. J., and Hubbard, S. J. (2009) Exploiting proteomic data for genome annotation and gene model validation in Aspergillus niger. *BMC Genomics* **10,** 61.

62. Taylor, C. F. (2006) Minimum reporting requirements for proteomics: a MIAPE primer. *Proteomics* **6 Suppl 2,** 39–44.

63. Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A., Julian, R. K., Jr., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., Seymour, S. L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T. M., Whitelegge, J. P., Wilkins, M. R., Xenarios, I., Yates, J. R., 3rd, and Hermjakob, H. (2007) The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* **25,** 887–93.

64. Mead, J. A., Shadforth, I. P., and Bessant, C. (2007) Public proteomic MS repositories and pipelines: available tools and biological applications. *Proteomics* **7,** 2769–86.

Chapter 4

Scoring and Validation of Tandem MS Peptide Identification Methods

Markus Brosch and Jyoti Choudhary

Abstract

A variety of methods are described in the literature to assign peptide sequences to observed tandem MS data. Typically, the identified peptides are associated only with an arbitrary score that reflects the quality of the peptide-spectrum match but not with a statistically meaningful significance measure. In this chapter, we discuss why statistical significance measures can simplify and unify the interpretation of MS-based proteomic experiments. In addition, we also present available software solutions that convert scores into sound statistical measures.

Key words: Peptide identification, Statistical significance, False discovery rate, *q*-Value, Posterior error probability, Percolator, PeptideProphet, Qvality

1. Introduction

Mass spectrometry (MS) has become the method of choice for protein identification and quantification offering high-throughput analysis at high sensitivity (1). Most proteomics studies are based on shotgun sequencing, wherein proteins are proteolytically digested into peptides and subsequently analyzed by tandem MS. In this way, peptide mass to charge ratios are determined, and selected ions are isolated and fragmented to generate product ion mass spectra (2).

A large number of computational tools have been developed to support high-throughput peptide and protein identification by automatically assigning sequences to tandem MS spectra ((3), table 1). Three types of approaches are used: (a) database searching, wherein all peptide candidate sequences selected from an in-silico digested protein sequence database are investigated at the MS/MS

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_4, © Humana Press, a part of Springer Science+Business Media, LLC 2010

level by correlating the experimental and the theoretical peptide fragmentation patterns; (b) de novo sequencing, wherein peptides are inferred from spectra without prior knowledge of protein sequences; (c) hybrid approaches, wherein short sequence tags are inferred by de novo methods and subsequently searched against protein databases.

With the constant advances in instrument technology and improved algorithms, de novo and hybrid methods may have a more important role in the future; however, database searching remains the most widely used method for peptide identification. Most of these algorithms provide one or more peptide spectrum match (PSM) scores that correlate with the quality of the match, but are typically hard to interpret and are not associated with any valid statistical meaning. Researchers face the problem of computing identification error rates or PSM significance measures and need to deal with post-processing software that converts search scores into meaningful statistical measures. This chapter focusses on scoring and assessment of database search results and gives a brief overview of common methods, their advantages and disadvantages, and presents alternative statistical concepts that deal with some of the shortcomings of standard methods in a non-mathematical language. We also present some examples using freely accessible software in order to demonstrate the ease with which sound statistics can be calculated today.

2. Scores and Thresholds

Sequest (4) was the first tandem MS search algorithm available and is today, together with Mascot (5), one of the most widely used database search tools. These are representative of the numerous database search algorithms that report for every PSM, a score that reflects the quality of the cross correlation between the experimental and the computed theoretical peptide spectrum. Although Sequest and Mascot scores are fundamentally different in their calculation, they facilitate good relative PSM ranking: all peptide candidates that were matched against an experimental spectrum are ranked according to the PSM score, and only the best matches are reported.

Often, only the top hit is considered for further investigation, and some search engines such as X!Tandem (6) exclusively report that very best match. However, not all these identifications are correct. Sorting all top hit PSMs (absolute ranking) according to their score enables the selective investigation of the very best matched PSMs. This approach was initially used to aid manual interpretation and validation. As the field of MS-based proteomics moved towards high-throughput methods, researchers started to define empirical score thresholds: PSMs scoring above these thresholds were accepted and assumed to be correct, while anything else was classified as incorrect. Depending on how well the underlying PSM score discriminates, the correct and incorrect scores overlap significantly (Fig. 1), and therefore thresholding is always a trade-off between sensitivity (fraction of true positive identifications) and the acceptable error rate (fraction of incorrect identifications). Low score thresholds will accept more PSMs at the cost of a higher error rate, and on the other hand, a high score threshold reduces the error rate at the cost of sensitivity. Many groups also apply heuristic rules that combine the score threshold with some other validation properties such as charge state, the difference in score to the second best hit, amongst others. The problem with these methods is that the actual error rate remains unknown, and the decision of accepting assignments is only based on judgement of an expert. Moreover, results between laboratories or even between experiments cannot be reliably compared since different search algorithms, protein databases, search parameters, instrumentation and sample complexity etc., require adaptation of acceptance criteria. A recent HUPO study (7) investigated the reproducibility between laboratories. Amongst the 18 laboratories, each had their own criteria of what was considered a high and low

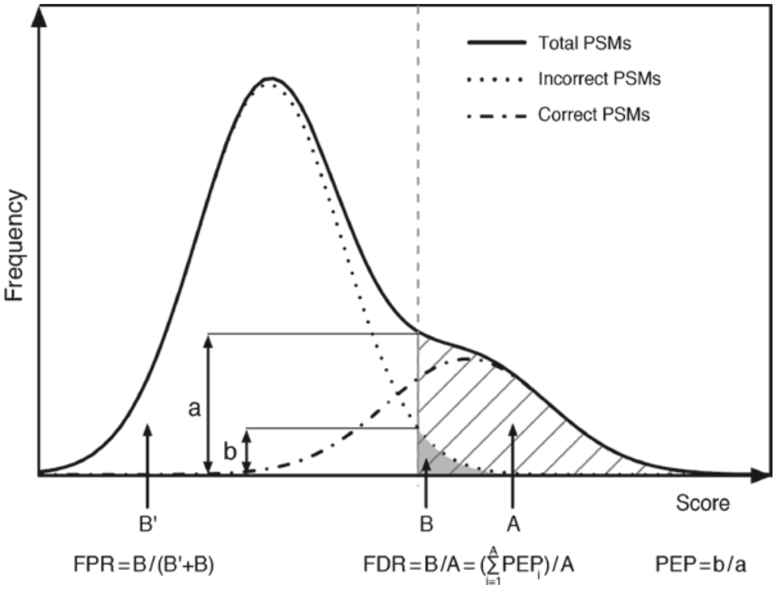


Fig. 1. A score distribution (*solid line*) typically consists of a mixture of two underlying distributions, one representing the correct PSMs (*dash-dot line*) and one the incorrect PSMs (*dotted line*). Above a chosen score threshold (*dashed grey line*), the *shaded area* (**a**) represents all PSMs that were accepted, while the *solid grey filled area* (**b**) represents the fraction of incorrectly identified PSMs with the chosen acceptance criteria. B together with B′ sum up all incorrect PSMs for the whole dataset. The false positive rate (FPR) and the false discovery rate (FDR) can be calculated when the numbers of PSMs in B, B′ and A are counted using the presented formulas. The posterior error probability (PEP) can be calculated from the height of the distributions at a given score threshold

confidence protein identification, which were mostly based on simple heuristic rules and score thresholds ((7), supplementary table 1). It was found that the number of high confidence assignments between two different laboratories could vary by as much as 50%, despite being based on the same data. As a result, many proteomic journals require the validation and assessment of score thresholds, ideally with significance measures such as presented below.

3. From PSM Scores to Meaningful Significance Measures

The expected error rates associated with individual or sets of PSMs can be reported as standard statistical significance measures. This allows transformation of specific scoring schemes into generic and unified measures, enabling comparability across any experiment in a consistent and easy to interpret format. In this section, we discuss and explain commonly used statistical measures that ideally are reported by every database search algorithm or post-processing software; focusing on (a) the false discovery rate (FDR), its derived *q*-value and (b) the Posterior Error Probability (PEP), also sometimes called local FDR. In the subsequent section, we focus on how these measures can be calculated with available software tools.

3.1. About p-Values and q-Values

The *p*-value is a widely used statistical measure for testing the significance of results in the scientific literature. The definition of the *p*-value in the context of MS database search scores is the probability of observing an incorrect PSM with a given score or higher by chance, hence a low *p*-value indicates that the probability is small of observing an incorrect PSM. The *p*-value can be derived from the false positive rate (FPR), which is calculated as the proportion of incorrect PSMs above a certain score threshold over all incorrect PSMs (Fig. 1). The simple calculation of the *p*-value however is misguiding when this calculation is performed for a large set of PSMs. In this case, we would expect to observe a certain proportion of small *p*-values simply by chance alone. An example: given 10,000 PSMs at a score threshold that is associated with a *p*-value of 0.05, we expect $0.05 \times 10{,}000 = 500$ incorrect PSMs simply by chance. This leads to the well known concept of multiple testing correction, which can be found in its simplest, but conservative form in the Bonferroni correction. Bonferroni suggested to correct the *p*-value by the number of tests performed, leading to a *p*-value of 5E-5 in our example above. However, we have only corrected for the number of spectra, but not for the number of candidate peptides the spectrum was compared against. A correction of taking into account both factors leads to extremely conservative score thresholds. However, an alternative well established method for multiple testing correction for large-scale data (e.g. genomics and proteomics) is to calculate the false discovery rate (FDR) (8).

The FDR is defined as the expected proportion of incorrect predictions amongst a selected set of predictions. Applied to MS, this corresponds to the fraction of incorrect PSMs within a selected set of PSMs above a given score threshold (Fig. 1). As an example, say 1,000 PSMs score above a prearranged score threshold, and 100 PSMs were found to be incorrect, the resulting FDR would be 10%. On the other hand, the FDR can be used to direct the trade-off between sensitivity and error rate, depending on the experimental prerequisites. If, for example, a 1% FDR were required, the score threshold could be adapted accordingly.

To uniquely map each score and PSM with its associated FDR, the notion of *q*-values should be used. This is because two or more different scores may lead to the same FDR indicating that the FDR is not a function of the underlying score. Storey and Tibshirani (9) have therefore proposed a new metric, the *q*-value, which was introduced into the field of MS proteomics by Käll et al. (10, 11). In simple terms, the *q*-value can be understood as the minimal FDR threshold at which a PSM is accepted, thereby transforming the FDR into a monotone function: increasing the score threshold will always lower the FDR and *vice versa*. This property enables the mapping of scores to specific *q*-values. In Fig. 2, the *q*-value is shown for a Mascot search on a high

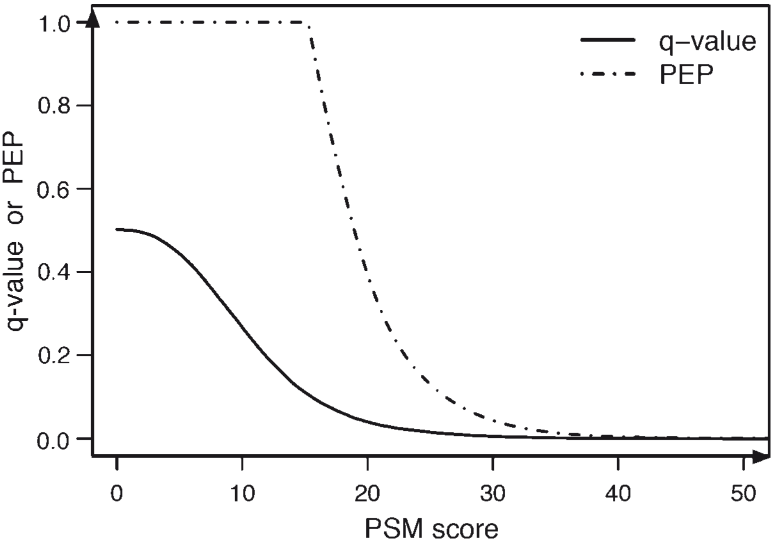


Fig. 2. PSM scores were transformed into *q*-values and Posterior Error Probabilities (PEP) using Qvality (*see* Subheading 4). A score cut-off of 30 demonstrates the fundamental difference of the two significance measures: the *q*-value would have reported about 0.5% of all the PSMs as incorrect above that score threshold, whereas the PEP would have reported 4% chance of a PSM being incorrect at this specific score threshold. *Note*. The maximum *q*-value for this dataset is 0.5 since only half of the PSMs are incorrectly assigned even without any score threshold applied because of the use of high quality and high mass accuracy data stemming from an LTQ-FT Ultra instrument

accuracy dataset. At a Mascot Ionscore of 10, 20 and 30, the corresponding *q*-values were 0.26, 0.04, 0.005 with 19,967, 14,608, 10,879 PSM identifications, respectively. It is important to note that for other datasets, instruments and parameter setting etc., the *q*-value could be significantly different for the same score and hence the *q*-value analysis should be performed for any individual search.

3.2. The Posterior Error Probability

The *q*-value is associated with individual PSM scores although this measure is always a result of all PSMs in a dataset. For illustration, imagine we remove from a large dataset half of the spectra that were incorrectly matched above a given score threshold; after spectral removal, the *q*-value for this same score threshold would be only about 50% of its original value even though the underlying spectrum and PSM remain the same. Moreover, in an extreme case, a *q*-value of 1% could be taken to mean that 99 PSMs are perfectly correct and 1 PSM is incorrect. More likely, the majority of these PSMs are good, but not perfect matches, and a few are weaker matches. Clearly, when the focus of an experiment is based on individual peptide identifications (e.g. for biomarker discovery, genome annotation, a key peptide for interesting and extensive follow-up research etc.), then it would be useful to compute spectrum specific significance measures that can be represented as the posterior error probability (PEP).

The global FDR or *q*-value reflects the error rate which is associated with a set of PSMs, whereas the PEP (or sometimes referred to as local FDR) measures the significance of a single spectrum assignment with a specific PSM score (11, 12). The PEP is simply the probability of the PSM being incorrect, thus a PEP of 0.01 means that there is 1% chance of that PSM being incorrect. For the example from above where 100 PSMs resulted in a *q*-value of 1%, the PEPs would have reflected the stronger and weaker matches.

Unlike the FDR and *q*-value calculations that require minimal distributional assumptions, the PEP can only be calculated with knowledge of the underlying score distributions representing the correct and incorrect PSM identifications (*see* Subheading 4) since the PEP is inferred from the height of the distributions at a given PSM score. Figure 1 illustrates again that the PEP is specific to one PSM score, whereas the FDR accounts for the whole set of PSMs that scored at least as good as the PSM at hand. This leads to the fact that the sum of the PEPs above a chosen score threshold divided by the number of selected PSMs results in an alternative way of computing the FDR (13).

Figure 2 shows the results of the PEPs as well as the *q*-values calculations for a high mass accuracy dataset that was searched with Mascot. For a PSM score threshold of 10, 20 and 30, the associated *q*-values were 0.26, 0.04 and 0.005, whereas the PEPs were 1.0, 0.39 and 0.04, respectively. This clearly demonstrates the difference between the significance measures: a Mascot score threshold of 30 (this is all PSMs with Mascot scores of 30 and above) led to only 0.5% incorrect PSMs in this dataset, whereas the individual Mascot score of 30 was associated with a 4% PEP.

4. Software and Methods to Compute Statistical Measures

Some database search algorithms report statistical measures, but these should be carefully validated and fully understood before being used and interpreted since their significance calculations are often based on pseudo statistical principles. It is however very easy to obtain well founded significance measures with free post-processing software packages and methods as presented below. We want to stress that their absolute accuracy still depends on the underlying models and assumptions. Finally, the well known effect of "garbage-in/garbage-out" is also true for MS data analysis, but when tools and methods are applied sensibly, they can be extremely valuable and represent some of the latest developments in shotgun proteomics.

4.1. Target/Decoy Database Searching for Validation

Moore et al. (14) pioneered the concept of target/decoy database searching (*see* Dedicated chapter in this book for a more complete review of this subject), where data is not only searched against the standard sequence database (target), but also against a randomised, shuffled or reversed database (decoy). There are two accepted concepts of target/decoy database searching and different groups favour one or the other method (15): either data is searched against a concatenated compound target/decoy database or alternatively data is separately searched against the target and decoy database.

The idea is that PSMs obtained from the decoy database can be used to estimate the number of incorrect target PSMs for any given criteria such as score thresholds or heuristic methods (16). This enables the calculation of the FDR by simply counting the number of decoy and target PSMs that meet the chosen acceptance criteria (Fig. 1, FDR formula for separate target/decoy searches). It should be noted that more accurate FDRs can be obtained when the fraction of incorrect PSMs (pi0) matching the target database can be estimated and incorporated. This is discussed in depth by Käll et al. in Refs. (10) and (12).

4.2. Qvality: Target/Decoy

Qvality (12), is a generic post-processing software tool that allows transformation of raw PSM scores into *q*-values and PEPs. It utilises separate target/decoy database searching together with nonparametric logistic regression, where decoy PSM scores are used as an estimate of the underlying null score distribution. Since no explicit assumptions of the type of the score distributions are made, the method was shown to be robust for many scoring systems and hence is not limited to one specific database search algorithm. Qvality is expected to calculate more accurate *q*-values than the standard approach discussed in Subheading 4.1, since it incorporates pi0 estimates into the FDR calculation.

Application of Qvality is straightforward; it only expects two disjoint sets of raw PSM scores as input, one stemming from the target and one from the decoy database. Data for Fig. 2 was for example computed with Qvality using the target and decoy Mascot Ion scores. Qvality is a small stand-alone command-line application without any external dependencies and is readily applicable. It can be downloaded under http://noble.gs.washington.edu/proj/qvality/.

4.3. PeptideProphet and Percolator

PeptideProphet and Percolator both report not only the FDR/*q*-value and PEP, but also attempt to improve the discrimination performance between correct and incorrect PSMs (Fig. 3) by employing an ensemble of features, several of which are used by experts for manually validating PSMs.

"PeptideProphet" (13), developed by Keller and Nesvizhskii et al., was the first software in the field of MS-based proteomics that reported probabilities (P) of the peptide assignment being correct, akin to the PEP, and FDRs. In order to improve the discrimination performance between correct and incorrect PSMs,

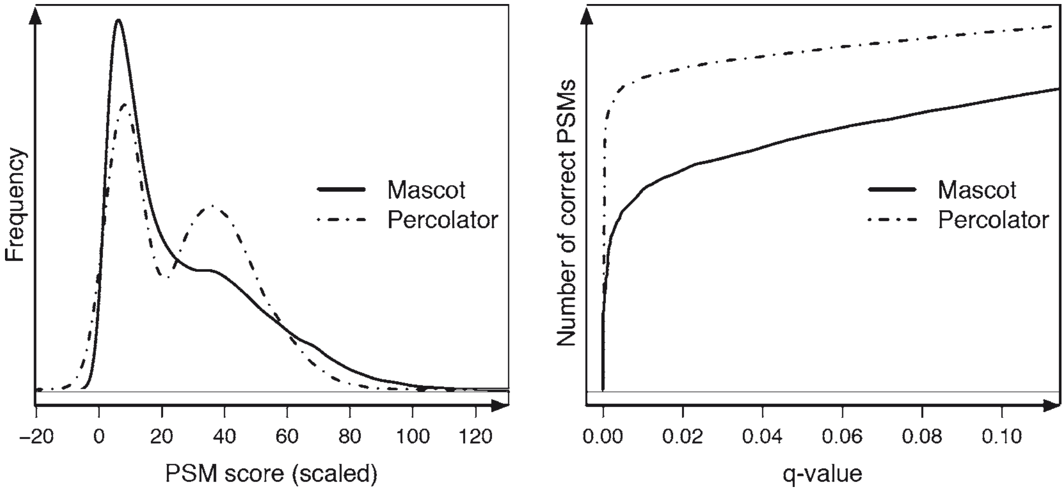


Fig. 3. Distributions of Mascot and Percolator scores were generated from a high accuracy LTQ-FT Ultra dataset (*left*). This illustrates the bi-modal nature of PSM matching scores as simulated in Fig. 1 and further demonstrates the discrimination performance improvement between correct and incorrect PSMs for post-processing tools such as Percolator over Mascot. *Note*. These scores are not on the same scale, but have been normalised and scaled for this illustration

PeptideProphet learns from a training dataset a discriminant score which is a function of Sequest specific scores such as XCorr, deltaCn, Sp amongst others. PeptideProphet makes extensive use of the fact that PSM scores as well as discriminant scores represent a mixture distribution from the underlying superimposed correct and incorrect score distributions (Figs. 1 and 3). The original PeptideProphet algorithm is based on the assumption that the type of these distributions remain the same across experiments and hence were determined from training datasets. However, using an Expectation Maximation algorithm, the parameters of these distributions are adapted for each dataset individually, enabling calculation of the corresponding FDR and P significance measures. Recent versions of PeptideProphet supplemented this parametric model with a variable component mixture model and a semi-parametric model (17, 18) that incorporate decoy database search results. The rational of this was to provide more robust models for a greater variety of analytical platforms where the type of distribution may vary. PeptideProphet is a widely used and accepted method to compute confidence measures and is available at http:// tools.proteomecenter.org. However, PeptideProphet is not a small stand-alone application, but is part of a large software package (Trans-Proteomic Pipeline) that comprises a collection of tools for tandem MS data analysis.

Percolator (19) is an alternative post-processing software that relies on target/decoy database search results rather than on distributional assumptions to infer the FDR/*q*-value and PEP. This system also improves the discrimination performance between correct and incorrect PSMs (Fig. 3) by employing a large ensemble of features including mass accuracy, enzyme specificity, score difference between top hit and second best hit, peptide and protein properties amongst many other features. Percolator calculates these features from the target and a separate decoy database search result to iteratively learn a classifier. The basic procedure is as follows: initially, the target and decoy PSMs are discriminated by the most relevant feature and filtered to a fixed FDR (e.g. 1%). This PSM subset (positive training set) together with the decoy PSMs (negative training set) is used to train a machine learning algorithm (Support Vector Machine). The learnt classifier is then applied to all target/decoy PSMs, followed by FDR filtering to continue the procedure as before. It was shown that after a few iterations, the system converges and results in a robust classifier that results in significantly better discrimination between correct and incorrect PSMs when compared with raw PSM scores (Fig. 3). Moreover, this system specifically and dynamically adapts for each dataset, which means that the used features and learnt classifiers are tuned to data quality, protocols and instrumentation.

Percolator is available under http://per-colator.com/ and similar to Qvality does not depend on any external dependencies and hence can be readily used. It offers a simple command line interface that requires Sequest results as input and outputs the *q*-value, PEP, as well as the peptide and associated protein(s) information for each spectrum. We have developed a Mascot module for Percolator that uses similar features but supplements these with Mascot specific features as well as intensity and ion-series information (20). It is available for download under http://www.sanger.ac.uk/Software/analysis/MascotPercolator/.

References

1. de Godoy, L.M., Olsen, J.V., de Souza, G.A., Li, G., Mortensen, P., and Mann, M. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. Genome Biol, 7(6), R50.
2. McCormack, A.L., Schieltz, D.M., Goode, B., Yang, S., Barnes, G., Drubin, D., and Yates, J.R. III. (1997) Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. Anal Chem, 69(4), 767–776.
3. Nesvizhskii, A.I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Meth, 4(10), 787–797.
4. 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(11), 976–989.
5. 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(18), 3551–3567.
6. Craig, R., and Beavis, R.C. (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 20(9), 1466–1467.
7. States, D.J., Omenn, G.S., Blackwell, T.W., Fermin, D., Eng, J., Speicher, D.W., and Hanash, S.M. (2006) Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. Nat Biotechnol, 24(3), 333–338.
8. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B, 57(1), 289–300.
9. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA, 100(16), 9440–9445.
10. Käll, L., Storey, J.D., MacCoss, M.J., and Noble, W.S. (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res, 7(1), 29–34.
11. Käll, L., Storey, J.D., MacCoss, M.J., and Noble, W.S. (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res, 7(1), 40–44.
12. Käll, L., Storey, J.D., and Noble, W.S. (2008) Nonparametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. Bioinformatics, 24(16), i42–i48.
13. 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(20), 5383–5392.
14. Moore, R.E., Young, M.K., and Lee, T.D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. J Am Soc Mass Spectrom, 13(4), 378–386.
15. Fitzgibbon, M., Li, Q., and McIntosh, M. (2007) Modes of inference for evaluating the confidence of peptide identifications. J. Proteome Res, 7(1), 35–39.
16. Brosch, M., Swamy, S., Hubbard, T., and Choudhary, J. (2008) Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. Mol Cell Proteomics, 7(5), 962–970.
17. Choi, H., and Nesvizhskii, A.I. (2008) Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res, 7(1), 254–265.
18. Choi, H., Ghosh, D., and Nesvizhskii, A.I. (2008) Statistical validation of peptide identifications in large-scale proteomics using the

target-decoy database search strategy and flexible mixture modeling. J Proteome Res, 7(1), 286–292.

19. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods, 4(11), 923–925.

20. Brosch, M., Yu, L., Hubbard, T., and Choudhary, J. (2009) Accurate and sensitive peptide identification with Mascot Percolator. J Proteome Res, 8(6), 3176–3181.

Chapter 5

Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics

Joshua E. Elias and Steven P. Gygi

Abstract

Accurate and precise methods for estimating incorrect peptide and protein identifications are crucial for effective large-scale proteome analyses by tandem mass spectrometry. The target-decoy search strategy has emerged as a simple, effective tool for generating such estimations. This strategy is based on the premise that obvious, necessarily incorrect "decoy" sequences added to the search space will correspond with incorrect search results that might otherwise be deemed to be correct. With this knowledge, it is possible not only to estimate how many incorrect results are in a final data set but also to use decoy hits to guide the design of filtering criteria that sensitively partition a data set into correct and incorrect identifications.

Key words: Proteomics, Target-decoy, False positive, False discovery, Mass spectrometry, Estimation

1. Introduction

Peptide and protein identifications made in most mass spectrometry-based proteomic work flows first involve acquiring a set of tandem mass (MS/MS) spectra and then interrogating each spectrum against spectra predicted from a list of protein sequences by search engines, such as SEQUEST (1), Mascot (2), OMSSA (3), and X!Tandem (4). The output of these programs indicates the best theoretical peptide matches to the input spectra, which are then used to infer the source protein that was present in the biological sample. Unfiltered sets of peptide identifications produced in this manner are necessarily imperfect for three reasons: (1) not all peptide species in a sample are represented in the search space; (2) spectra derived from background nonpeptide species will often be given a peptide assignment; and (3) incorrect candidate peptide sequences occasionally may outscore correct sequences.

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_5, © Humana Press, a part of Springer Science+Business Media, LLC 2010

For many search engines, nearly all input MS/MS spectra will be assigned a peptide match if there are any that lie within the supplied mass tolerance. Thus, the primary task of proteomics researchers is to distinguish incorrect from correct peptide assignments.

When working with very small data sets, such as those produced from a single spot on a 2D gel or a gel band representing a component of an isolated protein complex, identifying correct peptide identifications is almost trivial: they are the ones with the highest scores and tend to map to the same protein. It is also reasonable and appropriate to manually examine individual peptide-spectrum matches (PSMs) to verify that they are correct. However, the increasingly large data sets created by modern tandem mass spectrometers in global proteomic efforts are not amenable to these strategies. Simple filtering criteria based on score magnitude or numbers of peptides per protein tend to be neither sensitive nor accurate (5), and the staggering amount of information that can be produced in a single experiment renders the manual validation of peptide assignments impractical. Consequently, high-throughput protein sequencing efforts must rely on methods for estimating the frequencies of incorrect peptide and protein identifications among correct ones. The "target-decoy" search strategy is a simple yet powerful way to deliver false positive estimations and can be applied to nearly any MS/MS workflow. Here, we present several methods for preparing decoy sequences and strategies for selecting correct peptide identifications.

2. Materials

2.1. MS/MS Spectra

MS/MS spectra can be acquired on any number of tandem mass spectrometers, including the LTQ family of ESI-ion trap instruments from ThermoFisher, the QSTAR from Applied Biosystems, and the FLEX family from Bruker Daltonics. Alternatively, several public sources of MS/MS spectra are freely available on the internet, including PeptideAtlas (6) and the Open Proteomics Database (7). It is recommended that the target-decoy approach be applied to data sets consisting of several thousand MS/MS spectra (see Note 1).

2.2. Protein Sequences

MS/MS spectra are generally searched against peptides predicted from FASTA-formatted protein sequence lists. Sequence lists should be chosen such that any peptide that may have given rise to an observed spectrum is represented. For example, if a mouse-derived sample was sequenced by MS/MS, the spectra should be searched against a list of all known mouse proteins. Protein lists can be downloaded from numerous sources, including the International Protein Index (8) and UniProt/SwissProt (9). It is useful to also include sequences of known contaminants, such as trypsin and human keratins.

2.3. Search Engine

Numerous MS/MS search engines are in common usage. Some are commercially available, for example:

1. SEQUEST (http://www.thermo.com/com/cda/product/detail/0,1055,22209,00.html)
2. Mascot (http://www.matrixscience.com)
3. SpectrumMil (http://www.chem.agilent.com/scripts/pds.asp?lpage=7771)

   Other search engines are freely-distributed via the internet:
4. OMSSA (http://pubchem.ncbi.nlm.nih.gov/omssa/)
5. X!Tandem (http://www.thegpm.org/tandem/)

All of these produce some form of a score indicating the degree to which observed and predicted MS/MS spectra agree. Several of these search engines' scores may be probability-based. See refs. (10–13) for more detailed descriptions and comparisons of these search engines. One principle benefit to target-decoy searching is its applicability to data generated by any search engine.

3. Methods

One deceptively simple way to estimate false positives is to manufacture "decoy" sequences that do not exist in nature, and then allow the search engine to consider these alongside "target" sequences derived from the organism being studied. Necessarily, incorrect decoy hits should be similar to incorrect but unknown hits derived from target sequences in terms of length, amino acid composition, mass accuracy, and search engine-assigned scores. Therefore, knowing the proportion of decoy versus target sequences in the search space allows one to estimate the number of incorrect target sequences in a reasonably large collection of PSMs. More than providing a means to estimate the number of incorrect target hits in a collection of PSMs, decoy hits can be used to guide researchers in the design of sensitive filtering criteria to precisely distinguish correct from incorrect PSMs.

Target-decoy searching is usually performed in the following steps:

1. Construct a concatenated target-decoy sequence list, marking decoy sequences with a text flag in their annotation.
2. Use a MS/MS search engine to interpret input MS/MS spectra using target-decoy sequence list.

3. Evaluate the relative proportion of target and decoy sequences in the search space to derive the multiplicative factor required to estimate false positives, if necessary.
4. Estimate false positive-related statistics.
5. Use decoy hits to guide the establishment of filtering criteria.
6. Report statistics for filtered data set.

Each of these steps will be discussed in further details below.

3.1. Decoy Sequence Construction

Several methods for creating decoy sequences have been described (14–16). Each has varying advantages and disadvantages, and it must be stressed that no single decoy type is perfect. Ideal decoy sequences should have the following characteristics:

1. Similar amino acid distributions as target protein sequences.
2. Similar protein length distribution as target protein sequence list.
3. Similar numbers of proteins as target protein list.
4. Similar numbers of predicted peptides as target protein list.
5. No predicted peptides in common between target and decoy sequence lists.

If each of these conditions are reasonably met, one can safely assume that decoy sequence selected by the search engine are incorrect, and that there is a one-to-one correspondence between incorrect target hits and decoy hits. By design or as a consequence of the decoy sequence construction method, conditions 3 or 4 may not be met. In this case, one should take into account the discrepancy between target and decoy sequences (*see* Subheading 3.3). This is particularly true when using stochastic means to generate decoy sequences based on target sequences demonstrating substantial amounts or repetition or homology.

3.1.1. Reversed Proteins

Protein reversal is by far the simplest and most widely used method for creating decoy sequences (see Note 2 for a simple Perl script to create a concatenated target-decoy sequence list based on an input target sequence list) (17, 18). By switching the amino-carboxyl orientation of a protein's amino acids, a negligible number of peptide sequences are preserved, particularly when imposing *in silico* digestion constraints with proteases like trypsin. Protein reversal has two main advantages: First, because it preserves the general features of the target sequence list, reversed protein sequences will share the same degree of interprotein redundancy as the input target sequences; Second, since it is a defined transformation, multiple research groups can generate the same decoy sequences. The main disadvantage to protein reversal is that it is not a random transformation as some may prefer. Consequently, it can be argued that it does not strictly represent a null random distribution, and for certain types of peptides (e.g., palendromic or low sequence complexity), it may not be possible to create a suitable decoy counterpart. In practice, however, protein reversal stands up to the five conditions listed above (14), and can therefore be used to faithfully estimate the occurrences of incorrect identifications.

3.1.2. Shuffled Proteins

Protein shuffling is another method used for creating decoy sequences (16) in which the amino acids of each input target protein are randomly rearranged to yield a new decoy protein. Like protein reversal, shuffling is fairly simple to implement programmatically, and it preserves both the amino acid composition and length of each input target protein sequence. Unlike sequence reversal, this transformation has desired stochastic properties. As is true of most random transformations though, redundancies and homologies between protein entries will not be preserved, resulting in a greater number of decoy peptides than originally present in the target sequence list. This imbalance must be measured and then taken into account when generating estimations of false positives.

3.1.3. Random Proteins

Proteins can also be generated in a completely random fashion. This is the method internally implemented by some search engines, such as Mascot, for performing target-decoy analyses. Ideally, randomized sequences should have the same amino acid biases and protein length distribution as an input target sequence list. One way to do this is to first evaluate the target sequence list to generate a frequency matrix of amino acids and a histogram of protein lengths. Decoy proteins are then constructed by randomly selecting amino acids according to the frequency matrix, and adding these to the growing decoy protein until it reaches a specified length, randomly determined from the length histogram.

Rather than relying on a simple amino acid frequency matrix, one can construct a Markov chain model of amino acid frequencies to better replicate small scale patterns found in the target sequence list, such as single or double amino acid repeats or highly basic or acidic regions. Essentially, this is done by generating a frequency matrix reflecting the likelihood of observing a particular amino acid given the preceding n amino acids (14). Another frequency matrix should be constructed consisting of only the n amino acids that initiate the protein sequence. After randomly selecting from the initiating sequence frequency matrix, the protein can be extended by randomly selecting from the conditional frequency matrix until the protein achieves a specified length.

With either randomization method, it is possible to modulate the number of decoy sequences with respect to the number of target sequences considered. This has been done to examine the effects of interrogating a set of MS/MS spectra against search spaces of varying sizes (19). As with shuffled decoy proteins, random proteins do not preserve redundancies and homologies, so care must be taken to measure the relative proportion of target and decoy sequences, and then account for any observed bias when generating false positive estimations (see Subheading 3.3).

3.1.4. Decoy Peptides

Rather than generating entire decoy proteins from which decoy peptides will be derived according to *in silico* enzymatic digestion rules, one can instead generate decoy peptides directly by altering each peptide sequence derived from the target sequence list. Alterations can take the form of reversals or shuffling. This procedure has the advantage of creating decoy peptides exactly matching the masses of all target peptides considered by the search engine. If reversal or nonrandom shuffling was the transformation applied, the number of target and decoy sequences will match exactly both in number and in mass distributions. Otherwise, decoy peptides may outnumber target peptides, as with stochastically created proteins. Since *in silico* digestion is usually performed by the search algorithm prior to querying observed spectra, the generation of decoy peptides directly is typically performed within the search algorithm. An example of a search engine with this feature is the Sorcerer-SEQUEST platform from SAGE-N.

3.2. Spectrum Search

Once a target-decoy sequence list has been generated, the analysis of a set of MS/MS spectra can begin. The generally accepted means to do this is to supply the search engine with a single protein sequence list consisting of both target and decoy sequences. For each spectrum, the search engine must then choose between target and decoy sequences. Correctly-identified peptides will exclusively be selected from target protein sequences, while incorrect peptide matches will be randomly drawn from target and decoy sequences. If the number of target and decoy sequences considered by the search engine are equal, there should be a one-to-one correlation between target and decoy sequences among incorrect identifications. If the number of target and decoy sequences are unequal, the correlation between target and decoy sequences should reflect this bias. It should be noted that some groups advocate searching target sequences separately from decoy sequences. For a variety of reasons, this procedure can lead to an overly conservative interpretation of search results (14) (see Note 3).

3.3. Measuring Decoy Bias

In order to properly estimate the number of false positive identifications in a set of peptide identifications, it is essential that one first knows the relative proportion of decoy to target hits in the search space. For reversed-decoy databases, it can generally be assumed that there is a 1:1 correlation between target and decoy sequences (14). For decoy sequence lists generated with a stochastic component, there are usually more decoy sequences than target sequences, particularly when there is a substantial degree of homology or redundancy among target sequences. One computational approach for measuring this proportion is to create *in silico* digests of each target and decoy component, and then ask how many peptides from each component are within a specified tolerance near a given mass. For example, one would determine how many target and decoy peptides are within 1.0 Da surrounding a mass of 1,000 Da. The proportion of target and decoy peptides should be consistent across all masses in the range of peptides one might consider (e.g., 600–5,000 Da).

More simply, one can examine the frequency with which a search engine returns target and decoy hits for incorrect identifications. Since correct peptide identifications usually achieve the top-ranked hit for a given MS/MS spectrum, it can be usually assumed that lower ranked peptide hits are incorrect (14, 20, 21). Alternatively, if one shifts the precursor masses of input MS/MS spectra outside of the specified mass tolerance, they cannot be correctly matched (14, 20). Comparing the frequencies of target and decoy hits for incorrect spectra reveals the effective proportion of target and decoy sequences in the search space and therefore the factor one should use to estimate the number of hidden incorrect target hits, given the observed decoy hits (Fig. 1) (14).

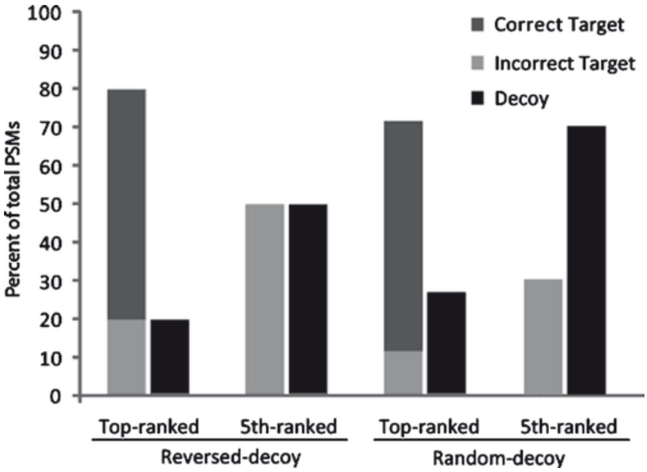


Fig. 1. Decoy PSMs indicate incorrect target PSMs, depending on the underlying proportion of target and decoy sequences. Under the reversed-decoy model, the proportion of target and decoy peptides considered are approximately equal (5th-ranked, reversed-decoy). Thus, the proportion of decoy PSMs observed in the presence of correct identifications equals the proportion of target PSMs that are incorrect (Top-ranked, reversed-decoy). When the underlying proportion of target and decoy sequences are not equal, as is usually the case with randomly created protein sequence lists, one must first measure this proportion (5th-ranked, random-decoy), and then apply it to the condition containing correct identifications (top-ranked, random-decoy). See ref. 14 for further details

Once the background frequencies of target and decoy hits are determined (t, d), one can determine the multiplicative factor (f) used to estimate the total (target + decoy) number of incorrect identifications:

$$f = \frac{1}{d} \tag{1}$$

where $d = 1 - t$. For reversed decoy sequences in which target and decoy search spaces are nearly equal, it can be assumed that t and d are both equal to 0.5, and f is therefore equal to 2. One can then estimate the total number of incorrect peptides by doubling the number of observed decoy hits. If t and d are determined to be 0.37 and 0.63, respectively, as can be the case for randomly-created decoy sequences (14), then f should be 1.6.

3.4. False Positive Statistics

In order to fairly compare data sets collected in different laboratories, acquired on different instruments, searched with different search engines, and representing different biological samples, it is crucial that they meet similar false positive-related constraints. The first step in this process is to estimate the total number of correct PSMs in the entire data set. One way to do this is as follows:

1. Sort all peptide hits by score, descending.
2. Count how many target hits are greater than or equal to a given score
3. Count how many decoy hits are greater than or equal to a given score
4. Estimate the number of correct hits (true positive, TP) from total (T) and decoy hits (d) greater than or equal to a given score:

$$\mathrm{TP} = T - df \tag{2}$$

5. Estimate the total number of correct hits in the data set from the maximum value of TP observed across all score thresholds.

Given the total number of correct identifications in the data set, the number of identifications being considered, and how many of these are incorrect, one can populate the Venn diagram shown in Fig. 2. Given estimations of false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN), one can generate the measurements shown in Table 1. Of these, precision and sensitivity are often the most useful for evaluating and comparing MS/MS data sets.

3.5. Designing Filtering Criteria

For several years, large MS/MS data sets were subject to predefined, general filtering constraints to attempt to separate correct from incorrect peptide identifications. Often, these constraints were learned from a training data set consisting of known proteins,

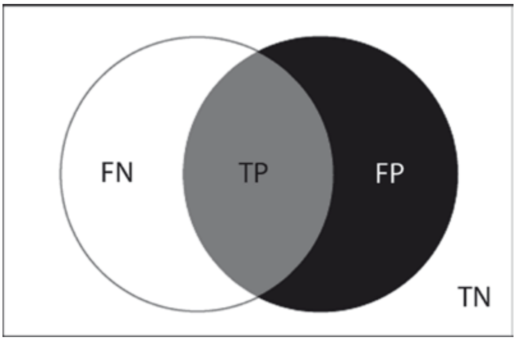


Fig. 2. Venn diagram of basic measurements related to estimated false positive identifications. The total number of identifications are contained within the rectangle. All correct identifications are contained within the *white circle*. All identifications passing a given set of selection criteria (positive identifications) are contained within the *black circle*. The overlap between these circles are true positives (TP). False positive identifications (FP) are the remaining positive identifications, and false negative identifications (FN) are the remaining correct identifications that do not meet the selection criteria. True negatives (TN) are the incorrect identifications that are correctly classified as such by the selection criteria. This Venn diagram scheme is elaborated in Fig. 3

Table 1
Measurements derived from target-decoy estimations of FP, TP, FN, TN

Measurement	Formula	Description
Precision	$\frac{\text{TP}}{\text{TP + FP}}$	Proportion of assignments passing selection criteria that are correct
False discovery rate (FDR)	$\frac{\text{FP}}{\text{TP + FP}}$, 1 – precision	Proportion of assignments passing selection criteria that are incorrect
Sensitivity	$\frac{\text{TP}}{\text{TP + FN}}$	Proportion of correct assignments passing selection criteria
Specificity	$\frac{\text{TP}}{\text{TP + FP}}$	Proportion of all incorrect assignments excluded by selection criteria
Accuracy	$\frac{\text{TP+TN}}{\text{TP + FP+TN+FP}}$	Proportion of all assignments correctly classified by selection criteria

and then applied to experimental data sets that were often orders of magnitude larger than the training data set. Through target-decoy searching, it was determined that the proportion of false positive identifications that surpass standard criteria varies with individual data sets, as does the proportion of correct identifications that fail to meet them (i.e., false negatives). Thus, application of identical filtering criteria across multiple data sets does not necessarily yield data sets with comparable sensitivity or precision rates. It is often desirable, therefore, to design filtering criteria that can accommodate the diversity of LC-MS/MS analyses while yielding optimized, comparable error profiles.

Since decoy peptide matches and incorrect target matches have similar properties, one can examine decoy hits to learn how all incorrect hits can be segregated from correct hits in a sensitive and precise manner. This is fairly easy to accomplish when one considers a single monotonic score provided by the search engines, such as SEQUEST's XCorr, Mascot's Ion Score, and the E-value from OMSSA and X!Tandem, or composite scores, such as the Discriminant Score, returned by Peptide Prophet's linear discriminant function (5):

1. Sort all peptide hits by score, descending.
2. Count how many target hits are greater than or equal to a given score
3. Count how many decoy hits are greater than or equal to a given score
4. Estimate the total number of incorrect hits (false positive, FP) from observed decoy hits (d) greater than or equal to a given score:

$$\mathrm{FP} = df \tag{3}$$

5. Calculate statistics related to FP for each given score threshold (see Subheading 3.5).
6. Select score threshold based on a desired statistic threshold.

Single scores are generally less able to sensitively separate correct from incorrect hits than consideration of multiple peptide measurements, such as mass accuracy, enzyme specificity, and alternate scoring methods. Composite scores are therefore superior to single scores, since they can incorporate these multiple lines of evidence that influence the likelihood that a peptide is correct. Another approach is to use the target-decoy strategy to examine multiple peptide measurements in a holistic fashion without condensing them into a single composite score. This is done by seeking an optimal (or several optimal) threshold combination(s) that maximizes the number of peptide identifications while minimizing the number of false positive identifications, or at least restricting them to a specified proportion of all positive identifications (Fig. 3). Evaluating and optimizing multiple candidate score threshold combinations can be tedious to perform manually; computational approaches for doing this have been described, however (22, 23).

3.6. Report Statistics for Filtered Data Set

Increasingly, journals are requiring an assessment of data quality when publishing MS/MS results (24–26). As previously stated,

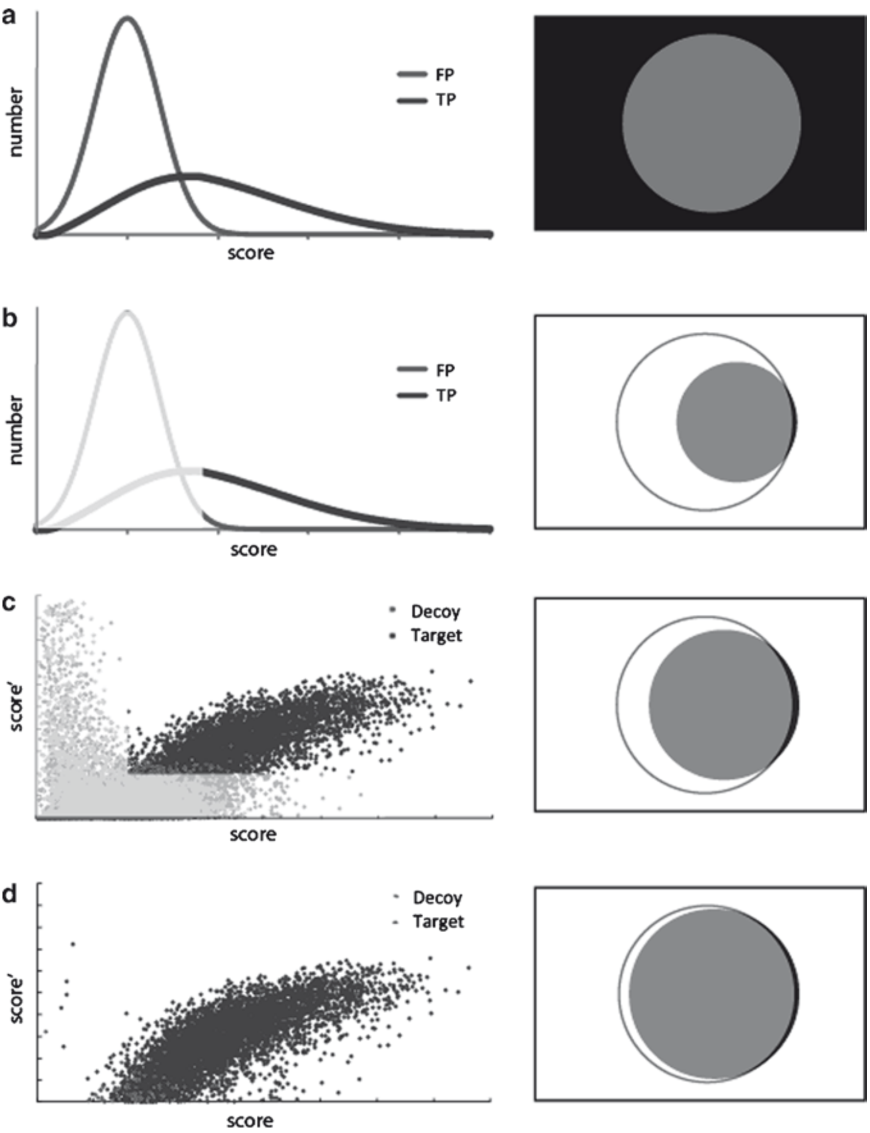


Fig. 3. Considering multiple selection criteria enhances accuracy. Selection criteria applied to score distributions (*left*) determine the form of the Venn diagrams (*right*). Venn diagram shapes and colors correspond with those in Fig. 2. (**a**) Distribution of FP and TP hits sorted by an arbitrary score. When no score criteria are applied, all selected correct identifications are denoted in *grey circle*, and all selected incorrect identifications are denoted in *black rectangle*. (**b**) Application of a single score threshold, which excludes most incorrect identifications (*lighter region*), can yield an acceptable precision rate, but yields sub-optimal sensitivity. (**c**) Considering two scores allows for greater separation between correct and incorrect identifications. The distribution of incorrect identifications is indicated by the distribution of decoy hits. Application of global criteria that excludes most decoy hits in two score dimensions (*lighter region*) provides greater sensitivity than one score alone. (**d**) Designing selection criteria that take into account numerous peptide measurements, such as mass accuracy, charge, enzymatic specificity, and peptides per protein, can yield far greater sensitivity while maintaining acceptable precision

the most useful measurements are usually precision (or FDR) and sensitivity. Although it is convenient to include decoy hits in a data set during analysis (see Note 4), decoy hits should not contribute to the final tally of incorrect hits since they can be easily recognized and removed. Thus, the reported number of FP and corresponding precision rate should be:

$$\mathrm{FP}_{\mathrm{final}} = d\,(f - 1) \tag{4}$$

$$\mathrm{precision}_{\mathrm{final}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

It must be stressed that the above calculations apply to the aggregate of all identifications that meet or exceed a given set of filtering criteria. The final precision rate represents the proportion of the final data set that is likely to be correct; it does not indicate the likelihood of any particular identification of being correct (see Note 5).

These statistics may also be applied at the protein level. However, protein inference from multiple peptides poses additional challenges beyond the scope of this chapter (see Note 6). Protein precision is often worse than the precision measured from PSMs. This usually can be attributed to proteins that are incorrectly identified by just one peptide. In contrast, proteins identified by multiple peptides are usually correct. Thus, correct peptide identifications map to fewer proteins than incorrect peptides, reducing the final protein precision. This situation can be addressed by paying specific attention to single peptide identifications (see Note 7).

4. Notes

1. It is important to emphasize that the target-decoy search strategy is a tool for *estimating* the number of incorrect target PSMs. It is therefore useful to place confidence intervals on these estimations. If one assumes that target and decoy hits follow a binomial distribution (27), the theoretical standard deviation σ of target-decoy estimations can be calculated explicitly, given estimated precision and the observed number of PSMs being considered (*N*):

$$\sigma = \sqrt{\frac{1 - \mathrm{precision}}{N}} \tag{6}$$

Given σ, precision and *N*, one can estimate the confidence interval *C* containing a given proportion of repeated measurements of the precision, assuming a two-tailed normal distribution:

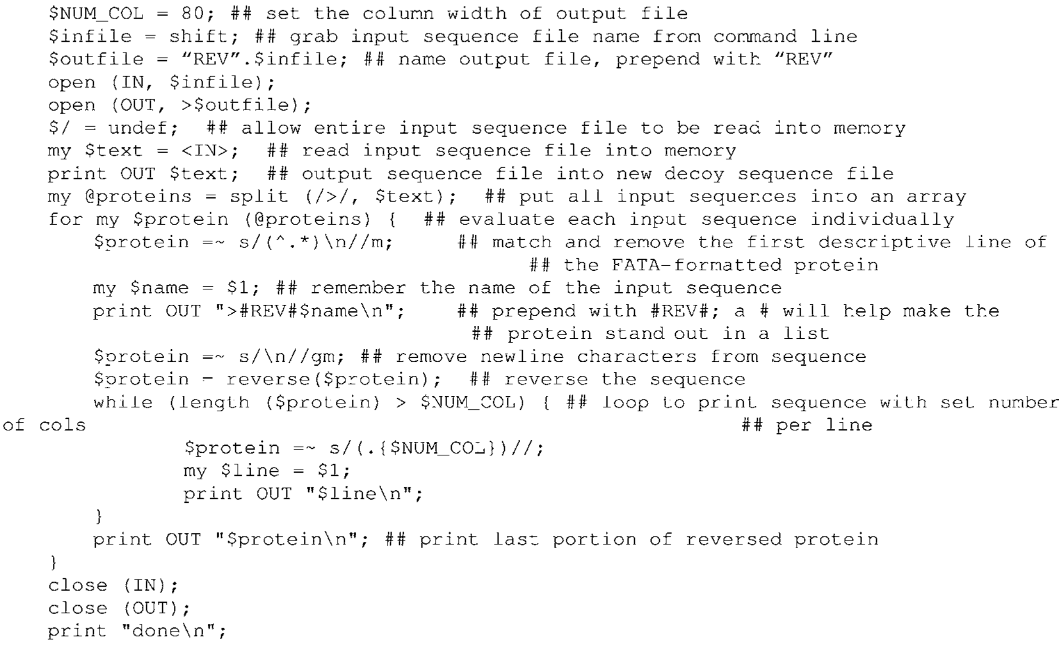
$$C = (1 - \text{precision}) \pm \frac{Z\sigma}{\sqrt{N}} \quad (7)$$

Combining Eqs. 5.6 and 5.7 gives

$$C = (1 - \text{precision}) \pm \frac{Z\sqrt{1 - \text{precision}}}{N} \quad (8)$$

For example, a confidence level of 0.99 indicates a Z value of 2.58; given an observed precision level of 0.9500, from 2000 PSMs, one would calculate the confidence interval to be ±0.000288. However, for 200 PSMs, this interval would be wider at ±0.00288. If the precision rate were decreased to 0.8000 from 2000 PSMs, this interval would also be larger at ±0.000576. Thus, these equations indicate that estimation confidence increases with larger sample sizes and fewer incorrect spectra in the underlying data. Considering more extreme values, the target-decoy approach is usually not very effective on small (tens) sets of PSMs or sets of PSMs that are largely incorrect (14).

2. A simple Perl script for generating a target-reversed decoy sequence list:



```
    $NUM_COL = 80; ## set the column width of output file
    $infile = shift; ## grab input sequence file name from command line
    $outfile = “REV”.$infile; ## name output file, prepend with “REV”
    open (IN, $infile);
    open (OUT, >$outfile);
    $/ = undef;  ## allow entire input sequence file to be read into memory
    my $text = <IN>;  ## read input sequence file into memory
    print OUT $text;  ## output sequence file into new decoy sequence file
    my @proteins = split (/>/, $text);  ## put all input sequences into an array
    for my $protein (@proteins) {  ## evaluate each input sequence individually
        $protein =~ s/(^.*)\n//m;      ## match and remove the first descriptive line of
                                             ## the FATA-formatted protein
        my $name = $1; ## remember the name of the input sequence
        print OUT ">#REV#$name\n";     ## prepend with #REV#; a # will help make the
                                        ## protein stand out in a list
        $protein =~ s/\n//gm; ## remove newline characters from sequence
        $protein - reverse($protein);  ## reverse the sequence
        while (length ($protein) > $NUM_COL) { ## loop to print sequence with set number
of cols                                                              ## per line
                $protein =~ s/(.{$NUM_COL})//;
                my $line = $1;
                print OUT "$line\n";
        }
        print OUT "$protein\n"; ## print last portion of reversed protein
    }
    close (IN);
    close (OUT);
    print "done\n";
```

3. Several groups recommend first searching MS/MS spectra against decoy sequences to derive a null distribution of scores, and then basing filtering criteria on the null distribution. Furthermore, by restricting the target database search to just target sequences, scores that are dependent on the search space will often be greater for correct identifications in comparison to the combined target-decoy search. While the practice of separate searches is reasonable in principle, it creates a variety of situations that must be accounted for in the final analysis. These include, but are not limited to:

   (a) *Correct/incorrect PSM noncompetition*: A high-quality MS/MS spectrum will often receive an elevated score compared to a low-quality spectrum, even if both corresponding PSMs are incorrect. When searching against a concatenated target-decoy sequence list, a correct target PSM necessarily competes with an incorrect decoy PSM, and is then returned by the algorithm. Under the separate searches paradigm, high-scoring decoy PSMs will indicate setting an exceptionally stringent filtering threshold that undermines sensitivity, unless these PSMs are secondarily compared to their target PSM counterparts following the search.

   (b) *Imbalanced incorrect target and decoy numbers*: Typical search results consist of a mixture of correct and incorrect PSMs. Under the concatenated target-decoy paradigm, incorrect PSMs are distributed between target and decoy sequences according to their background frequency (i.e., 1:1 for reversed sequences). When searching target and decoy sequences separately, decoy PSMs will necessarily outnumber incorrect PSMs, since spectra that can be correctly assigned to target sequences will be matched to decoy sequences. For example, if 20% of all spectra are correctly assigned, the proportion of incorrect target to incorrect decoy will be 0.8:1, even if the underlying target and decoy sequences were equal in number. Further complicating matters, the larger decoy distribution presents the opportunity for them to achieve a wider range of scores, inappropriately suggesting more stringent filtering thresholds.

4. Even after a set of filtering criteria have been arrived at, it is often useful to leave decoy PSMs mixed among the target ones. Should one choose to revisit the data analysis, one can derive further filtering/selection criteria involving additional parameters not considered in the original analysis.

5. False positive statistics applied to entire data sets can obscure scoring data, which indicate that some PSMs are assigned with

greater confidence than others. Particularly with very large data sets with a set precision threshold, it is possible that a small number of PSMs with a very low likelihood of being correct will be included. Recently, it has been proposed to restrict PSM selection based on the likelihood that a particular identification is correct (16, 23, 28, 29). This can be a highly useful practice, particularly when there is little tolerance for error, such as the submission of PSMs to a reference data set. However, many research applications are tolerant of some error, since it can allow for much greater sensitivity. A data set composed of PSMs with a minimum likelihood of being correct of 0.99, for example, may have an overall precision rate of 0.999, but nearly half the sensitivity of a data set restricted to have a precision of 0.99.

6. Although peptide identifications can be correct, it is possible to incorrectly infer the proteins that gave rise to them, due to sequence homologies. These proteins should be considered to be false positives, since the identified proteins were not actually present in the experimental sample. The target-decoy system cannot be used to estimate this source of error. Programs, such as Protein Prophet (30), can be used to formally identify the protein(s) that are most likely given the observed peptides. However, it is worth noting that despite some protein ambiguity, often, peptides restrict the protein identifications to a narrow group, often consisting of highly related isoforms.

7. Proteins identified by single peptides (“one-hit-wonders”) represent a special class of peptide and protein identifications. It is generally true that the vast majority of incorrect peptide identifications are in this category. As a result, the precision rate measured at the protein level is usually less than that observed at the protein level. It is often tempting, therefore to remove single peptide identifications from a final data set. While this practice certainly improves the precision rate at the protein level, it is usually accompanied by a substantial loss in sensitivity. Often, more than half of all correct peptide identifications fall into the one-hit-wonder category. Rather than removing these PSMs from the final data set, a more measured approach would be to apply filtering criteria tailored to just this subset.

Acknowledgments

This work was supported in part by National Institutes of Health (NIH) GM67945 and HG00041 (S.P.G.).

References

1. Eng, J. K., McCormack, A. L., and Yates, J. R. I. (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, 976–89.
2. 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, 3551–67.
3. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. J Proteome Res 3, 958–64.
4. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–7.
5. 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, 5383–92.
6. Deutsch, E. W., Lam, H., and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep 9, 429–34.
7. 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, 471–2.
8. 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, 1985–8.
9. (2008) The universal protein resource (UniProt). Nucleic Acids Res 36, D190–5.
10. Bakalarski, C. E., Haas, W., Dephoure, N. E., and Gygi, S. P. (2007) The effects of mass accuracy, data acquisition speed, and search algorithm choice on peptide identification rates in phosphoproteomics. Anal Bioanal Chem 389, 1409–19.
11. Balgley, B. M., Laudeman, T., Yang, L., Song, T., and Lee, C. S. (2007) Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. Mol Cell Proteomics 6, 1599–608.
12. Elias, J. E., Haas, W., Faherty, B. K., and Gygi, S. P. (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods 2, 667–75.
13. Sadygov, R. G., Cociorva, D., and Yates, J. R., III (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat Methods 1, 195–202.
14. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4, 207–14.
15. Higdon, R., Hogan, J. M., Van Belle, G., and Kolker, E. (2005) Randomized sequence databases for tandem mass spectrometry peptide and protein identification. OMICS 9, 364–79.
16. Kall, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 7, 29–34.
17. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. J Am Soc Mass Spectrom 13, 378–86.
18. 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.
19. Haas, W., Faherty, B. K., Gerber, S. A., Elias, J. E., Beausoleil, S. A., Bakalarski, C. E., Li, X., Villen, J., and Gygi, S. P. (2006) Optimization and use of peptide mass measurement accuracy in shotgun proteomics. Mol Cell Proteomics 5, 1326–37.
20. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 24, 1285–92.
21. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., and Gygi, S. P. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol 22, 214–19.
22. Jiang, X., Han, G., Ye, M., and Zou, H. (2007) Optimization of filtering criterion for SEQUEST database searching to improve proteome coverage in shotgun proteomics. BMC Bioinformatics 8, 323.
23. Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 4, 923–5.
24. Binz, P. A., Barkovich, R., Beavis, R. C., Creasy, D., Horn, D. M., Julian, R. K., Jr., Seymour, S. L., Taylor, C. F., and Vandenbrouck, Y. (2008) Guidelines for reporting the use of mass

spectrometry informatics in proteomics. Nat Biotechnol 26, 862.

25. Bradshaw, R. A., Burlingame, A. L., Carr, S., and Aebersold, R. (2006) Reporting protein identification data: the next generation of guidelines. Mol Cell Proteomics 5, 787–8.

26. Taylor, C. F. (2006) Minimum reporting requirements for proteomics: a MIAPE primer. Proteomics 6 Suppl 2, 39–44.

27. Huttlin, E. L., Hegeman, A. D., Harms, A. C., and Sussman, M. R. (2007) Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. J Proteome Res 6, 392–98.

28. Kall, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res 7, 40–4.

29. Tang, W. H., Shilov, I. V., and Seymour, S. L. (2008) Nonlinear fitting method for determining local false discovery rates from decoy database searches. J Proteome Res 7(9):3661–7.

30. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75, 4646–58.

Chapter 6

Understanding and Exploiting Peptide Fragment Ion Intensities Using Experimental and Informatic Approaches*

Ashley C. Gucinski, Eric D. Dodds, Wenzhou Li, and Vicki H. Wysocki

Abstract

Tandem mass spectrometry is a widely used tool in proteomics. This section will address the properties that describe how protonated peptides fragment when activated by collisions in a mass spectrometer and how that information can be used to identify proteins. A review of the mobile proton model is presented, along with a summary of commonly observed peptide cleavage enhancements, including the proline effect. The methods used to elucidate peptide dissociation chemistry by using both small groups of model peptides and large datasets are also discussed. Finally, the role of peak intensity in commercially available and developmental peptide identification algorithms is examined.

Key words: Peptide fragmentation, Data mining, Tandem mass spectrometry, Mobile proton model, Intensity-based algorithms, Dissociation pattern, Intensity, Statistical analysis

1. Introduction

Mass spectrometry (MS), which allows for measurement of peptide, protein, and fragment ion mass-to-charge ratios (*m/z*), is widely used in studies that aim to identify peptides and proteins. Often, these studies involve high-throughput, large-scale identification of proteins from complex mixtures (1, 2). MS is expected to continue serving an important function in this arena for many years to come due to the sensitivity, selectivity, and speed of MS-based analyses (3). The further optimization and enhancement of MS

* This chapter is dedicated to the memory of Katheryn A. Resing: colleague, collaborator and friend, who left us January 8th, 2009 after a courageous battle with cancer.

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_6, © Humana Press, a part of Springer Science+Business Media, LLC 2010

technology and data analysis capabilities for proteomics remain a highly active area of research (4–6).

While single stage mass spectrometry does play a role in protein identification, many protein identifications are performed by tandem mass spectrometry (MS/MS) of peptides derived from protein digests (7–9). In a common "bottom-up" MS/MS approach to proteomics for large-scale protein identification, peptides are produced by enzymatic digestion of a mixture of proteins. The specificity of the protease determines the sites at which peptide bonds are hydrolyzed and thus dictates the numbers, lengths, and terminal residue identities of peptides produced from a given protein. The peptides produced by digestion of a mixture of proteins are commonly separated by one or two stages of high-performance liquid chromatography (HPLC), ionized (typically by electrospray ionization, ESI) (10), and mass-selected for MS/MS fragmentation analysis. After peptide ion activation and subsequent dissociation, product ions are analyzed by *m/z* and relative intensity. This MS/MS spectral information must then be converted into peptide sequence information and in turn, protein identification. A schematic for this process is shown in Fig. 1.

Several algorithms are available that perform peptide sequencing and protein identification from MS/MS data (11–14), and additional software tools have been developed to help users consolidate and interpret database search results (15). These various protein identification algorithms have differing success rates, and current algorithms assign sequence matches to only a minority of acquired spectra. Therefore, it would be appealing to obtain sequence matches for a larger percentage of peptide spectra submitted to a given algorithm. This would allow additional proteins to be identified from a given dataset and would also

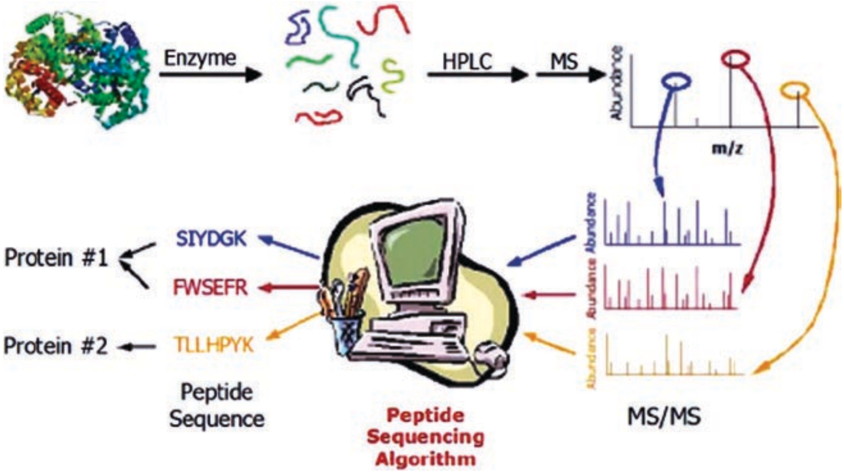


Fig. 1. Schematic of tandem mass spectrometry based protein analysis

provide a larger number of matching peptides per identified protein. Together, these improvements would lend greater confidence to protein identifications while minimizing the potential for false positive associations. It should also be noted that simplistic proteomic approaches are impractical in certain situations for a variety of reasons. For instance, genome data may not be available for a particular species (16), posttranslational modifications may require characterization (17), or the peptides being analyzed may not be protein derived (e.g., neuropeptides or peptide hormones) (18).

Some types of MS/MS scoring routines involve production of a list of expected fragment ions or generation of a predicted MS/MS spectrum. These theoretical predictions are then used to rank potentially matching sequences that lie within a given *m/z* tolerance of known sequences derived from genomic data. To date, knowledge of residue- or peptide-specific dissociation chemistry has been only sparingly incorporated into the process of spectrum prediction and match scoring. Moreover, those algorithms that do include chemically relevant criteria involve only the most simplistic implementations. For example, experimentally observed fragment ions corresponding to the neutral loss of ammonia would require the presence of arginine, lysine, glutamine, or asparagine in the fragment ion. The inclusion of these very simple and qualitative chemical dissociation rules is typically the only extent to which knowledge of peptide ion chemistry informs the informatic aspect of a proteomic experiment.

At present, fragment ion *intensity* information is disregarded or only minimally accounted for by proteomic database search algorithms. The overwhelming majority of these algorithms are based on *m/z* values only, with none of the popular approaches to database searching presently employing a sophisticated model of relative peak intensities among peptide dissociation products. Generally, this means that ion abundance information, including strong enhancement or suppression of particular ions, is not used by the algorithms. Thus, the current paradigm for MS/MS database searching in proteomics is based on only one dimension of inherently two dimensional datasets. The incomplete use of the available spectral information is largely attributed to the fact that it is not yet fully known how to most appropriately determine and exploit peptide product ion intensity information. Considering that sequence information is also encoded within the intensity dimension of an MS/MS spectrum, a chemically meaningful incorporation of fragment ion abundance into tools for proteome informatics has significant potential to improve the success rate and confidence level of sequence and protein identifications. The development of this type of platform is expected to provide a rich and thus far relatively untapped source of sequence relevant information.

A large body of research has established that the relative intensity of peptide fragment ions is remarkably sensitive to peptide composition, sequence, charge state, and the location of charges, as well as the type of instrument and activation method used (19, 20). This complex and nuanced behavior presents major challenges for the design of rigorous predictive models for peptide product ion abundances. Because our research and the research of others has shown that certain structural motifs lead to enhanced or diminished MS/MS cleavage, it is logical to consider whether inclusion of selective cleavage information for particular structural motifs into protein identification algorithms might improve identification rates. Recently, we and several other authors have made the suggestion that greater knowledge of gas-phase peptide dissociation patterns and the underlying chemical reasons for the dissociation patterns might lead to the development of improved algorithms. In order to realize the potential benefits of relative intensity information in a proteomic context, multifaceted and interdisciplinary research will be essential. First, understanding of the chemical basis for cleavage selectivity and fragment ion abundance must be advanced and refined through systematic study of model peptide systems. Second, large databases of peptide MS/MS data must be explored for distinctive spectral features that can be related to peptide sequence. Finally, these insights must be used to inform the design and implementation of improved sequencing algorithms. This chapter will address each of these areas in turn.

2. The Mobile Proton Model of Peptide Dissociation

Peptides are usually analyzed by MS as singly protonated (i.e., $[M+H]^{+}$) and multiply protonated (i.e., $[M+nH]^{n+}$) molecules. The most common method of dissociating peptides in MS/MS is collision-induced dissociation (CID), which involves the conversion of peptide ion kinetic energy into vibrational energy upon impact with neutral, inert target gas atoms or molecules. Peptides may also be subjected to tandem mass spectrometry using surface-induced dissociation (SID), which deposits vibrational energy into precursor ions by means of colliding them with a surface. Although this chapter is primarily focused on peptide ion dissociation as a result of vibrational activation, it is important to note some important alternative activation methods. In recent years, electron capture dissociation (ECD) and electron transfer dissociation (ETD) have proven to be effective dissociation methods for proteomics (21, 22). These activation techniques involve the capture of a low-energy electron by a multiply protonated peptide (in the case of ECD) or transfer of a low-energy electron from an anionic reagent to a multiply protonated peptide (in the case of ETD). While CID and SID MS/MS spectra contain predominantly b and y sequence ions, ECD and ETD MS/MS spectra contain mainly c and z ions. These collisional and electronic activation methods produce very different MS/MS spectra with a high level of complementarity. The combination of complementary activation methods, such as CID and ETD, can often provide more protein identifications than either method alone (21, 23).

Because peptides are polyfunctional molecules, the charge-carrying proton or protons may potentially occupy a number of basic sites on the side chains of amino acid residues (e.g., the side chain guanidino group of arginine residues) or along the peptide backbone (e.g., carbonyl oxygen atoms). Given a sufficient internal energy, an activated peptide ion will undergo unimolecular decay to yield fragment ions. In CID and SID, these are most commonly sequence ions of the b and y types, which are formed through dissociation mechanisms that involve the participation of a charge-carrying proton. Thus, the location of protons exerts a strong influence on the sites of cleavage (24–27). While some potential protonation sites are more favored than others, it should not be overlooked that at a given point in time and for a given distribution of internal energies, a population of ostensibly identical protonated peptides is actually a collection of variously protonated isoforms. That is, a population of protonated peptides can, in reality, be a collection of distinct ions, with the proton or protons occupying different sites. Moreover, a given protonated peptide is not static; rather, protons can be intramolecularly transferred to a number of potential sites.

The foregoing considerations serve to illuminate a general qualitative framework for describing peptide fragmentation behavior on the basis of proton mobility. While the mobile proton model alone does not provide for quantitative prediction of fragment ion intensities, the model does furnish sound chemical rationale for several well known types of enhanced and diminished cleavage. One influence of proton mobility on peptide fragmentation can be dramatically demonstrated by comparing the collision energies required to dissociate peptide ions having differing numbers of charge-carrying protons in relation to the number of basic amino acid side chains (28, 29). Those peptide ions with a number of protons greater than the number of basic amino acid residues tend to dissociate at relatively low collision energies. In these cases, each basic residue is considered to harbor a proton, leaving at least one additional, mobile proton. Dissociation of these precursor ions generally yields product ions with good sequence coverage, as under such circumstances, there are many roughly equivalent sites of protonation that may be occupied by the mobile proton. By contrast, peptide ions with a number of charge-carrying protons less than or equal to the number of basic amino acid residues (particularly, arginine residues) require significantly greater collision energies in order to efficiently dissociate. In these cases, all available protons are most favorably localized at the basic side chains, thus not allowing for a readily mobile proton. In this case, additional energy is required to mobilize these sequestered protons or to reduce basicity by an intermediate neutral loss and thus allow the participation of these protons in backbone cleavage mechanisms.

Proton mobility not only plays a role in the overall activation energy required to bring about peptide ion dissociation but also serves to explain some well-known types of selective cleavage. For example, cleavage C-terminal to aspartic acid residues (and, to a lesser extent, glutamic acid residues) is highly favored in the absence of mobile protons (30, 31). This type of enhanced cleavage has been attributed to the participation of an acidic side chain proton in the dissociation mechanism. Because the proton participating in the dissociation chemistry is not the charge-carrying proton, this type of cleavage is often described as a charge-remote pathway. When mobile protons are available, cleavage C-terminal to acidic residues becomes an essentially nonselective process. Selective cleavage is also commonly observed at the C-terminus of histidine residues, although the behavior of this cleavage is different from that seen at the C-terminus of acidic residues (32). For these peptide ions, the fragmentation occurs preferentially only in the presence of mobile protons. This observation has been interpreted as evidence that a charge-carrying proton must occupy the histidine side chain imidazole group in order to bring about the selective cleavage. By contrast, histidine-containing peptide ions with no mobile protons cleave in a nonselective manner. While these examples do not constitute an exhaustive discussion of mobile proton related selective cleavage types, they do serve to illustrate the exquisite sensitivity of peptide dissociation patterns to the chemistry of each specific ion.

3. Elucidation of Chemical Trends from Collections of Fragmentation Spectra

As mentioned previously, proteomics experiments use algorithms, such as Sequest or Mascot, to assign peptide sequences to peptide fragmentation spectra in order to identify the corresponding proteins present in a sample (12, 14). While these programs have greatly enabled progress in proteomics, they are still limited from both a practical and chemical perspective. Of the thousands of tandem mass spectra acquired in a given experiment, only a small percentage of the spectra are identified by the algorithms (33–36). This may be due in part to the simplicity of the chemical fragmentation models these algorithms use, as mentioned in the previous section (12, 14). One limitation of the fragmentation models used is that cleavages are predicted to occur almost exclusively at the amide bond between neighboring residues, regardless of amino acid residues present. As many groups have identified several reproducible residue-dependent cleavage enhancements (19, 31, 37, 38), it is clear that the algorithms do not take into account all of the chemical information available to describe a peptide fragmentation spectrum. Incorporating more chemically detailed information may help to improve the ability of an algorithm to correctly identify a peptide based on a fragmentation spectrum if a robust, fast, and sophisticated model can be developed.

A wide variety of chemical properties have been shown to affect the fragmentation pattern of a peptide. Some of those explored include size, charge state, and residues present (28, 30, 37–42). The way in which all of these factors act together to give a certain fragmentation spectrum is complex and not yet fully understood. Two main approaches have been taken in order to understand the effect of different characteristics on peptide fragmentation: systematic studies using model peptides and data mining applied to large datasets.

4. Model Peptide Studies

Several groups have used small subsets of model peptides to demonstrate trends in peptide fragmentation spectra. Tsaprailis et al. used a small set of angiotensin peptide analogs to systematically explore the effect of the neighboring residue on enhanced cleavage at histidine residues (32). Dongre et al. demonstrated the role of residue basicity, peptide length, and peptide sequence on fragmentation patterns using systematically modified leucine enkephalin analogs, polyalanine analogs, and des-Arg bradykinin derivatives (28). Figure 2 shows the fragmentation efficiency curves for a series of singly protonated polyalanine analogs with different N-termini. As the gas phase basicity of the first residue increases, additional collision energy is required to achieve the same fragmentation efficiency. The increase in energy required to achieve fragmentation within the given timescale demonstrates the ability of more basic residues to more tightly sequester the ionization proton, a result that played a role in development of the mobile proton model.

Vaisar and Urban used a similar method to examine the proline effect on peptide fragmention by looking at a series of five different peptides of the sequence Ala-Val-X-Leu-Gly (43). These studies and others clearly indicate that multiple factors are responsible for the overall fragmentation behavior of a peptide. While each of these examples can describe differences in fragmentation behavior in relation to other peptides in the study that have been varied with a systematic intent, it is not possible to either fully elucidate all of the contributions to the fragmentation spectrum, nor is it possible to draw more general conclusions of how these

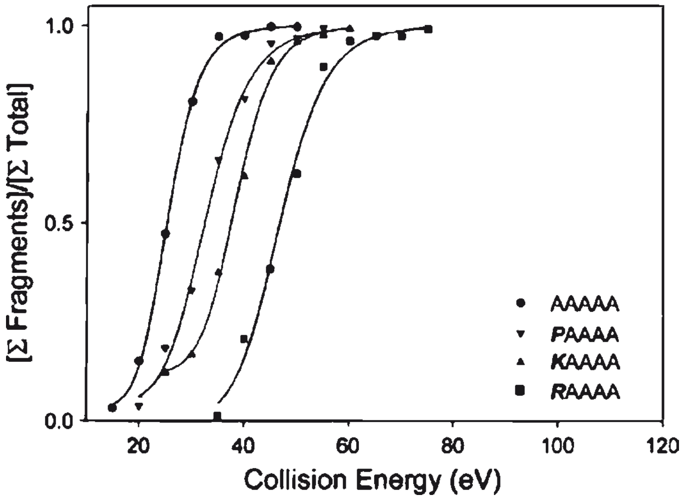


Fig. 2. Influence of gas-phase basicity on fragmentation efficiency. As gas-phase basicity increases (from A to P to K to R), the ionizing proton is more easily sequestered so that more energy is required to achieve the same fragmentation efficiency. Reproduced with permission from *J. Am. Chem. Soc.* 1996, *118*, 8365–8374. Copyright 1996 Am. Chem. Soc

factors can be applied to larger sets of spectra. Because proteomics readily generates a large number of spectra to be interpreted, and because large numbers of spectra are needed to achieve statistically valid numbers of combinations of various residues, methods that seek to discern fragmentation patterns from large sets of data may be more appropriate tools.

5. Introduction to Data Mining

Tandem mass spectrometry data are aptly suited for data mining as a typical proteomics experiment will quickly generate several thousands of widely varied MS/MS spectra. The goal of data mining is to identify underlying patterns from the spectra that can ideally be correlated to chemical phenomena that will help describe the ways in which peptides fragment. Generally, data mining can be broken down into two approaches after data acquisition: classification and pattern analysis, and/or clustering and pattern analysis.

It is important to note here that a major requirement of data mining is the availability of large, high quality datasets in which there is great certainty that the peptide sequences are correctly identified based on the corresponding fragmentation spectra. Datasets consisting of a few thousands to a few million spectra have been studied via data mining in order to elucidate trends

(33, 35, 36, 38, 40, 42, 44). Small sets of model peptides have an advantage in terms of the ease of assembling the data set because peptides and their desired analogs can readily be synthesized and easily characterized using basic MS and MS/MS measurements. Assembling a dataset with thousands of spectra in the same manner would be extremely time intensive and lacking in practicality. Rather than synthesizing thousands of peptide analogs, proteolytic digests of complex protein mixtures are analyzed via LC-MS and the corresponding MS/MS spectra are collected. As stated previously, in a given experiment of this type, as few as 10–35% of the spectra can be correctly identified. In order to trim these data sets to include only spectra that have had their sequences identified with high certainty, the data are first run through an algorithm, and the spectra that are matched to a peptide/protein with an acceptable cutoff score are saved (33, 35, 38, 42). In order to further validate a dataset, Smith and coworkers ran a complex digest through two types of mass spectrometers, an FT-ICR and an ion trap, which were coupled with identical chromatographic conditions (45). The combination of the accurate mass measurements from the FT-ICR and the fragmentation spectra from the ion-trap was paired with the use of Sequest; when Sequest identified the peptide that was within 1 ppm of the accurate mass and correlated to the fragmentation spectrum at the same retention time within a margin of error, then the spectrum was considered to be identified with very high confidence. However, this approach necessarily introduces bias because those spectra and sequences that are not identified are not represented in the database. While this method would not eliminate all incorrectly assigned peptide fragmentation spectra, it would identify a large number of high quality spectra in a relatively small amount of time.

The motivation for using a larger dataset as opposed to a set of systematically altered model peptides is that a larger distribution and variability of peptides and their corresponding fragmentation spectra will be present. With a greater distribution, the goal is to identify underlying trends in the fragmentation spectra that can be universally applied to future systems. However, many subsets are limited to one charge state or one type of peptide. Many have focused their studies on doubly charged tryptic peptides, as they are a common type of peptide ion seen (35, 38, 40, 42). Only a few researchers, including Wysocki and Zhang, have investigated the role of a variety of charge states (41, 46). While some other charge states and nontryptic peptides are less common in proteomics experiments, it is nonetheless important to acknowledge the specific bias a given dataset may contribute to the outcome of a data mining effort.

Once the dataset is assembled, data mining may proceed through two main approaches: classification and pattern analysis or clustering and pattern analysis. One common approach is to first include a preclassification step. Based on previously understood chemical principles, Huang et al. preliminarily separated data from 28,311 spectra into nine subsets based on structural features, such as proline content and basic residue content, and the charge state (41). In each subset of this study, pairwise fragmentation maps were generated to describe cleavages between all possible residue pairs. An example of this fragmentation map is shown in Fig. 3, which illustrates the y (top) and b (bottom) ion intensity patterns among doubly charged arginine (left) and lysine (right) terminated peptides. These fragmentation maps yield a plethora of information that may be integrated into future peptide identification algorithms.

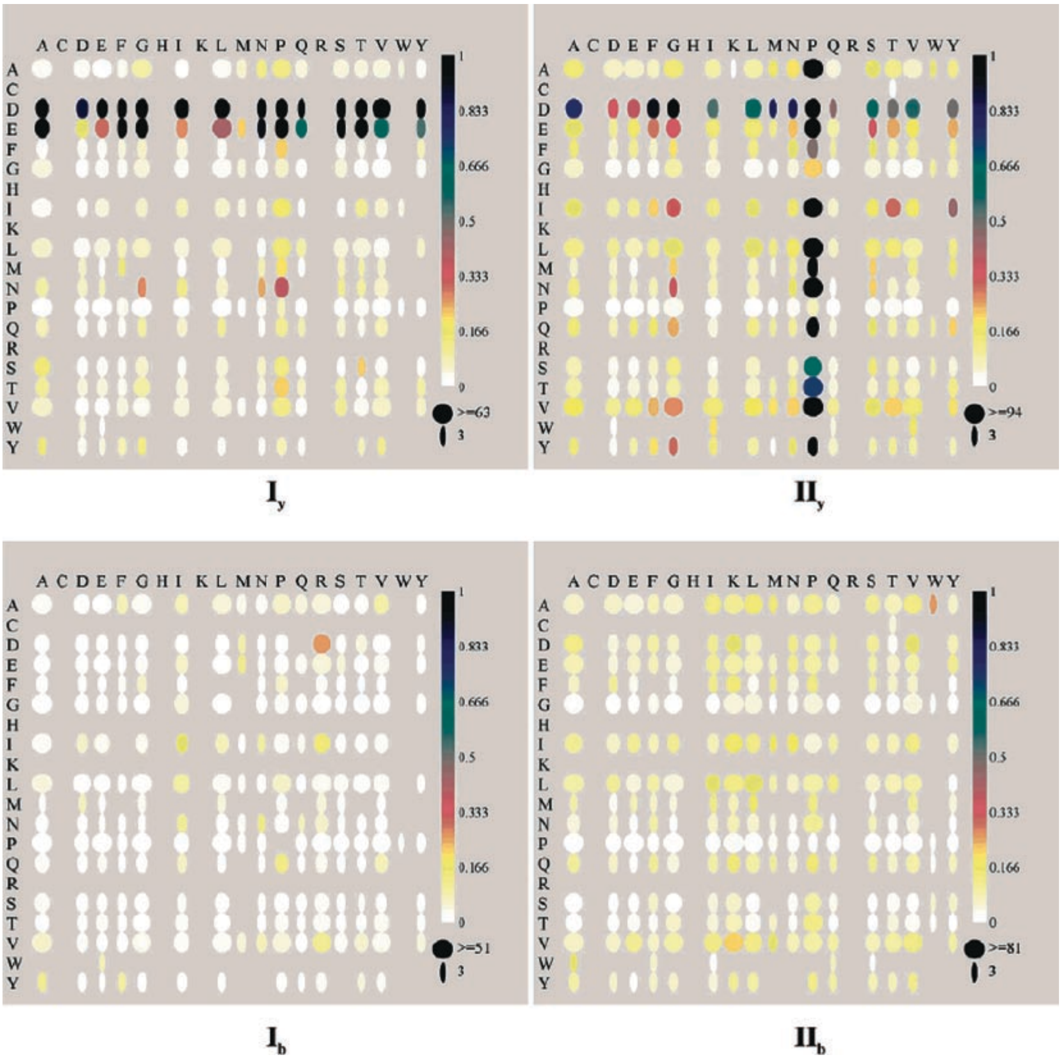


Fig. 3. Pairwise fragmentation map for singly charged peptides ending in arginine (Iy and Ib) or lysine (IIy and IIb). Reproduced with permission from *Anal. Chem.* 2005, *77*, 5800–5813. Copyright 2005 Am. Chem. Soc

In a similar approach, Tabb et al. examined trends in a database of 1,465 doubly charged tryptic peptides (35). Initially, they refined their dataset to include only doubly charged ions whose spectra contained at least 50% of the theoretically predicted ions that were fully tryptic; that is, ending in Arg or Lys without any internal Arg or Lys residues. They then examined the relationship between fragment intensity and ion series origin, fragment mass, residue type and effect on the neighboring amide bond cleavage, and the link between peptide amino acid composition and neutral fragment loss. In another study by Tabb et al., proteinase K was used to generate 2,568 nontryptic doubly charged peptides so that the role of basic residue location in a peptide could be correlated to fragmentation efficiency (42). A similar method was used by Kapp et al. to investigate trends using a dataset of 5,500 peptides. The authors demonstrated that the incorporation of a proton mobility factor could greatly improve algorithm identification success (36).

Others have used data mining to focus on specific fragmentation patterns, such as Huang's investigation of the influence of internal basic residues on the fragmentation C-terminal of the acidic residues Asp and Glu and Breci's look at fragment ion intensities due to cleavage N-terminal to Pro (37, 38). Through an examination of the b and y fragment ion intensity C-terminal to Asp when an internal His was present, Huang and coworkers were able to demonstrate that cleavage C-terminal to Asp was enhanced because of the ability of a basic His internal residue to sequester protons for doubly charged tryptic peptides. Breci et al. used a measure of the relative bond cleavage, which compares the intensity of the ions from cleavage at Pro to the intensity of all ions present in the spectrum, to determine that while cleavage N-terminal to Pro is reproducible for a certain residue, there is not enough chemical understanding as of yet to fully elucidate the entire fragmentation mechanism.

An alternative approach taken by Huang et al. was to use a penalized K-means algorithm to allow for unsupervised clustering of 28,330 spectra (47). This allowed for the peptide fragmentation spectra to cluster into four groups without the introduction of any prior chemical knowledge into the algorithm, as shown in Fig. 4. After the clustering, a decision tree was used in order to correlate the clusters to specific chemical properties. A fifth cluster for noise and outlier peaks was also generated using a method developed by Tseng, to allow for cleaner clustering. This method is important because it bypasses the need to introduce any prior assumptions and instead provides a relatively unbiased overview of the fragmentation behavior observed in the dataset as a whole.

Whittaker and coworkers have employed an alternative data mining technique that they refer to as statistical modeling, which

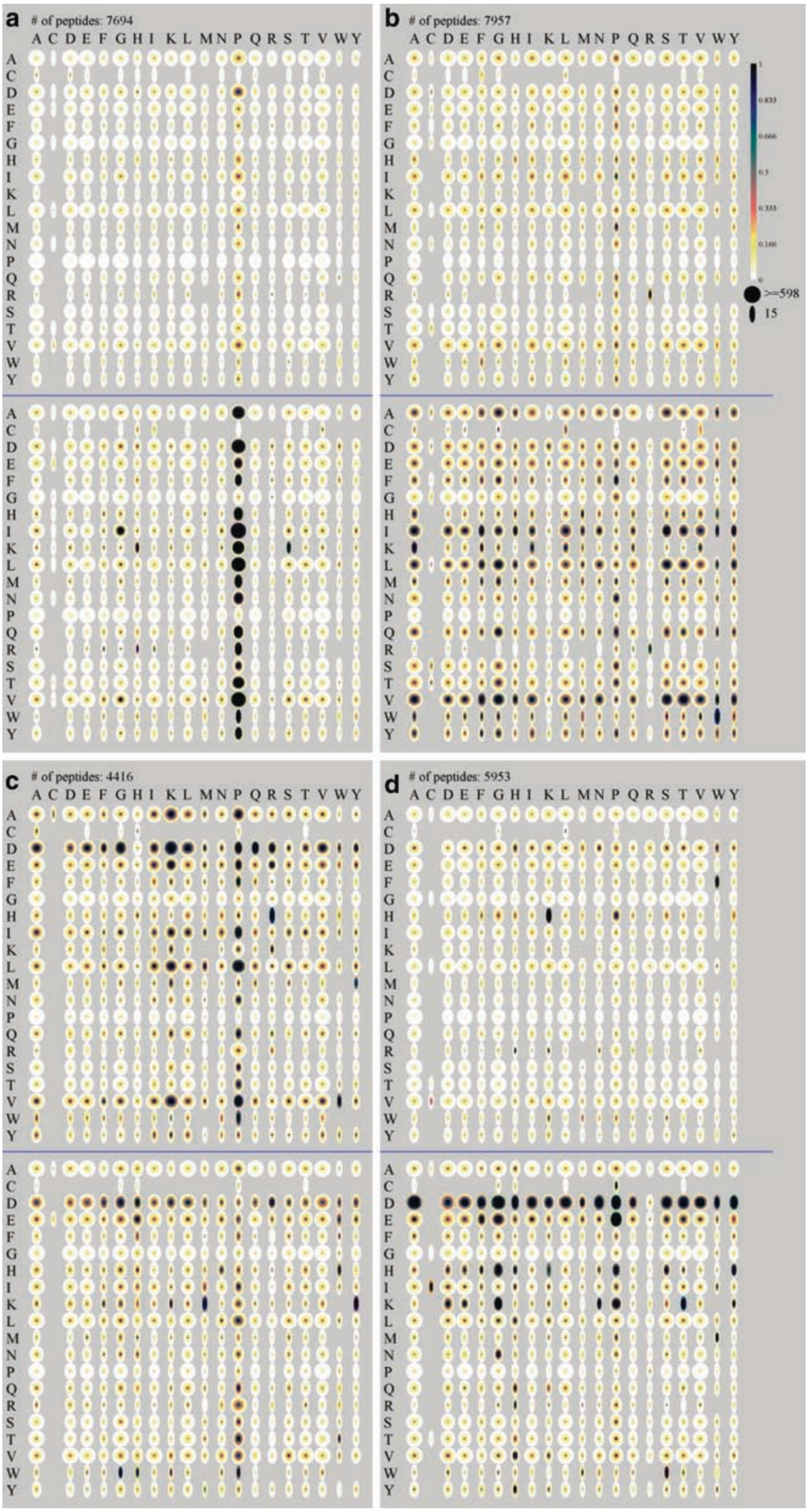


Fig. 4. Quantile maps of b (*above*) and y (*below*) ions for the four clusters identified from Huang's study using a penalized K-means algorithm for unsupervised clustering. The four clusters of spectra are characterized by the dominant cleavages patterns seen: (**a**) X–P, (**b**) I/L/V–X, (**c**) both D–X and X–P, and (**d**) D/E–X. Reproduced with permission from Proteome Res. 2008, 7, 70–79. Copyright 2008 Am. Chem. Soc

uses probabilistic models relating trends in fragmentation spectra to multiple predictor variables (39, 48). The key advantage of statistical modeling is in the ability to consider each factor simultaneously as opposed to independently. This is ideally suited for the interpretation of tandem mass spectra, as the factors dictating a particular fragmentation pattern are complex and multivariate in nature. For example, Barton et al. used models to describe b and y ion formation (separately, as they regarded different factors to influence the formation of each ion type) involving fragment ion mass, cleavage location and neighboring residues, and peptide residue composition (48).

Elias et al. used a machine learning approach to examine the ion intensities of 27,000 high quality fragmentation spectra to develop a model that can describe how likely it is that certain fragments would appear with a predicted relative intensity (33). They compared these predictions to a set of peptides that were either matched or mismatched to determine how the incorporation of ion intensity information could improve the success of the peptide identification algorithm. They saw improvements in peptide identification from 50 to 96%, suggesting that the incorporation of intensity is crucial to the improvement of these algorithms. This will be further discussed in the following section.

6. Incorporation of Fragment Ion Intensity in Peptide Sequencing Algorithms

As mentioned previously, various factors, including size, charge state, amino acid content, and charge location, can contribute to the process of gas phase peptide dissociation, making the resulting fragmentation spectra difficult to fully predict or interpret (19). This problem is compounded by the fact that most current algorithms rely on models that oversimplify the fragmentation process, thus causing valuable spectral information to be discarded. Introducing more of the available chemical information and fragmentation patterns into a sequencing algorithm could therefore allow the algorithm to more efficiently and more accurately match a peptide fragmentation spectrum to its correct matching peptide. This section will examine how several popular algorithms use the available peptide fragmentation information to predict spectral matches.

Some of the popular algorithms that are used to perform peptide sequencing or protein identification from MS/MS data include MS-Tag, SEQUEST, MASCOT, X!Tandem, OMSSA, and Phenyx (14). MS-Tag is an algorithm that was originally developed for the interpretation of MS/MS spectra that do not contain a contiguous ion series; that is, not all characteristic b and y ions are present (11). Figure 5 shows an experimental spectrum

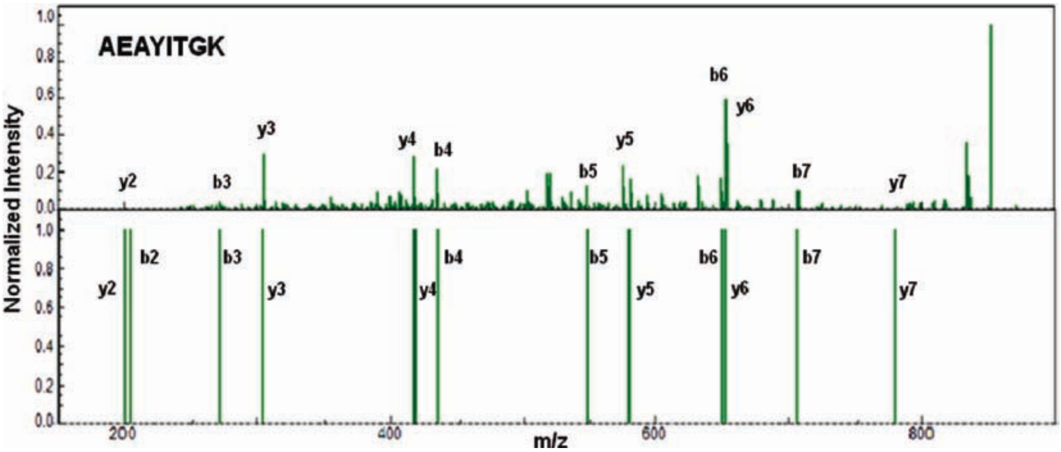


Fig. 5. Comparison of actual peptide fragmentation spectrum (*top*) to contiguous ion series (*bottom*)

and the theoretical contiguous ion series that would correspond to the sequence of the peptide AEAYITGK.

Assignment of a peptide sequence to a spectrum involves calculating the theoretical fragment ion *m/z* values for all candidate peptide sequences. MS-Tag ranks the candidate sequences in the order of increasing number of unmatched experimental fragment ions.

SEQUEST is an algorithm that correlates a given uninterpreted MS/MS spectrum with candidate sequences through the use of scoring and ranking methods based on spectral similarity by cross-correlation of the theoretically predicted spectra and the experimental spectrum (11). However, SEQUEST does not compare the raw spectra with predictions. Instead, it divides the spectrum into 10 bins and normalizes each to the most intense peak in the bin, effectively removing relative ion intensity across the entire fragmentation spectrum as a strong determinant of a match. This approach has been very successful in matching spectra to candidate sequences despite the lack of detailed rules for predicting fragment ion intensities.

MASCOT is an algorithm that contains multiple approaches to database searching, of which two use MS/MS data (MS/MS Ion Search and Sequence Query) (14). MS/MS Ion Search calculates theoretical fragment ion masses in a similar manner to that of MS-Tag before matching them to experimental spectra. Sequence Query requires some manual interpretation of the MS/MS data during which molecular weight, residue composition, and sequence qualifiers are determined for the candidate sequences. Both MASCOT strategies use the same probability-based scoring routine based on the MOWSE algorithm in which peptide size distributions (or peptide fragment size distributions) are considered with respect to protein masses (or peptide masses) in the searched database. A cutoff score for the probability that a match is a purely random event is given for each search.

X!Tandem, the most popular open source algorithm, uses intensity in its preliminary score, or hyperscore (49). This score is similar to ion intensity current, which is the sum of the intensities of all b and y ions found in the experimental spectra. This is not the same as using peak intensity information that reflects chemical fragmentation suppression or enhancement; it only acknowledges the presence of a peak. Through a statistical analysis of the hyperscore of each candidate sequence, an expectation value (E-value) describing the significance of the difference between the top match and other matches is generated and used as the main score of X!Tandem. Because this idea is common to several algorithms, the use of a hyperscore alone is not enough to significantly improve the success of X!Tandem when compared to other algorithms that use additional information and scoring stages to assign peptide spectra.

OMSSA (Open Mass Spectrometry Search Algorithm) is another example of an open source algorithm that uses expectation values as criteria, similar to X!Tandem. The older version of OMSSA only uses intensity as a threshold to filter noisy peaks (13), while the newer version has improved how intensity is used (50). In the newer edition, each peak in the experimental spectrum is ranked. The sum of the ranks of the matched peaks is compared with a normal distribution of ranks of random peak sums to calculate an expectation value. Like X!Tandem, OMSSA is complementary to Sequest because it gives an identification a probability component, whereas Sequest matches do not include probability.

Lastly, Phenyx is a platform that generates its score based on an extended match, which matches a peptide using a combination of and comparison between theoretical and experimental spectra. (51). In other words, this method incorporates structural information such as intensity, ion series contiguity, and spectral signal-to-noise ratios in addition to *m/z* information, and the extended match score reflects the quality of a match. By analyzing a testing set of spectra with known sequences, Phenyx calculates the probability of observing the above extended match information when the match is correct or if the match is purely random; the ratio of these two probabilities is the Phenyx score. When attempting to identify a peptide sequence from an unknown spectrum, similar extended match information can be generated against candidate sequences in a given database to determine the ratio score. Evaluation of the score will enable true matches to be distinguished from false.

While these algorithms are popular and successful in proteomics studies worldwide, they are not without limitations. Because every spectrum is assigned to a sequence candidate, a variety of studies have shown that in a typical MS/MS run, over 80% of the peptide identifications by SEQUEST are false and filters are necessary to eliminate those low confidence matches; programs have been developed, such as DTASelect, Peptide Prophet, and Protein Prophet, that remove these low confidence matches (52–54). However, scoring cut-off filters may also require that some correctly identified spectra are discarded in order to remove a majority of the false positive identifications. Though many proteins can still be identified using current algorithms, and the use of multiple algorithms can be combined to increase protein identification confidence as demonstrated by Searle et al. (15), these algorithms are still far from optimally meeting the rapid identification demands of the proteomics experiments that generate large volumes of peptide fragmentation spectra.

One common characteristic for all of these widely used algorithms is that they mainly utilize the mass-to-charge ratio information from a mass spectrum while ignoring the intensity component beyond the intensity threshold (12, 14). This is generally a result of insufficient knowledge of the peptide dissociation process, as we mentioned previously, though some efforts have been made recently to include intensity into peptide identifications algorithms (46, 47, 55–58). As discussed previously, reproducible intensity patterns have been identified for several residues, such as the study by Breci and coworkers on the enhanced cleavages N-terminal to proline (37). The integration of intensity is emphasized in certain algorithms not because it is more critical than *m/z*, but because it can provide additional correlating information that can assist with the peptide identification. Studies have shown that the incorporation of intensity can reduce peptide fragmentation identification error by 50–96% (33). Clearly, the use of intensity to improve peptide identification rates is an attractive prospect. Indeed, while this chapter has placed strong emphasis on the relevance of fragment ion intensity to proteomic strategies, the importance of *m/z* values cannot be minimized. Because a wide variety of MS platforms are being applied to proteomics, it is of utmost importance that proteome researchers be aware of the mass resolution and mass accuracy performance characteristics of the mass analyzer being used. Such information is essential for the appropriate setting of precursor and fragment ion mass tolerances, and the specification of average versus monoisotopic masses at the database search stage.

Different from the popular algorithms mentioned above, algorithms incorporating intensity do not work under the assumption that the all amino acid pairs and peptide patterns dissociate non-selectively to generate peaks without discrimination in intensity. Though the appearance of a given spectrum is difficult to predict, results have shown that given the same experimental conditions mass spectra are reproducible (33, 37, 46, 57). Schutz and colleagues assessed this reproducibility by using an ion trap dataset produced by the same instrument and parameters via three different methods: correlation between the intensities of two spectra as a measure of their similarity, normalized dot product of both the peak intensities from pairs of spectra, and the square root of the intensities (59). They found that MS/MS spectra, especially of peptides with low charge states, exhibit reproducible fragmentation intensities and patterns, which enables the prediction of peak intensity. Newer algorithms that incorporate complex intensity models that are based on either probability or chemical properties will be discussed below.

7. Probability Based Algorithms

Elias and coworkers used a probabilistic decision tree – specifically, a treelike feather extracting graph, which requires the members of each branch to have similar properties – to model the probability of observing certain peak intensities in a mass spectrum from 27,266 high quality spectra (33). The most confident true matches from SEQUEST were selected and decision trees were generated using 63 different attributes, including b ion length, y ion length, fraction of basic residues, and peptide length. Each node of the tree represents a chemical property that can separate the intensity into different bins, and the likelihood that a certain fragment ion peak will have a certain intensity that can be calculated from the distribution of the sizes of the resulting branches. With the input of a predicted ion from a candidate sequence, the likelihood of yielding the measured intensity in the experimental spectrum can be obtained from the decision tree. For both correctly matched and mismatched peptides, the decision trees are made and compared to serve as a guideline as to whether an identification is correct or incorrect. More than a 50% decrease in peptide identification error rate was achieved when using this method in conjunction with SEQUEST.

Another intensity based algorithm is Narasimhan's Multinomial Algorithm for Spectral Profile-based Intensity Comparison (MASPIC) scorer (60). Though based on a popular random match assumption that the correct match should have the least likelihood to be achieved randomly by chance only, MASPIC considered the possibility of random intensity matches as an alternative to using *m/z* only. This method divides the whole experimental spectrum into +1, +2, and +3 zones according to the charge of the fragment. In each zone, peaks are binned into classes with descending intensity, where lower intensity classes have more peak members. This process converts the experimental spectrum into a probability profile along the *m/z* axis. It is more likely to randomly match a predicted peak from a candidate sequence into the lower intensity class because this class has more members, thus decreasing the importance of a match with decreasing intensity. When all predicted peaks from a candidate sequence are compared with this probability profile, the number of matched and unmatched peaks for each class is counted, and further calculations are performed to give a probability of matching.

8. Chemical Property Based Algorithms

Zhang reported a kinetic model for prediction of low-energy CID spectra from sequence in 2004, with a general idea to abandon the traditional statistics model used by intensity prediction efforts and mimic the peptide dissociation process based on kinetics and the mobile proton model (57). The key assumption is that the intensity of a fragment ion is determined by the rate of the dissociation pathway generating this fragment; if the rate constants for all fragment ion pathways are known, then the relative intensity of each fragment can be predicted. Collision energy, proton density, fragmentation rate, ion cooling rate, activation energy, and gas-phase basicity are considered and incorporated into the rate calculation of eleven different backbone cleavage pathways as well as side-chain cleavages and neutral losses. Based on this iterative calculation model, Zhang developed an algorithm called MassAnalyzer, which uses a Sim score to evaluate the similarity of a simulated and experimental spectrum (57).

The kinetic model is mainly used to confirm the results from popular algorithms rather than to provide independent protein identification. This is due to various limitations, including variability between spectra acquired on different instruments under different experimental conditions and the large number of parameters that must be considered, as mentioned above. The Resing group later used this model as one part of the Manual Analysis Emulator (MAE), a program intended to improve the validation of tandem mass spectra (61). Another part of this MAE program takes into account the proportion of the ion current (PIC), which represents the percentage of intensities in an experimental spectrum that can be derived from the peptide sequence. A higher PIC score means that the program was using the most intense peaks for peptide identification as opposed to noise and low abundance peaks. With the incorporation of these two intensity-related scores, MAE yielded a better discrimination between true and false matches of SEQUEST and Mascot results.

Clearly, peptide searching algorithms utilize a variety of spectral and chemical information to assign peptide sequences to spectra. Selecting a single algorithm over another will likely lead to different sets of peptide and protein assignments based on the criteria that an algorithm uses. As briefly mentioned earlier, the use of multiple search algorithms has been shown to improve confidence of a peptide identification. Programs such as Scaffold, available from Proteome Software, provide an interface for direct comparison of MS/MS data analyzed using a variety of algorithms (15). As new algorithms are developed, it is important to understand what spectral characteristics allow the algorithm to more accurately match certain spectra to peptide sequences while the matches for other spectra with different characteristics are poor. Programs such as Scaffold will allow algorithms to be more readily compared.

9. Prospectus

We can imagine a time in the future when our fundamental knowledge and computational capabilities are sufficiently advanced to rapidly and accurately predict theoretical MS/MS spectra for any given peptide sequence. This will ultimately require that different protonation motifs, their relative probabilities of existence, their relative propensities for interconversion, and their overall contribution to dissociation kinetics all be taken into account. This would be a significant advance, as theoretical sequences could be generated to match a measured accurate mass and the corresponding synthetic tandem mass spectra could be generated and compared to the experimental spectrum. This should, in principle, allow peptide sequence identification to be obtained even in the absence of protein level information and even in the absence of genomic information. In approaching this goal, it will be necessary to continue systematic investigation of peptide structure and gas-phase unimolecular ion chemistry of protonated peptides and to incorporate the forthcoming insights into the next generation of proteomic search algorithms.

References

1. Aebersold, R., and Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chemical Reviews* **101**, 269–96.
2. Resing, K. A., and Ahn, N. G. (2005) Proteomics strategies for protein identification. *FEBS Letters* **579**, 885–89.
3. Griffin, T. J., and Aebersold, R. (2001) Advances in proteome analysis by mass spectrometry. *Journal of Biological Chemistry* **276**, 45497–500.
4. Mo, W., and Karger, B. L. (2002) Analytical aspects of mass spectrometry and proteomics. *Current Opinion in Chemical Biology* **6**, 666–75.
5. Smith, R. D. (2002) Trends in mass spectrometry instrumentation for proteomics. *Trends in Biotechnology* **20**, s3–s7.
6. Boutilier, K., Ross, M., Podtelejnikov, A. V., Orsi, C., Taylor, R., Taylor, P., and Figeys, D. (2005) Comparison of different search engines using validated MS/MS test datasets. *Analytica Chimica Acta* **534**, 11–20.
7. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winton, S., and Hauer, C. R. (1986) Protein

sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 6233–37.

8. Wysocki, V. H., Resing, K. A., Zhang, Q., and Cheng, G. (2005) Mass spectrometry of peptides and proteins. *Methods* **35**, 211–22.
9. Yates, J. R., 3rd, Speicher, S., Griffin, P. R., and Hunkapiller, T. (1993) Peptide mass maps: a highly informative approach to protein identification. *Analytical Biochemistry* **214**, 397–408.
10. 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.
11. 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. *Analytical Chemistry* **71**, 2871–82.
12. 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. *Journal of the American Society for Mass Spectrometry* **5**, 976–89.
13. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *Journal of Proteome Research* **3**, 958–64.
14. 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**, 3551–67.
15. Searle, B. C., Turner, M., and Nesvizhskii, A. I. (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *Journal of Proteome Research* **7**, 245–53.
16. 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.
17. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. *Nature Biotechnology* **21**, 255–61.
18. Fricker, L. D., Lim, J., Pan, H., and Che, F. (2006) Peptidomics: identification and quantification of endogenous peptides in neuroendocrine tissues. *Mass Spectrometry Reviews* **25**, 327–44.
19. Paizs, B., and Suhai, S. (2005) Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews* **24**, 508–48.
20. Wysocki, V. H., Tsaprailis, G., Smith, L. L., and Breci, L. A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry* **35**, 1399–406.
21. Good, D. M., Wirtala, M., McAlister, G. C., and Coon, J. J. (2007) Performance characteristics of electron transfer dissociation mass spectrometry. *Molecular and Cellular Proteomics* **6**, 1942–51.
22. Savitski, M. M., Kjeldsen, F., Nielsen, M. L., and Zubarev, R. A. (2006) Complementary sequence preferences of electron-capture dissociation and vibrational excitation in fragmentation of polypeptide polycations. *Angewandte Chemie International Edition* **45**, 5301–03.
23. Molina, H., Matthiesen, R., Kandasamy, K., and Pandey, A. (2008) Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Analytical Chemistry* **80**, 4825–35.
24. Harrison, A. G., and Yalcin, T. (1997) Proton mobility in protonated amino acids and peptides. *International Journal of Mass Spectrometry* **165–166**, 339–47.
25. Cox, K. A., Gaskell, S. J., Morris, M., and Whiting, A. (1996) Role of the site of protonation in the low-energy decompositions of gas-phase peptide ions. *Journal of the American Society for Mass Spectrometry* **7**, 522–31.
26. Johnson, R. S., Martin, S. A., and Biemann, K. (1988) Collision-induced fragmentation of [M+H]+ ions of peptides: side chain specific sequence ions. *International Journal of Mass Spectrometry and Ion Processes* **86**, 137–54.
27. Tsaprailis, G., Nair, H., Somogyi, A., Wysocki, V. H., Zhong, W., Futrell, J. H., Summerfield, S. G., and Gaskell, S. J. (1999) Influence of secondary structure on the fragmentation of protonated peptides. *Journal of the American Chemical Society* **121**, 5142–54.
28. Dongre, A. R., Jones, J. L., Somogyi, A., and Wysocki, V. H. (1996) Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: evidence for the mobile proton model. *Journal of the American Chemical Society* **118**, 8365–74.
29. Gu, C., Somogyi, A., Wysocki, V. H., and Medzihradszky, K. F. (1999) Fragmentation of protonated oligopeptides XLDVLQ (X=L, H, K or R) by surface induced dissociation: additional evidence for the 'mobile proton' model. *Analytica Chimica Acta* **397**, 247–56.
30. Gu, C., Tsaprailis, G., Breci, L., and Wysocki, V. H. (2000) Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in

fixed-charge derivatives of Asp-containing peptides. *Analytical Chemistry* **72**, 5804–13.

31. Rozman, M. (2007) Aspartic acid side chain effect-experimental and theoretical insight. *Journal of the American Society for Mass Spectrometry* **18**, 121–27.
32. Tsaprailis, G., Nair, H., Zhong, W., Kuppannan, K., Futrell, J. H., and Wysocki, V. H. (2004) A mechanistic investigation of the enhanced cleavage at histidine in the gas-phase dissociation of protonated peptides. *Analytical Chemistry* **76**, 2083–94.
33. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., and Gygi, S. P. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology* **22**, 214–19.
34. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Jonscher, K. R., Pierce, K. G., Old, W. M., Cheung, H. T., Russell, S., Wattawa, J. L., Goehle, G. R., Knight, R. D., and Ahn, N. G. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Analytical Chemistry* **76**, 3556–68.
35. 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. *Analytical Chemistry* **75**, 1155–63.
36. Kapp, E. A., Schuetz, F., Reid, G. E., Eddes, J. S., Moritz, R. L., O'Hair, R. A., Speed, T. P., and Simpson, R. J. (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Analytical Chemistry* **75**, 6251–64.
37. 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. *Analytical Chemistry* **75**, 1963–71.
38. Huang, Y., Wysocki, V. H., Tabb, D. L., and Yates, J. R. (2002) The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *International Journal of Mass Spectrometry* **219**, 233–44.
39. Barton, S. J., and Whittaker, J. C. (2008) Review of factors that influence the abundance of ions produced in a tandem mass spectro-meter and statistical methods for discovering these factors. *Mass Spectrometry Reviews* **28(1)**, 117–187.
40. Huang, Y., Triscari, J. M., Pasa-Tolic, L., Anderson, G. A., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2004) Dissociation behavior of doubly-charged tryptic peptides: correlation of gas-phase cleavage abundance with ramachandran plots. *Journal of the American Chemical Society* **126**, 3034–35.
41. Huang, Y., Triscari, J. M., Tseng, G. C., Pasa-Tolic, L., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2005) Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Analytical Chemistry* **77**, 5800–13.
42. Tabb, D. L., Huang, Y., Wysocki, V. H., and Yates, J. R. (2004) Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry* **76**, 1243–48.
43. Vaisar, T., and Urban, J. (1996) Probing the proline effect in CID of protonated peptides. *Journal of Mass Spectrometry* **31**, 1185–87.
44. Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., and Pevzner, P. A. (2008) Clustering millions of tandem mass spectra. *Journal of Proteome Research* **7**, 113–22.
45. Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y. F., Conrads, T. P., Veenstra, T. D., and Udseth, H. R. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2**, 513–23.
46. Zhang, Z. (2005) Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Analytical Chemistry* **77**, 6364–73.
47. Huang, Y., Tseng, G. C., Yuan, S., Pasa-Tolic, L., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2008) A data-mining scheme for identifying peptide structural motifs responsible for different MS/MS fragmentation intensity patterns. *Journal of Proteome Research* **7**, 70–79.
48. Barton, S. J., Richardson, S., Perkins, D. N., Bellahn, I., Bryant, T. N., and Whittaker, J. C. (2007) Using statistical models to identify factors that have a role in defining the abudnace of ions produced by tandem MS. *Analytical Chemistry* **79**, 5601–07.
49. Craig, R., and Beavis, R. C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry* **17**, 2310–16.
50. Geer, L. Y., Bai, D. L., Kowalak, J. A., Chi, A., Xu, M., Shabanowitz, J., Markey, S. P., Hunt, D. F., and Bryant, S. H. (2008), National Library of Medicine, NIH, Bethesda, MD. University of Virginia, Charlottesville, VA, National Institute of Mental Health, NIH, Bethesda, MD.
51. Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: towards high-throughput tandem mass spec-

trometry data identification. *Proteomics* **3**, 1454–63.

52. Tabb, D. L., McDonald, W. H., and Yates, J. R. (2002) DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *Journal of Proteome Research* **1**, 21–26.
53. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifiactions made by MS/MS and database search. *Analytical Chemistry* **74**, 5383–92.
54. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry* **75**, 4646–58.
55. Gibbons, F. D., Elias, J. E., Gygi, S. P., and Roth, F. P. (2004) SILVER helps assign peptides to tandem mass spectra using intensity-based scoring. *Journal of the American Society for Mass Spectrometry* **15**, 910–12.
56. Havilio, M., Haddad, Y., and Smilansky, Z. (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Analytical Chemistry* **75**, 435–44.
57. Zhang, Z. (2004) Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry* **76**, 3908–22.
58. Zhou, C., Bowler, L., and Feng, J. (2008) A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC Bioinformatics* 9325.
59. Schütz, F., Kapp, E. A., Simpson, R. J., and Speed, T. P. (2003) Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochemical Society Transactions* **31**, 1479–83.
60. Narasimhan, C., Tabb, D. L., VerBerkmoes, N. C., Thompson, M. R., Hettich, R. L., and Uberbacher, E. C. (2005) MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Analytical Chemistry* 77, 7581–93.
61. Sun, S., Meyer-Arendt, K., Eichelberger, B., Brown, R., Yen, C.-Y., Old, W. M., Pierce, K., Cios, K. J., Ahn, N. G., and Resing, K. A. (2007) Improved validation of peptide MS/MS assignments using spectral intensity prediction. *Molecular and Cellular Proteomics* **6**, 1–17.
62. Barton, Sheila J. Whittaker, John C. (2009) Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrometry Reviews*, **28**(1), 177-187.
63. Zhou Cong, Bowler Lucas D., Feng Jianfeng (2008) A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC bioinformatics*, 9 325.

Chapter 7

Spectral Library Searching for Peptide Identification via Tandem MS

Henry Lam and Ruedi Aebersold

Abstract

Spectral library searching is a new approach in proteomic data analysis that promises to address some of the shortcomings of sequence database searching, currently the dominant method for inferring peptide identifications from tandem mass spectra. In spectral searching, a spectral library is first meticulously compiled from a large collection of previously observed and identified peptide MS/MS spectra. The unknown spectrum can then be identified by comparing it to all the candidates in the spectral library for the best match. It offers the benefits of tremendous speed gain and increase in sensitivity and selectivity, compared to sequence searching. This article provides a concise roadmap for the proteomics researchers to start using spectral library searching in their data analysis workflow.

Key words: Peptide identification, Mass spectrometry, Spectral library, Spectral searching

1. Introduction

Traditionally, the inference of the peptide sequence from its characteristic tandem mass spectra is most often achieved by sequence (database) searching (Chapter 3). Several popular computational tools developed for this purpose have emerged over the years, each employing different algorithms and heuristics to achieve an acceptable balance of sensitivity and accuracy (1). Unfortunately, traditional sequence searching is a challenging, error-prone, and computationally expensive exercise. Despite the tremendous improvement in computer hardware and software over the past decade, this step often remains the bottleneck of any given proteomics experiment. The requirement of computational resources is also substantial, limiting the use of this powerful

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_7, © Humana Press, a part of Springer Science+Business Media, LLC 2010

technique to only those research groups that can afford the costly computational infrastructure (2, 3).

Spectral searching is an alternative approach that promises to address some of the shortcomings of sequence searching. In spectral searching, a spectral library is meticulously compiled from a large collection of previously observed and identified peptide MS/MS spectra, either by a centralized library builder, or by individual researchers. The unknown spectrum can then be identified by comparing it to all the candidates in the spectral library for the best match (4, 5). This approach has been commonly employed for mass spectrometric analysis of small molecules with great success but has only become possible for proteomics very recently. The chief difficulty, that of generating enough high-quality experimental spectra for compilation into spectral libraries, has been overcome by the recent explosion of proteomics data and the availability of public data repositories. Several attempts at creating and searching spectral libraries in the context of proteomics have been published within the past few years, all demonstrating the tremendous improvement in search speed and the great potential of this method in complementing sequence searching in many proteomics applications (6–8).

The defining features of spectral searching are (1) the use of experimental, as opposed to theoretical, spectra to match query spectra, and (2) a much reduced search space compared to traditional sequence searching. Both factors contribute to the improved performance of spectral searching.

Spectral searching compares experimental spectra to experimental spectra, whereas sequence searching compares experimental spectra to theoretical spectra. In general, the theoretical spectra considered in sequence searching are very simplistic (e.g., only including b- and y-type ions, at a fixed intensity), and do not resemble the experimental spectra that they are supposed to match. On the other hand, armed with previously observed experimental spectra compiled into spectral libraries, spectral searching can take full advantage of all spectral features, including actual peak intensities, neutral losses from fragments, and various uncommon or even uncharacterized fragments, to determine the best match. The similarity scoring of spectral searching is therefore more precise, and will generally provide better discrimination between good and bad matches. This usually results in much superior statistics (e.g., sensitivity, false discovery rates) for the search results, compared to sequence searching.

Spectral searching also benefits from a much reduced search space. Because spectral libraries are compiled from previously observed and identified peptide ions of a proteome, a spectral search engine only considers as candidates a small fraction of all putative peptide sequences derivable from a sequence database. It is well known that most of these putative peptide ions considered in sequence searching are never observed in practice for a variety of reasons, ranging from the absence or the scarcity of the protein in the sample, to the inability to the peptide to ionize efficiently to be observed in the mass spectrometer. With typical search parameters, the search space of spectral searching can be several orders of magnitude smaller than sequence searching, yielding a corresponding speed gain. In addition, the reduced search space also means that the search engine will have fewer candidates to consider, which often translates to an improvement in discrimination power.

Needless to say, the narrowing of the search space to previously identified peptides also limits the application of spectral searching to situations where discovery of novel peptides or proteins is not the goal. This perhaps is the biggest drawback of the spectral searching approach, albeit one that is sometimes overstated. Today, spectral libraries are being compiled from multitudes of large-scale experiments that cover a wide range of sample sources, sampling techniques, instrumentation, and data analysis methods. It is becoming increasingly unlikely that an ordinary proteomic experiment studying commonly studied systems and employing well-known techniques will yield many newly discovered peptides not covered in these libraries. On the other hand, more and more opportunities of scientific discovery lie in understanding how these known segments of the proteome change with time and circumstances and how they interact in concert to produce biological function. Consequently, there is a shifting emphasis from discovery-oriented endeavors to targeted and quantitative proteomics in which one is merely interested in studying known and previously observed peptides (2, 9). Spectral searching is well suited to this type of workflows.

There are also ample opportunities to use spectral searching as a part of an integrated data analysis pipeline. For instance, it can be used as a first-pass search to identify all previously known peptides, before one resorts to more expensive and brute-force approaches for yet unidentified spectra, which can perhaps be aided in some manner by the identifications already made by spectral searching. Spectral searching can also be part of an iterative approach in which sequence searching is employed first on a reference sample to construct a reference spectral library, followed by quick spectral searching to confirm identifications of interest in many parallel experiments. This would be ideal for quantitative proteomic experiments with many samples and replicates, such as a time series experiment, or a clinical study involving many subjects.

There have been several published reports on various aspects of spectral library building and searching in the context of proteomic data analysis, with associated software tools made available freely to the research community (Table 1). A detailed review and comparison of these options is not the intent of this article, which is aimed at providing a concise roadmap to the reader to start using spectral library searching in their data analysis workflow. Nor it is intended to provide evidence for the effectiveness of

Table 1
Free or open-source software for spectral library searching

	NISTMS	X!Hunter	Bibliospec	SpectraST
Supported libraries (format)	NIST libraries (.msp)	X!Hunter libraries (.hlf)	Bibliospec libraries (.ms2)	.msp .hlf .ms2 SpectraST libraries (.splib)
Supported data formats	.dta .pkl .mgf .msp	.mzXML .mzData .dta .pkl .mgf .bioml	.dta	.mzML .mzXML .mzData .msp .mgf .dta
Platform	Windows	Windows LINUX Remote webserver	Windows LINUX	Windows LINUX Remote webserver
Remarks	Extension of well-known MS search program for small molecules; One-click installer	X!Tandem-like scoring Need web server and Perl installation		Integrated with Trans-Proteomic Pipeline (TPP); One-click Windows installer
Reference	–	(6)	(7)	(8)

spectral searching, for which the reader is referred to the publications listed in Table 1. Finally, it should also be emphasized that spectral searching in proteomics is a new and rapidly developing field. Many aspects of its implementation and usage are likely to evolve over time.

2. Methods

2.1. Obtaining Software

The websites from which to download spectral searching software, as well as any helpful sites with instructions and documentation, are listed in Table 2. Easy one-click installation is available for NISTMS and SpectraST (see Note 1) on the Windows platform.

Table 2
Useful websites

NISTMS	Software download http://peptide.nist.gov/ Library download http://peptide.nist.gov/ http://www.peptideatlas/speclib/ Instructions http://peptide.nist.gov/
X!Hunter	Software download ftp://ftp.thegpm.org/projects/xhunter/binaries ftp://ftp.thegpm.org/projects/xhunter/source Library download ftp://ftp.thegpm.org/projects/xhunter/libs Instructions http://h201.thegpm.org/docs/xhunter_system.html Web client to X!Hunter on remote server http://xhunter.thegpm.org/
Bibliospec	Software download http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/bibliospec.php Library download http://proteome.gs.washington.edu/software/bibliospec/documentation/libs.html Instructions http://proteome.gs.washington.edu/software/bibliospec/documentation/index.html
SpectraST	Software download http://sourceforge.net/project/showfiles.php?group_id=69281 Library download http://www.peptideatlas/speclib/ http://peptide.nist.gov/ Instructions http://tools.proteomecenter.org/wiki/index.php?title=SpectraST Information on Trans-Proteomic Pipeline and open XML data formats http://tools.proteomecenter.org/software.php http://tools.proteomecenter.org/wiki/ Web client to SpectraST on remote server http://www.peptideatlas.org/spectrast/

For the other tools, please refer to the associated README files for installation instructions.

2.2. Obtaining Spectral Libraries

There are two main sources of spectral libraries, public and custom-built. Public libraries are compiled by dedicated library builders, such as the National Institute of Standards and Technology (NIST) (see Note 2) and the Global Proteome Machine (GPM) (see Note 3). These public libraries are built from a large number of contributed datasets from many researchers.

They are comprehensive and high-coverage libraries of popular model organisms that are suitable for general use. Table 2 contains several websites from which public libraries can be downloaded for each spectral search engine. However, please be reminded that currently not all library formats and spectral search tools are compatible (*see* Table 1) or designed to work well together.

Custom-built libraries are built by individual researchers and are specialized to their biological systems of interest and their instrumentation setup. A custom-built spectral library is essentially a concise summary of the individual research group's observed proteomes or subproteomes of interest, and can be a very useful living resource for the research effort. X!Hunter, Bibliospec, and SpectraST all provide means for the user to construct spectral libraries from sequence search results. The actual procedures are however quite involved and will not be covered here.

2.3. Data Conversion

Prior to searching, data files containing the query spectra have to be converted to supported formats by the search engine. Unfortunately, raw data straight from the mass spectrometer is often encoded in vendor-specific proprietary formats and cannot be directly processed by spectral search engines. An additional step is required to convert or export them into one of the many open formats in use in the community. These can be simple text-based peak lists (SEQUEST's .dta format, Mascot's .mgf format) or XML-based standardized formats (.mzXML (10), .mzData, . recently mzML). All engines accept simple text-based peak lists with minor differences, which can be readily manipulated and interconverted with simple scripts. For larger-scale experiments, popular open XML that encapsulate entire MS runs are more compact and convenient. Please *see* Table 1 for supported formats for each spectral search engine. For information on data formats, please *see* Chapter 11.

2.4. Library Searching

Once the data is in supported formats, spectral library searching is usually launched within a graphical user interface (NISTMS, web interfaces for X!Hunter and SpectraST), or with commands issued on the LINUX or Windows command-line. For detailed instructions on how to use each tool, please refer to the respective user manuals or websites listed in Table 2.

Typical search parameters that the user needs to specify are: library to be searched, precursor mass or *m/z* tolerance, output format, and sometimes charge states allowed. Note that unlike sequence searching, for which the user needs to limit the search space by various means (e.g., enzymatic cleavage rules, modifications considered), the search space for spectral searching is fixed by the library coverage. Because spectral libraries are already limited to previously observed peptides and are therefore quite manageable, it is possible and often advisable to cast a wide net by setting a wide precursor mass or *m/z* tolerance (at least 3 Da or Th). This is true even for high-mass accuracy instruments. (A search space that is too small can be false positive-prone. It is preferable to use the mass accuracy constraints in the subsequent statistical validation.).

Due to the difference in peptide fragmentation patterns, there is some penalty in performance if the query and library spectra are from different types of mass spectrometers. Fortunately however, preliminary studies have shown that spectral searching is still highly effective even in those circumstances (6, 8).

2.5. Statistical Validation and Result Interpretation

It is important to note that spectral searching, very much like sequence searching, is but one step in the data analysis of proteomic data. At the end of spectral searching, each query spectrum is matched to a highest-scoring library spectrum, and various similarity and significance metrics are provided along with the putative identification. Statistical validation refers to the subsequent step in which these putative identifications are assigned confidence and the error rates estimated. There are three major approaches to statistical validation in proteomics, each with different assumptions. Expectation value-based assessment is available for NISTMS and X!Hunter (*see* Chapter 5), mixture model-based probability assignment is available for SpectraST (with PeptideProphet (11), *see* Chapter 19), and decoy searching (*see* Chapter 6) is available for all, provided one can identify a suitable decoy library. The most straightforward, and probably easiest, approach for decoy-based validation is to use library spectra of a different organism as decoys. The obvious caveat of this approach is the probable presence of some identical or homologous peptides in both target and decoy libraries, which can lead to errors in false discovery rate estimation. It is therefore wise to minimize this by using phylogenetically distant species for this purpose, or to detect and remove identical and homologous peptides from the decoy library beforehand (see Note 4).

3. Notes

1. The open-source software SpectraST is developed at the Institute for Systems Biology as part of the Trans-Proteomic Pipeline (TPP, *see* Chapter 19) software suite (12), available freely with ongoing technical support and periodic updates. The TPP provides other useful components of a typical proteomic data analysis pipeline, including raw data conversion, statistical validation, quantification and data visualization, etc. It is also available for Windows and LINUX platforms.

TPP can be obtained at the website http://sourceforge.net/project/showfiles.php?group_id=69281. Select the latest release for Trans-Proteomic Pipeline. A native Windows installer for TPP is available, along with a zip archive of the source code. For LINUX installation, unzip the archive, and follow the compilation and installation instructions in the file README in the top directory. All components of TPP, along with SpectraST, will be compiled and installed.

2. Currently, NIST's spectral libraries are available for six major model organisms: human (*Homo sapiens*), yeast (*Saccharomyces cerevisiae*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), fruit fly (*Drosophila melanogaster*), and *Escherichia coli*. NIST libraries can be searched with NISTMS or SpectraST. For human and yeast, separate libraries are provided for ion trap (IT) instruments and time-of-flight (TOF) instruments, although the latter are much smaller at the moment. The library building effort is ongoing, and updates of substantial improvement are expected every year. Visit http://peptide.nist.gov/ to download the libraries in .msp format directly from NIST, or http://www.peptideatlas.org/speclib/ to download libraries in either the .msp or SpectraST's .splib formats from the PeptideAtlas portal. (For .msp files, a SpectraST command needs to be run to convert them into SpectraST's format – mainly to index the library entries for fast search.) NIST libraries can be viewed simply by opening the .msp files (or the similar SpectraST's .sptxt files) in a text editor. Refer to NIST's documentation for a description of the fields contained therein.

3. Currently, GPM's spectral libraries are available for human, yeast, mouse, rat, several other vertebrates, two plants (*A. thaliana* and *A. fumigatus*), and several bacteria. They can be downloaded at ftp://ftp.thegpm.org/projects/xhunter/libs in .hlf formats, which are machine-readable files handled by X!Hunter only. (SpectraST can convert .hlf files into SpectraST's format and search GPM libraries; however, the search engine will not perform as well as with NIST libraries in terms of discriminating power.) Notably, GPM libraries only keep the top 20 peaks of a spectrum, whereas NIST libraries keep the full spectrum.

4. SpectraST provides a feature ("Subtract homolog") to remove any decoy spectrum with an identical or homologous identification to a spectrum in the target library.

References

1. Steen, H. and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711.
2. Domon, B. and Aebersold, R. (2006) Challenges and opportunities in proteomics data analysis. *Mol. Cell. Proteomics.* **5**, 1921–1926.

3. Patterson, S. D. (2003) Data analysis – the Achilles heel of proteomics. *Nat. Biotechnol.* **21**, 221–222.
4. Yates, J. R., III, Morgan, S. F., Gatlin, C. L., Griffin, P. R. and Eng, J. K. (1998) Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal. Chem.* **70**, 3557–3565.
5. Stein, S. E. and Scott, D. R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866.
6. Craig, R., Cortens, J. C., Fenyo, D. and Beavis, R. C. (2006) Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* **5**, 1843–1849.
7. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. and MacCoss, M. J. (2006) Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **78**, 5678–5684.
8. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N. et al. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667.
9. Kuster, B., Schirle, M., Mallick, P. and Aebersold, R. (2005) Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583.
10. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelsang, M., Deutsch, E. W. et al. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22**, 1459–1466.
11. 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.
12. Keller, A., Eng, J., Zhang, N., Li, X. J. and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**, 17.

Chapter 8

De Novo Sequencing Methods in Proteomics

Christopher Hughes, Bin Ma, and Gilles A. Lajoie

Abstract

The review describes methods of *de novo* sequencing of peptides by mass spectrometry. De novo methods utilize computational approaches to deduce the sequence or partial sequence of peptides directly from the experimental MS/MS spectra. The concepts behind a number of de novo sequencing methods are discussed. The other approach to identify peptides by tandem mass spectrometry is to match the fragment ions with virtual peptide ions generated from a genomic or protein database. De novo methods are essential to identify proteins when the genomes are not known but they are also extremely useful even when the genomes are known since they are not affected by errors in a search database. Another advantage of de novo methods is that the partial sequence can be used to search for posttranslation modifications or for the identification of mutations by homology based software.

Key words: Proteomics, Tandem mass spectrometry, De novo sequencing, Sequence tags, Peptide fragmentation, Homology

1. Introduction

Mass spectrometry (MS) plays an increasingly important role in the biological sciences. MS has long been used for the analysis of small molecules but more recently has become a tool of choice for the characterization of proteins (1–4). Modern mass spectrometers are capable of analyzing, in a relatively short time, mixtures of thousands of peptides to derive sequence information. Because of its sensitivity, accuracy, and robustness, MS has replaced earlier chemical and enzymatic methods for sequencing proteins. The recent developments in MS, separation sciences, and bioinformatics have all contributed to the rapid emergence of proteomics. In fact, proteomics, or the global study of the entire protein complement of cells or tissues, has become a field of its own. Due to these technical advances, it is now possible to perform, in a very

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_8, © Humana Press, a part of Springer Science+Business Media, LLC 2010

high throughput manner, the analysis of very complex protein mixtures. Proteomics can provide answers to questions that cannot be probed by genomics analysis especially in regard to protein modifications, localization, and quantification.

1.1. MS Instrumentation

Early forms of mass spectrometry were not amenable to the analysis of protein samples. Ionization techniques such as electron ionization (EI) and chemical ionization (CI) would cause high levels of in-source fragmentation in biomolecules. The discovery of soft ionization techniques such as matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) revolutionized the analysis of biomolecules by MS (5–7). Both of these ionization types are classified as "soft" ionization techniques because there is little fragmentation of the ionized species, such that the mass of the intact molecular ion can be measured. While MALDI produces mainly singly charged ions, ESI will typically generate multiply charged ions for peptides and proteins. One distinct advantage of ESI is its compatibility with a number of online separation techniques. Several online methods for separating complex peptide mixtures prior to their entrance into the mass spectrometer have been developed. These include high pressure liquid chromatography (HPLC), nano-HPLC (LC), and strong ion exchange chromatography (SCX). Some can be used in an orthogonal fashion, i.e., SCX/LC, or by LC in basic conditions followed by LC in acidic conditions (2D LC/LC). These separation methods allow for better peptide coverage of a sample by decreasing the complexity of the sample entering the MS.

A variety of mass analyzers and instrument designs are currently available for proteomics experiments. Most instrument designs found in research laboratories today are hybrid instruments (Fig. 1), combining two or more mass analyzers, such as:

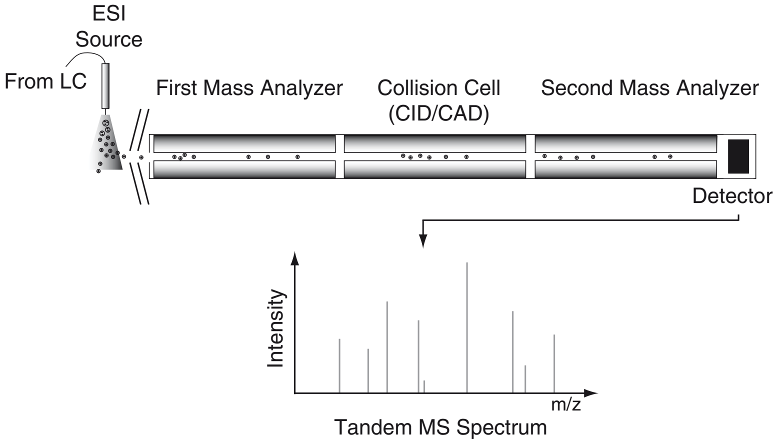


Fig. 1. Mass spectrometer

quadrupole-TOF (Q-TOF) (8, 9), triple quadrupole (Q3) instruments, some with linear ion traps (LIT, LTQ) (10), quadrupole-Fourier transform ion cyclotron resonance (qFT-ICR) (11) instruments, and LTQ-Orbitrap (12). All of these instrument platforms have advantages and disadvantages. LIT provide high sensitivity peptide spectra at a rapid rate but lack the mass accuracy and resolution for many applications. Q-TOF instruments provide good mass accuracy and resolution. Despite their relatively low duty cycles, Q-TOFs have been, until recently, the workhorse of most proteomics laboratories. The ultrahigh mass accuracy, mass range, and resolution of more recent qFT ICR and LIT-Orbitrap instruments make them the most amenable to a wide range of sample types, as well as to de novo sequencing techniques. For a comprehensive review of instruments types and their characteristics (*see* ref. 1, 10).

1.2. Proteomics Experiments and Tandem Mass Spectrometry

In a typical proteomics experiment, proteins are first digested with a protease, most commonly trypsin, and the resulting peptides analyzed by LC MS analysis. The mass spectrometry analysis takes place in three steps. First, the mass of the intact peptide ions eluting in a given time window is measured. In the second step, specific precursor ions are selected in the first mass analyzer (e.g., Q1), primarily based on their relative abundance, and fragmented in a collision cell (Q2). The mass, or more accurately the *m/z*, of each fragment ion, or product ion, is recorded in a second mass analyzer (e.g., TOF, LIT, Orbitrap, ICR cell), for each of the precursor ions. This type of experiment is referred to as a tandem mass spectrometry or MS/MS. The detailed analysis of the product ions will provide information on the peptide sequences and, by inference, the identification of the proteins. This strategy for protein identification is known as “bottom-up” proteomics (13, 14).

1.3. Peptide Fragmentation

The most common form of fragmentation is collisional-induced dissociation (CID), also known as collisionally activated dissociation (CAD) (15–17). CID/CAD will result primarily in the formation of *b* and *y* series ions from the precursor ion. These ions are formed through random cleavage of the peptide bonds, where *b*-ions are N-terminal ions, and *y*-ions are the C-terminal series of ions. A typical MS/MS spectrum contains *b* and *y* series ions, and an ideal spectrum contains a full set of both (Fig. 2). Most often, in CID generated spectra for tryptic peptides, the *y*-series ions give stronger signals than the *b*-series, especially at high *m/z*. As a result, *y*-series of ions will often be the main series of ions that are matched in a tandem mass spectrum. However, there are other characteristic fragments for both *b* and *y*-series ions that are frequently observed and that are useful for the identification of peptides. Ions in both the *y*- and *b*-series will often lose ammonia

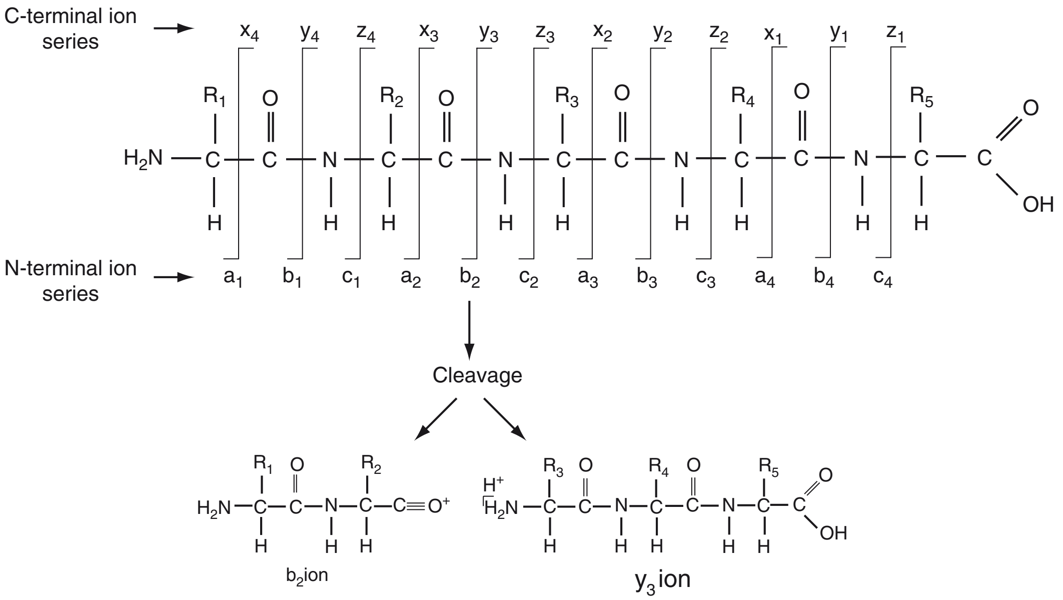

C-terminal ion series
x4 y4 z4 x3 y3 z3 x2 y2 z2 x1 y1 z1
N-terminal ion series
a1 b1 c1 a2 b2 c2 a3 b3 c3 a4 b4 c4
Cleavage
b2ion
y3ion

Fig. 2. Peptide fragmentation

on residues R, K, Q, and N, and will lose water at S, T, E and D residues (18). Immonium ions can also provide valuable information as to which amino acids are present in a peptide. Immonium ions are formed through the combination of *a*-type and *y*-type cleavage. Observation of these ions in the low mass region of the mass spectrum can give clues as to the presence or absence of specific amino acids. Other characteristic peaks such as *a*-series ions adjacent to *b*-series ions generated through the loss a carbon monoxide (–28) from the acylonium group of *b*-series ions can help in the assignment of ion identity. Peptide fragmentation mechanisms have been well characterized by numerous groups (19–29).

While CID is the most common fragmentation method, electron capture (ECD) and electron transfer dissociation (ETD) have been implemented in more recent mass spectrometers (30–32). ECD and ETD peptide ions are fragmented after reaction with electrons generated from filaments or anionic species, respectively. ECD and ETD form *c* and *z* type ions through cleavage of the peptide bond between the amino group and alpha carbon (Fig. 2). ECD and ETD are more efficient than CID for larger, multiply charged ions. The combination of both CID and ETD or ECD is now available on instruments such as Orbitrap and qFT-ICR types of mass spectrometer. This allows for the acquisition of spectra from the same sample using two fragmentation techniques generating complementary series of fragment ions.

1.4. Spectral Identification

With the ability to generate MS/MS spectra for thousands of peptides in a single experiment, modern mass spectrometers provide unparalleled efficiency for potential peptide identification (33). Along with spectral generation comes the requirement to accurately determine the sequence and identify the peptides present in a given sample (34). In early proteomics experiments, peptide identification was carried out manually because of the lack of software for automated interpretation. Identification was initially performed using an early incarnation of the de novo sequencing method whereby researchers would use mass differences between peptide ion peaks to determine sequence. Due to the ability of newer MS instrumentation to obtain thousands of MS/MS in a single experiment, it is no longer feasible to interpret spectra manually.

There are now many software platforms capable of performing data analysis using sophisticated algorithms and scoring schemes (35, 36). Many instrument vendors also offer software packages tailored to analyze data from their specific instrument. There are also several third party software packages available for identification of peptide MS/MS spectra. However, most require the use of the vendor-specific software for extraction of the spectral information from the instruments raw data format. Due to the proprietary nature of instruments' data formats, few third party software platforms are capable of accepting raw data. This is problematic because conversion of raw data files to text based peak lists does not conserve all features of the raw data format, and therefore reduces the depth and quality of the downstream analysis. A workaround solution for this problem exists through conversion of raw data formats to more open source data formats such as mzXML (37), mzData, or the newer mzML (37). Converters are available for most major mass spectra raw data formats to translate raw data files to mzXML. Among the mostly used third-party software, PEAKS (38) and MASCOT (39) can accept the raw data formats of most MS manufacturers.

2. Database-Dependent Methods

2.1. Database Searching

The main approaches for identification of peptides from MS/MS spectra can be divided into two broad categories, database dependent and database independent (Fig. 3). Database-dependent methods were the first to be widely adopted for large scale identification of peptide from MS/MS spectra (40–45). The database-dependent identification strategy attempts to match experimental spectra obtained from the MS, with theoretical spectra representing hypothetical peptides generated from

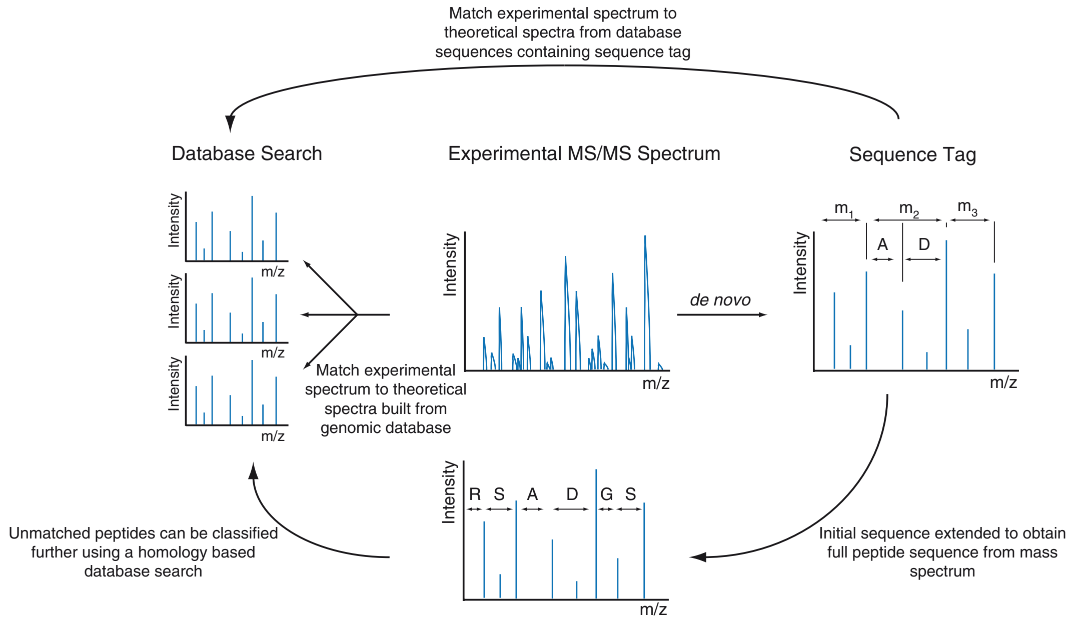

Match experimental spectrum to theoretical spectra from database sequences containing sequence tag
Database Search
Experimental MS/MS Spectrum
Sequence Tag
Intensity
m/z
Intensity
m/z
Intensity
m/z
Match experimental spectrum to theoretical spectra built from genomic database
Intensity
m/z
de novo
m_1
m_2
m_3
A
D
Intensity
m/z
Intensity
R S A D G S
m/z
Unmatched peptides can be classified further using a homology based database search
Initial sequence extended to obtain full peptide sequence from mass spectrum

Fig. 3. Peptide identification techniques

a protein or from genomic database that has been translated. Evidently, these databases are only available for organisms that have genomes that have been sequenced. Theoretical spectra are generated using fragmentation patterns known for specific series of amino acids. The first two widely used search engines using database searching were SEQUEST (40, 41, 46, 47) and MASCOT (39).

SEQUEST and MASCOT both perform the same function but use different scoring schemes to rank peptide matches. SEQUEST makes use of a cross-correlation score to match the hypothetical spectra to the experimental one, whereas MASCOT uses a probability score representing the probability that a spectral match was, or was not, generated by a random event. Each method of matching and scoring has shown advantages and disadvantages for identifying spectra exhibiting different characteristics. Comparisons of database search engines for analyzing different characteristic datasets have been reported (48–50).

There are many other search engines that also use database searching algorithms, such as X!TANDEM (51–53), OMSSA (54), ProbID (55), Phenyx (56), and SONAR (57). All of these search engines are based on database searching, but use different scoring schemes in order to determine the top hit for a peptide match. While database searching has proven to be efficient in its ability to identify peptides, the method does have several drawbacks. False positive identifications, from overly noisy spectra allowing background peaks to be mistaken as peptide peaks, are frequent. Scoring imbalances, where longer, lower quality peptides score higher than higher quality short peptides, are also problematic. Peptide identifications can also be missed if there are unsuspected posttranslational modifications (PTMs) or sequence polymorphisms in the target peptides. The one major limitation of the database-dependent method is that it requires a database for the organism of interest. If an organism has not been sequenced, it is not searchable using these methods.

2.2. Spectral Matching

Another approach relying on a database is spectral matching. In this case, the experimental spectrum is matched against a carefully assembled library of real MS/MS spectra previously obtained. This method has good performance in terms of speed and accuracy. However, no identification will result if the MS/MS spectra have not been previously and reliably identified (43, 58, 59).

2.3. Peptide Sequence Tags

A hybrid method for identifying peptide spectra combining database searching and partial de novo sequencing is called peptide sequence tag matching (PST). The original approach for peptide sequence tagging was first reported by Mann and Wilm (60). PST involves the identification of a short peptide sequence tags, typically 2 or 3 amino acid residues from a region of the MS/MS spectra where this identification can be made with high confidence. Mass regions bordering the sequence tag are extracted as well. Whether or not the sequence tag belongs to the *b*- or *y*-series of ions is usually not known, and this will not be used in scoring. The sequence tag will be used to narrow the list of possible peptides in the database search to only those containing the tag. The sequence tag will then be extended to attempt matching of the mass regions that were extracted along with the tag. Spectral scores are calculated based on the number and summed intensity of matched *b*- and *y*-ions, in combination with whether the C-termini match the specified enzymatic digest. Peptide hits are scored based on a random probability score calculated from matching the mass regions and sequence tag in a similarity search.

PST methods provide a good method for matching MS/MS spectra by combining both spectral matching and de novo methodologies, but are still reliant on database matching. Therefore, if the peptide sequence of interest is not represented in the database, the spectra will not be matched. PST methods also require identification of a sequence tag from every spectrum it wishes to search, which can be difficult for noisy spectra. Also, the generation of sequence tags does not improve the ability to detect peaks of PTMs that can be present in a spectrum. In summary, there are several factors that limit the capabilities of database dependent methods.

3. Database-Independent Methods

3.1. De Novo Sequencing

To overcome the limitations of database-dependent methods, software packages which make use of computerized methods for determining the sequences of peptides directly from the MS/MS spectra have been developed (61). De novo sequencing approaches hold a lot of potential due to their ability to identify previously unknown peptide sequences and post-translational modifications. De novo methods use the knowledge of the fragmentation methods employed in the MS, such as CID, ECD, or ETD, and basic search parameters input by the user, such as enzyme type and peptide modifications, to rebuild the peptide ion series (62–69). Algorithms use scoring schemes that give scores based on identification of the diagnostic ions mentioned before, i.e., the *b*- and *y*-series, *a*-ions, immonium ions, as well as ions from ammonia and water losses. The ability of software to derive a sequence from a MS/MS spectrum is dependent on many factors such as the instrument mass accuracy and resolution as well as spectral quality. Other difficulties like poor peptide fragmentation, peptide ion series directionality, and cleavage abnormalities in spectra are all issues that have to be addressed by de novo algorithms. There are fewer software packages that make use of de novo methods, but this is slowly changing. The more common de novo programs currently available are PEAKS (38), Lutefisk (63), PepNovo (70), and SHERENGA (62). Their efficiency has been evaluated by various groups (71, 72).

3.2. De Novo Sequencing with a Spectrum Graph

While MS/MS spectra can be de novo sequenced and interpreted manually, different types of computer algorithms are used to interpret spectra obtained in high throughput MS/MS. One method is referred to as the spectrum graph approach (73). This approach converts an MS/MS spectrum into a graph first before the actual computation. A graph is an abstract data representation that consists of vertices and edges that connect pairs of vertices. For each peak in a given mass spectrum, two vertices are created (one for the *b*-ion interpretation of the peak and the other for *y*-ion) in the spectrum graph. Each vertex is attributed a mass value according to the peak's *m/z* value. Two vertices from two different peaks are connected with an edge if their mass difference is equal to one of the 20 amino acid residues. Therefore, the ion ladders of the real peptide correspond to a path of connected edges in the spectrum graph. Thus, the de novo sequencing problem is reduced to the finding of a path that connects the N and C terminal vertices in the spectrum graph. Once such a path is found, it can be translated back to the peptide sequence.

In addition, the peak intensity and the existence of specific types of surrounding peaks are used to compute a score value for each vertex in the spectrum graph. Typically, this score function will reward high intensity peaks and the coexistence of peaks with neutral losses (such as the loss of water or ammonium). The score of a path is defined to be the total score of the vertices on the path. When multiple paths exist in the spectrum graph, the software tries to find the one with the highest score and uses that as the de novo sequencing result. Software packages such as SHERENGA and Lutefisk use the spectrum graph approach to score peptide candidates.

SHERENGA is based on the algorithm learning ion types and patterns from training data generated using known peptide sequences. Using the information obtained from the training data, the algorithm attempts to interpret spectra with the creation of a spectrum graph. The algorithm will use the highest scoring sequence path from the spectrum graph as the peptide sequence. Scoring is based on assigning a probability-based score, taking into account rewards/penalties for fragment ions that are present or missing. Lutefisk is based on the algorithm obtaining the longest partial sequence from the graph, where gaps in the sequence are permitted. The algorithm creates a list of significant fragment ions as well as lists of possible N and C-terminal ions. Lutefisk then generates a spectrum graph with *m/z* vs. cleavage site probabilities and uses it to rebuild the highest scoring peptide sequence from the N-terminus. Another program called PepNovo uses an algorithm that is based on peptide fragmentation rules in order to score whether observed peptide peaks are generated randomly or under the predicted fragmentation rules. This allows it to assign scoring rewards to vertices on the spectrum graph, which are representative of the predicted fragmentation pattern. Another algorithm called SeqMS (74) creates a spectrum graph generated from an experimental spectrum using probability scored ion types. Much like the other algorithms presented, the sequence is rebuilt by linking vertices differing by the mass of an amino acid, building from the N-terminus. While all these algorithms make use of different scoring and interpretation schemes, they are based on obtaining the highest scoring peptide sequence from a spectrum graph.

The advantage of the spectrum graph is that by converting an MS/MS spectrum to a graph, some well-known algorithms in graph theory can be utilized to perform the computation. However, the spectrum graph approach encounters difficulties when dealing with data of lesser quality. For example, when more than one adjacent *b*- and *y*-ions are missing from the ion ladders, there will be no path in the spectrum graph that represents the real peptide. Consequently, either no result will be provided, or an incorrect path and a wrong sequence will be deduced. This problem can be partially addressed by adding edges between two vertices with mass difference equal to the sum of more than one residue. But the complexity of the graph will increase and the quality of the result will decline.

In reality, adjacent *b*- and *y*-ions are frequently absent in MS/MS spectra, significantly reducing the usefulness of the spectrum graph methods. In fact, in many of these cases, partially correct sequences can still be identified manually if these spectra were analyzed by a trained researcher. A very different approach was developed in the PEAKS software.

3.3. De Novo Sequencing with the Spectrum Directly

The PEAKS software from Bioinformatics Solutions Inc. is currently the most popular software for de novo sequencing. PEAKS software was shown to have the best accuracy among all currently available de novo sequencing software packages (72).

PEAKS employs a novel algorithm that works on the spectrum directly instead of converting the spectrum to a graph. For each mass value *m*, whether there is a peak or not, PEAKS algorithm assigns two scores $f_y(m)$ and $f_b(m)$, reflecting the likelihood that the real peptide has a *y*-ion or a *b*-ion with mass value m, respectively. If there is a strong intensity peak at mass value *m* (or close to *m* and within the mass error tolerance), then the score is positive. Otherwise, the score is zero or negative. Moreover, the ions with neutral losses that are possibly generated from the *y*- ion or *b*-ion at mass value *m* are also taken into account to compute the two score functions. For a given peptide sequence, the score of the peptide is defined to be the sum of $f_y(m_i)$ and $f_b(m_j)$ for all *y*-ion mass values m_i and *b*-ion mass values m_j. Then PEAKS uses a carefully designed dynamic programming algorithm to compute a peptide sequence whose score is maximized.

In addition to this core algorithm, a few other techniques are employed to improve the accuracy of the de novo sequencing. The overall computational procedure is divided into four steps. First, the raw mass spectrometry data are preprocessed. The preprocessing step includes centroiding, charge recognition and deconvolution, deisotoping and noise filtering. In the second step, the dynamic programming algorithm mentioned above is used to compute the top 10,000 peptide candidates for the spectrum. Thirdly, a more stringent score function is used to evaluate each of the 10,000 candidates. Both immonium ions and the internal cleavage ions are used in this step. The algorithm also attempts to "recalibrate" the data by applying a systematic error to the *m/z* values of all peaks in the spectrum. A stricter mass error tolerance is also used in this step. The candidate sequence that achieves the highest score will be reported as the final output of the algorithm. Finally, a confidence score is computed for the output of the algorithm. In addition, a local confidence score is computed for each amino acid in the output peptide sequence.

This local confidence can be used to determine which portion of the peptide is more reliable than the others. Thus, when a completely correct peptide sequence cannot be obtained, one can still derive a highly confident partial sequence tag from the output of PEAKS.

These sequence tags can then be used by specialized homology search software, such as SPIDER, to identify homologous proteins (75, 76).

The general approach of PEAKS was also adopted by another software package, RAId, for de novo sequencing (77). RAId uses an initial process of identifying candidate peptides that closely resembles the first two steps in PEAKS. The second half of the algorithm uses a database search in an attempt to match the most significant peptides found through de novo to sequences found in the database. If a match is found, the sequence will be reported. However, if a match is not found, the sequence obtained through de novo will be reported.

3.4. De Novo Sequencing with CID and ECD/ETD

In order for a de novo algorithm to determine the full peptide sequence from a spectrum, it requires as much information from backbone fragmentation as possible. Complete peptide ion coverage in a CID experiment is rarely obtained. In a characteristic CID spectrum, the fragment abundance of the fifth most abundant fragment is often ten times less than that of the most abundant. Based on the average length of a peptide generated through a tryptic digest of ~10 residues, the relative abundance of the tenth fragment may be 100 times less intense (78). The end result is poor signal to noise ratio (S/N) for almost half of the ions in a typical CID spectra of a peptide. In addition to low S/N, fragment ions in peptide spectrum frequently overlap, resulting in incorrect peak assignment. Additional problems for CID are the inability to differentiate between isomeric amino acid residues, such as Leu/Ile, and poor suitability for identifying posttranslational modifications.

As mentioned previously, other methods for peptide backbone fragmentation are available on recently available MS instruments. ETD and ECD provide information in peptide spectra that cannot be obtained in CID experiments. The ability to differentiate between Leu/Ile residues as a result of secondary fragmentation, efficient fragmentation of large peptides, and conservation of posttranslational modifications are all possible with ETD/ECD (79–85). One drawback of ETD/ECD is its reduced efficiency with tryptic digests, which creates shorter and mostly doubly charged peptides. However, ETD/ECD works well with larger peptides generated from other proteases, such as Lys-C or Arg-C. Another concern, specifically in the use of de novo sequencing tools, is the propensity for *c*- and *z*-ion masses to vary through hydrogen rearrangement (86).

Recent work from Zubarev et al. illustrates the utility of using spectra obtained using CID with those from ECD/ETD (78). In a de novo sequencing approach, it is advantageous to have a large amount of information in each spectrum being processed. By combining the data obtained through both fragmentation techniques, information contained in one will complement the information present in the other. To this end, Zubarev developed a de novo sequencing algorithm that takes advantage of data obtained from both fragmentation types in order to significantly increase the number and confidence of peptide identifications (87, 88). The algorithm uses a method for creating reliable sequence tags (RSTs). RSTs are built using fragment masses which can be confirmed in both CID/CAD and ECD/ETD spectra. The RST is expanded using masses of increasing uncertainty from both sets of spectra until the full sequence is obtained. The use of a second fragmentation increases confidence in the tag since it can be confirmed by both sets of fragments. PEAKS de novo software analyze the ECD/ETD data files separately.

3.5. Accurate Mass Peptide Sequencing

With the ability of modern MS instrumentation to achieve mass accuracies in the sub-ppm range, algorithms that utilize this level of specificity are very beneficial for protein identification (89–92). As mentioned earlier, the identification of water and ammonia losses are frequently used to increase peptide match scores. However, with low mass accuracy instrumentation, due to the low resolution and mass accuracy, the distinction between the loss of water and ammonia cannot be reliably made. Adding to this problem is the probability for overlapping fragments of other ion peaks within a peptide spectrum. All these issues can compromise the validity of results when using database searching, spectra matching, PST as well as de novo methods. Since de novo experiments are based on the detection of mass differences between fragment peaks, high mass accuracy and resolution are even more critical.

Estimation of the false positive rates showed significant differences when using tighter mass tolerances in the processing of data (87). False positive rates range from 4 to 14% by increasing the allowable mass tolerance from 0.02 to 0.1 Da for fragment ions. If high mass accuracy, along with complementary fragmentation techniques is used to generate RSTs, the results from de novo sequencing was estimated to be >98% correct (87, 88). These results highlight the beneficial effects of quality data for improving output of the de novo algorithms.

3.6. Single Peptide Hits

A problem that affects database searching as well as de novo sequencing approaches is confidence levels in single peptide identifications (93–101). It is common practice to discard all single peptide hits from a set of search results due to the random chance that a single peptide can be assigned to the incorrect protein. However, due to the low abundance of many interesting proteins in complicated matrices, there may be very few detectable peptides for these species in solution. This has created a need to develop methods allowing for preservation of single peptide hits.

The use of complementary fragmentation methods, LC retention time (102), isoelectric point as well as accurate mass increase the confidence in single peptide hits. Another common technique is to use multiple search engines. The "inChorus" function offered in PEAKS allows users to search data files using multiple search engines, combining the results of all methods to create a final results list. By independently identifying peptides in multiple engines, confidence in all hits, including single peptides, is increased, allowing for their inclusion in a dataset. Other statistical methods for peptide and spectrum validation are commonly used to increase confidence in the entire dataset (103–105).

4. Conclusion

With the ever increasing efficiencies of mass spectrometers for proteomics analyses came the need for software to analyze and validate data obtained from these instruments. Initial solutions involved the use of database searching software that was reliant on a collection of genomic sequences within a database. Despite the increasing number of organisms sequenced, there still remain a large number of species for which the genetic code is not known. In addition, database searching methods leave large portions of the MS/MS data uninterpreted due to the absence of corresponding sequences in the database or due to unsuspected posttranslational modifications.

Because of their unbiased assumptions, the approaches using de novo sequencing for the interpretation of MS/MS spectra are rapidly gaining in importance. Efficient de novo software is best achieved with high quality spectra obtained using high resolution MS instrumentation. The combination of ETD or ECD with CID fragmentation, also improve the de novo sequencing approaches. Although more computationally demanding than database search approaches, de novo sequencing will play an increasing role in large scale proteomics experiments. In the near future, a practical strategy may be to use database searching in a first step followed by a de novo method on low scoring peptides as well as on uninterpreted MS/MS spectra. An alternative would be a de novo sequencing step followed by a homology type search to identify mutated or modified peptides.

References

1. Aebersold, R. and M. Mann, (2003). *Mass spectrometry-based proteomics.* Nature, **422**(6928): p. 198–207.
2. Domon, B. and R. Aebersold, (2006). *Mass spectrometry and protein analysis.* Science, **312**(5771): p. 212–7.
3. Wisniewski, J.R., (2008). *Mass spectrometry-based proteomics: principles, perspectives, and challenges.* Arch Pathol Lab Med, **132**(10): p. 1566–9.
4. Reinders, J., et al., (2004). *Challenges in mass spectrometry-based proteomics.* Proteomics, **4**(12): p. 3686–703.

5. Fenn, J.B., et al., (1989). *Electrospray ionization for mass spectrometry of large biomolecules.* Science, **246**(4926): p. 64–71.

6. Karas, M. and F. Hillenkamp, (1988). *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Anal Chem, **60**(20): p. 2299–301.

7. Whitehouse, C.M., et al., (1985). *Electrospray interface for liquid chromatographs and mass spectrometers.* Anal Chem, **57**(3): p. 675–9.

8. Chernushevich, I.V., A.V. Loboda, and B.A. Thomson, (2001). *An introduction to quadrupole-time-of-flight mass spectrometry.* J Mass Spectrom, **36**(8): p. 849–65.

9. Roepstorff, P., (2000). *MALDI-TOF mass spectrometry in protein chemistry.* EXS, **88**: p. 81–97.

10. Yost, R.A., and R.K. Boyd, (1990). *Tandem mass-spectrometry – quadrupole and hybrid instruments.* Methods Enzymol, **193**: p. 154–200.

11. Peterman, S.M., C.P. Dufresne, and S. Horning, (2005). *The use of a hybrid linear trap/FT-ICR mass spectrometer for on-line high resolution/high mass accuracy bottom-up sequencing.* J Biomol Tech, **16**(2): p. 112–24.

12. Gizzi, G., et al., (2005). *Determination of dioxins (PCDDs/PCDFs) and PCBs in food and feed using the DR CALUX (R) bioassay: results of an international validation study.* Food Addit Contam, **22**(5): p. 472–81.

13. Delahunty, C. and J.R. Yates, (2005). *Protein identification using 2D-LC-MS/MS.* Methods, **35**(3): p. 248–55.

14. Steen, H. and M. Mann, (2004). *The ABC's (and XYZ's) of peptide sequencing.* Nat Rev Mol Cell Biol, **5**(9): p. 699–711.

15. Hayes, R.N. and M.L. Gross, (1990). *Collision-Induced Dissociation.* Methods Enzymol, **193**: p. 237–63.

16. McLuckey, S.A., D.E. Goeringer, and G.L. Glish, (1992). *Collisional activation with random noise in ion trap mass spectrometry.* Anal Chem, **64**(13): p. 1455–60.

17. Morris, H.R., et al., (1996). *High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer.* Rapid Commun Mass Spectrom, **10**(8): p. 889–96.

18. Savitski, M.M., et al., (2007). *Relative specificities of water and ammonia losses from backbone fragments in collision-activated dissociation.* J Proteome Res, **6**(7): p. 2669–73.

19. Biemann, K., (1988). *Contributions of mass spectrometry to peptide and protein structure.* Biomed Environ Mass Spectrom, **16**(1–12): p. 99–111.

20. Breci, L.A., et al., (2003). *Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra.* Anal Chem, **75**(9): p. 1963–71.

21. Huang, Y., et al., (2005). *Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns.* Anal Chem, 77(18): p. 5800–13.

22. Mann, M., C.K. Meng, and J.B. Fenn, (1989). *Interpreting mass-spectra of multiply charged ions.* Anal Chem, **61**(15): p. 1702–08.

23. Paizs, B. and S. Suhai, (2004). *Towards understanding the tandem mass spectra of protonated oligopeptides. 1: mechanism of amide bond cleavage.* J Am Soc Mass Spectrom, **15**(1): p. 103–13.

24. Paizs, B. and S. Suhai, (2005). *Fragmentation pathways of protonated peptides.* Mass Spectrom Rev, **24**(4): p. 508–48.

25. Roepstorff, P. and J. Fohlman, (1984). *Proposal for a common nomenclature for sequence ions in mass spectra of peptides.* Biomed Mass Spectrom, **11**(11): p. 601.

26. Tabb, D.L., et al., (2003). *Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides.* Anal Chem, **75**(5): p. 1155–63.

27. Wysocki, V.H., et al., (2000). *Mobile and localized protons: a framework for understanding peptide dissociation.* J Mass Spectrom, **35**(12): p. 1399–406.

28. Zhang, Z., (2004). *Prediction of low-energy collision-induced dissociation spectra of peptides.* Anal Chem, **76**(14): p. 3908–22.

29. Zhong, H., et al., (2004). *Protein sequencing by mass analysis of polypeptide ladders after controlled protein hydrolysis.* Nat Biotechnol, **22**(10): p. 1291–6.

30. Appella, E. and C.W. Anderson, (2007). *New prospects for proteomics – electron-capture (ECD) and electron-transfer dissociation (ETD) fragmentation techniques and combined fractional diagonal chromatography (COFRADIC).* FEBS J, **274**(24): p. 6255.

31. Zubarev, R.A., (2004). *Electron-capture dissociation tandem mass spectrometry.* Curr Opin Biotechnol, **15**(1): p. 12–6.

32. Zubarev, R.A., et al., (2000). *Electron capture dissociation for structural characterization of multiply charged protein cations.* Anal Chem, **72**(3): p. 563–73.

33. Lin, D., D.L. Tabb, and J.R. Yates, (2003). *Large-scale protein identification using mass spectrometry.* Biochim Biophys Acta, **1646** (1–2): p. 1–10.

34. MacCoss, M.J., (2005). *Computational analysis of shotgun proteomics data*. Curr Opin Chem Biol, **9**(1): p. 88–94.

35. Johnson, R.S., et al., (2005). *Informatics for protein identification by mass spectrometry*. Methods, **35**(3): p. 223–36.

36. Forner, F., L.J. Foster, and S. Toppo, (2007). *Mass spectrometry data analysis in the proteomics era*. Curr Bioinform, **2**(1): p. 63–93.

37. Pedrioli, P.G., et al., (2004). *A common open representation of mass spectrometry data and its application to proteomics research*. Nat Biotechnol, **22**(11): p. 1459–66.

38. Ma, B., et al., (2003). *PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry*. Rapid Commun Mass Spectrom, **17**(20): p. 2337–42.

39. Perkins, D.N., et al., (1999). *Probability-based protein identification by searching sequence databases using mass spectrometry data*. Electrophoresis, **20**(18): p. 3551–67.

40. Eng, J.K., A.L. Mccormack, and J.R. Yates, (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**(11): p. 976–89.

41. Yates, J.R., III, J.K. Eng, and A.L. McCormack, (1995). *Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases*. Anal Chem, **67**(18): p. 3202–10.

42. Craig, R., J.P. Cortens, and R.C. Beavis, (2005). *The use of proteotypic peptide libraries for protein identification*. Rapid Commun Mass Spectrom, **19**(13): p. 1844–50.

43. Frewen, B.E., et al., (2006). *Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries*. Anal Chem, **78**(16): p. 5678–84.

44. Sadygov, R.G., D. Cociorva, and J.R. Yates, III, (2004). *Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book*. Nat Methods, **1**(3): p. 195–202.

45. Sadygov, R.G., H.B. Liu, and J.R. Yates, (2004). *Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases*. Anal Chem, **76**(6): p. 1664–71.

46. Yates, J.R., III, et al., (1995). *Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database*. Anal Chem, **67**(8): p. 1426–36.

47. Eng, J.K., et al., (2008). *A fast SEQUEST cross correlation algorithm*. J Proteome Res, 7(10): p. 4598–602.

48. Elias, J.E., et al., (2005). *Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations*. Nat Methods, **2**(9): p. 667–75.

49. Stein, S.E. and D.R. Scott, (1994). *Optimization and testing of mass-spectral library search algorithms for compound identification*. J Am Soc Mass Spectrom, **5**(9): p. 859–66.

50. Kapp, E.A., et al., (2005). *An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis*. Proteomics, **5**(13): p. 3475–90.

51. Craig, R. and R.C. Beavis, (2004). *TANDEM: matching proteins with tandem mass spectra*. Bioinformatics, **20**(9): p. 1466–7.

52. Craig, R. and R.C. Beavis, (2003). *A method for reducing the time required to match protein sequences with tandem mass spectra*. Rapid Commun Mass Spectrom, **17**(20): p. 2310–6.

53. Craig, R., J.P. Cortens, and R.C. Beavis, (2004). *Open source system for analyzing, validating, and storing protein identification data*. J Proteome Res, **3**(6): p. 1234–42.

54. Geer, L.Y., et al., (2004). *Open mass spectrometry search algorithm*. J Proteome Res, **3**(5): p. 958–64.

55. Zhang, N., R. Aebersold, and B. Schwilkowski, (2002). *ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data*. Proteomics, **2**(10): p. 1406–12.

56. Colinge, J., et al., (2003). *OLAV: towards high-throughput tandem mass spectrometry data identification*. Proteomics, **3**(8): p. 1454–63.

57. Field, H.I., D. Fenyo, and R.C. Beavis, (2002). *RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database*. Proteomics, **2**(1): p. 36–47.

58. Nesvizhskii, A.I., O. Vitek, and R. Aebersold, (2007). *Analysis and validation of proteomic data generated by tandem mass spectrometry*. Nat Methods, **4**(10): p. 787–97.

59. Lam, H., et al., (2007). *Development and validation of a spectral library searching method for peptide identification from MS/MS*. Proteomics, 7(5): p. 655–67.

60. Mann, M. and M. Wilm, (1994). *Error-tolerant identification of peptides in sequence databases by peptide sequence tags*. Anal Chem, **66**(24): p. 4390–9.

61. Xu, C. and B. Ma, (2006). *Software for computational peptide identification from MS-MS data*. Drug Discov Today, **11**(13–14): p. 595–600.

62. Dancik, V., et al., (1999). *De novo peptide sequencing via tandem mass spectrometry*. J Comput Biol, **6**(3–4): p. 327–42.

63. Johnson, R.S. and J.A. Taylor, (2002). *Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry.* Mol Biotechnol, **22**(3): p. 301–15.
64. Taylor, J.A. and R.S. Johnson, (1997). *Sequence database searches via de novo peptide sequencing by tandem mass spectrometry.* Rapid Commun Mass Spectrom, **11**(9): p. 1067–75.
65. Taylor, J.A. and R.S. Johnson, (2001). *Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry.* Anal Chem, **73**(11): p. 2594–604.
66. Lu, B. and T. Chen, (2003). *A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry.* J Comput Biol, **10**(1): p. 1–12.
67. Samgina, T.Y., et al., (2008). *De novo sequencing of peptides secreted by the skin glands of the Caucasian Green Frog Rana ridibunda.* Rapid Commun Mass Spectrom, **22**(22): p. 3517–25.
68. Standing, K.G., (2003). *Peptide and protein de novo sequencing by mass spectrometry.* Curr Opin Struct Biol, **13**(5): p. 595–601.
69. Xu, C. and B. Ma, (2006). *Complexity and scoring function of MS/MS peptide de novo sequencing.* Comput Syst Bioinformatics Conf: p. 361–9.
70. Frank, A. and P. Pevzner, (2005). *PepNovo: de novo peptide sequencing via probabilistic network modeling.* Anal Chem, **77**(4): p. 964–73.
71. Pitzer, E., A. Masselot, and J. Colinge, (2007). *Assessing peptide de novo sequencing algorithms performance on large and diverse data sets.* Proteomics, **7**(17): p. 3051–4.
72. Pevtsov, S., et al., (2006). *Performance evaluation of existing de novo sequencing algorithms.* J Proteome Res, **5**(11): p. 3018–28.
73. Chen, T., et al., (2001). *A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry.* J Comput Biol, **8**(3): p. 325–37.
74. Fernandez-de-Cossio, J., et al., (1998). *Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by 'SeqMS', a software aid for de novo sequencing by tandem mass spectrometry.* Rapid Commun Mass Spectrom, **12**(23): p. 1867–78.
75. Han, Y., B. Ma, and K. Zhang, (2004). *SPIDER: software for protein identification from sequence tags with de novo sequencing error.* Proc IEEE Comput Syst Bioinform Conf: p. 206–15.
76. Han, Y., B. Ma, and K. Zhang, (2005). *SPIDER: software for protein identification from sequence tags with de novo sequencing error.* J Bioinform Comput Biol, **3**(3): p. 697–716.
77. Alves, G. and Y.K. Yu, (2005). *Robust accurate identification of peptides (RAId): deciphering MS2 data using a structured library search with de novo based statistics.* Bioinformatics, **21**(19): p. 3726–32.
78. Zubarev, R.A., A.R. Zubarev, and M.M. Savitski, (2008). *Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet?* J Am Soc Mass Spectrom, **19**(6): p. 753–61.
79. Kjeldsen, F., et al., (2003). *Distinguishing of Ile/Leu amino acid residues in the PP3 protein by (hot) electron capture dissociation in Fourier transform ion cyclotron resonance mass spectrometry.* Anal Chem, **75**(6): p. 1267–74.
80. Cantin, G.T. and J.R. Yates, III, (2004). *Strategies for shotgun identification of post-translational modifications by mass spectrometry.* J Chromatogr A, **1053**(1–2): p. 7–14.
81. Kelleher, N.L., et al., (1999). *Localization of labile posttranslational modifications by electron capture dissociation: the case of gamma-carboxyglutamic acid.* Anal Chem, **71**(19): p. 4250–3.
82. Nielsen, M.L., M.M. Savitski, and R.A. Zubarev, (2005). *Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry.* Mol Cell Proteomics, **4**(6): p. 835–45.
83. Savitski, M.M., M.L. Nielsen, and R.A. Zubarev, (2007). *Side-chain losses in electron capture dissociation to improve peptide identification.* Anal Chem, **79**(6): p. 2296–302.
84. Silivra, O.A., et al., (2005). *Electron capture dissociation of polypeptides in a three-dimensional quadrupole ion trap: Implementation and first results.* J Am Soc Mass Spectrom, **16**(1): p. 22–7.
85. Reinders, J. and A. Sickmann, (2005). *State-of-the-art in phosphoproteomics.* Proteomics, **5**(16): p. 4052–61.
86. Savitski, M.M., et al., (2007). *Hydrogen rearrangement to and from radical z fragments in electron capture dissociation of peptides.* J Am Soc Mass Spectrom, **18**(1): p. 113–20.
87. Savitski, M.M., et al., (2005). *Proteomics-grade de novo sequencing approach.* J Proteome Res, **4**(6): p. 2348–54.
88. Savitski, M.M., M.L. Nielsen, and R.A. Zubarev, (2005). *New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques.* Mol Cell Proteomics, **4**(8): p. 1180–8.

89. Clauser, K.R., P. Baker, and A.L. Burlingame, (1999). *Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching*. Anal Chem, **71**(14): p. 2871–82.

90. Falth, M., et al., (2007). *SwedCAD, a database of annotated high-mass accuracy MS/MS spectra of tryptic peptides*. J Proteome Res, **6**(10): p. 4063–7.

91. Zubarev, R. and M. Mann, (2007). *On the proper use of mass accuracy in proteomics*. Mol Cell Proteomics, **6**(3): p. 377–81.

92. Frank, A.M., et al., (2007). *De novo peptide sequencing and identification with precision mass spectrometry*. J Proteome Res, **6**(1): p. 114–23.

93. Benjamini, Y. and Y. Hochberg, (1995). *Controlling the false discovery rate – a practical and powerful approach to multiple testing*. J R Stat Soc Series B Stat Methodol, **57**(1): p. 289–300.

94. Balgley, B.M., et al., (2007). *Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy*. Mol Cell Proteomics, **6**(9): p. 1599–608.

95. Choi, H., D. Ghosh, and A.I. Nesvizhskii, (2008). *Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling*. J Proteome Res, 7(1): p. 286–92.

96. Elias, J.E. and S.P. Gygi, (2007). *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry*. Nat Methods, **4**(3): p. 207–14.

97. States, D.J., et al., (2006). *Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study*. Nat Biotechnol, **24**(3): p. 333–8.

98. Peng, J., et al., (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**(1): p. 43–50.

99. Beer, I., et al., (2004). *Improving large-scale proteomics by clustering of mass spectrometry data*. Proteomics, **4**(4): p. 950–60.

100. Carr, S., et al., (2004). *The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data*. Mol Cell Proteomics, **3**(6): p. 531–3.

101. Nesvizhskii, A.I. and R. Aebersold, (2005). *Interpretation of shotgun proteomic data: the protein inference problem*. Mol Cell Proteomics, **4**(10): p. 1419–40.

102. Klammer, A.A., et al., (2007). *Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions*. Anal Chem, **79**(16): p. 6111–8.

103. Fenyo, D. and R.C. Beavis, (2003). *A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes*. Anal Chem, **75**(4): p. 768–74.

104. MacLean, B., et al., (2006). *General framework for developing and evaluating database scoring algorithms using the TANDEM search engine*. Bioinformatics, **22**(22): p. 2830–2.

105. Nesvizhskii, A.I., et al., (2003). *A statistical model for identifying proteins by tandem mass spectrometry*. Anal Chem, **75**(17): p. 4646–58.

Chapter 9

Cross Species Proteomics

J.C. Wright, R.J. Beynon, and S.J. Hubbard

Abstract

Proteomics has advanced in leaps and bounds over the past couple of decades. However, the continuing dependency of mass spectrometry-based protein identification on the searching of spectra against protein sequence databases limits many proteomics experiments. If there is no sequenced genome for a given species, then cross species proteomics is required, attempting to identify proteins across the species boundary, typically using the sequenced genome of a closely related species. Unlike sequence searching for homologues, the proteomics equivalent is confounded by small differences in amino acid sequences, leading to large differences in peptide masses; this renders mass matching of peptides and their product ions difficult. Therefore, the phylogenetic distance between the two species and the attendant level of conservation between the homologous proteins play a huge part in determining the extent of protein identification that is possible across the species boundary. In this chapter, we review the cross species challenge itself, as well as various approaches taken to deal with it and the success met with in past studies. This is followed by recommendations of best practice and suggestions to researchers facing this challenge as well as a final section predicting developments, which may help improve cross species proteomics in the future.

Key words: Proteomics, Mass spectrometry, Cross species, PMF, Tandem MS, De novo sequencing

Abbreviations

PMF	Peptide mass fingerprint
ESI	Electrospray ionisation
MALDI	Matrix assisted laser desorption/ionisation
MS	Mass spectrometry
m/z	Mass to charge
PTM	Post translational modification

1. Introduction

Cross species proteomics has been a problem since mass spectrometry was first applied to identify proteins and continues to be troublesome even today. *De novo* based methods notwithstanding,

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_9, © Humana Press, a part of Springer Science+Business Media, LLC 2010

nearly all protein identification methods from mass spectra require a sequence database to generate theoretical peptide spectra and compare it to the experimental ones. This is all very well for model species studied regularly in the biological sciences as they have well-annotated and fully-sequenced genomes (and hence proteomes). The problem arises when working with less well-studied species that have not had their genomes sequenced and perhaps are unlikely to in the near future. Our capability to sequence genomes is advancing rapidly with the recent developments in so-called “next generation” sequencing methods such as “Pyrosequencing”, “Polymerase-based sequencing-by-synthesis” and “Ligation-based sequencing” (1). Although this technology makes it easier and faster to sequence new species and will no doubt lead to a huge increase in the number of sequenced genomes over the next few years, there is still a limit to these methods and the number of different species around the world is estimated to be in the millions (2). Moreover, much of the new sequencing effort is dedicated to re-sequencing further examples of extant species (i.e. 1,000 human genomes) to characterise polymorphisms, strains and populations. Besides there being a limit to the number of species sequenced, there is also the problem of annotation. A raw genome is not necessarily of any use for making protein identifications, and the full annotation of a genome can be a slow process of gene prediction and functional identification although proteomics can also help in the annotation process (3). Therefore, the demand for conducting cross species proteomics experiments using unsequenced species is one that is unlikely to go away anytime soon as proteomics scientists will continue to want to investigate proteins involved in unique and exciting phenotypes in species with no currently sequenced genome. The cross species problem is not just limited to making protein identifications across the species boundaries; sometimes, even trying to match proteins using a database populated with proteins from a different strain of the same species can be difficult since single polymorphic variations (i.e. a single base change) can radically alter properties at the peptide level.

The cross species problem is a simple one to understand but much more complex to solve. It comes down to the fact that a single amino acid change in a peptide can change the mass of that peptide beyond the search tolerance of any proteomics experiment. For example, a single substitution of a glycine residue for a tryptophan residue leads to a mass difference of 129.06 Daltons. Given that the standard search tolerance in a proteomics experiment is kept below 1 Dalton, this is a massive relative shift in mass. This principle is further demonstrated in Fig. 1 showing two homologous sequences which are easily matched using sequence comparison and alignment tools as they have a sequence identity of 90%. However, in a peptide matching context, the few

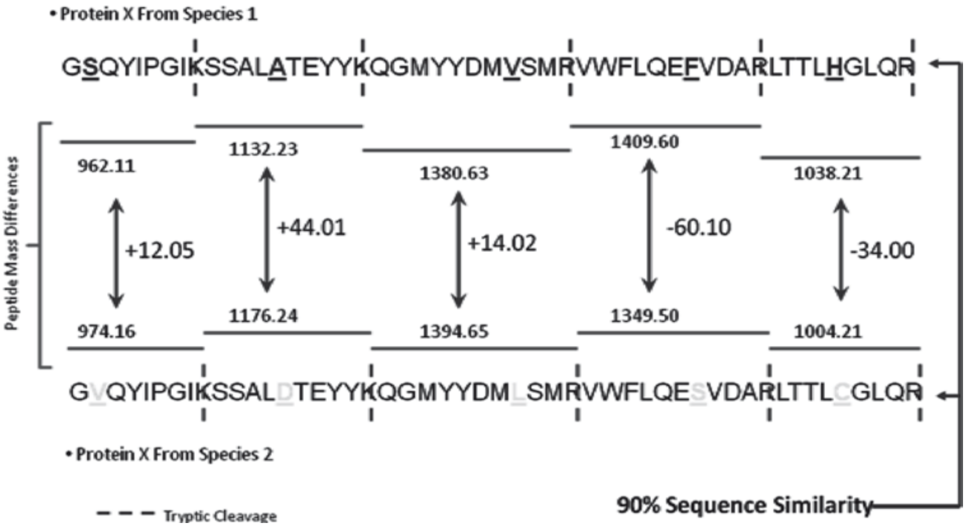


Fig. 1. The cross species proteomics problem – these two homologous sequences represent a protein of the same function in two different species. These two species are closely related on an evolutionary scale and these proteins have 90% sequence similarity. However, because of the positioning of the amino acid difference, when we theoretically digest the two protein sequences using tryptic specificity rules, we find that the peptides are much less conserved. Each peptide represented in the upper part of the figure has had its molecular mass calculated and then compared to its equivalent peptide in the lower protein. The numbers next to the vertical arrows in the figure show the mass difference between each peptide of the two proteins. Considering that most mass spectrometry database searches use a mass tolerance of less than 1 Dalton, these differences in peptide mass would mean that not a single peptide would match from one species to the other

amino acid differences between the two sequences shift the mass of each peptide beyond the tolerance of a standard proteomics database search. These slight changes in peptide sequence may even change the chemistry of the peptide enough to alter its "detectability" in the mass spectrometer so that it may not be seen in the mass spectra at all. Several studies have looked at the level of homology required for cross species identification to be possible; these studies found that at a minimum, 70% sequence identity is required (4, 5), and as Fig. 1 demonstrates, a match is still not guaranteed even at 90% identity. On top of the difference in amino acid composition, different species-specific post translational modifications can also hinder cross species identification.

Since the problem is not new, approaches have appeared in the literature since the early 1990s. One of the first published cross species peptide mass fingerprint experiments was the analysis of *Spiroplasma melliferum* proteins from 2D gels (6, 7). This study managed to identify several proteins from the unsequenced *S. melliferum* species by combining the PMF data with other data collected on the sample proteins. This included amino acid composition data, overall estimated intact protein mass and pI estimates. They demonstrated that PMF alone is not capable of making cross species identifications, but when combined with sequence composition data, the number of identifications made significantly improves. This method has been applied to other species such as *Mycoplasma genitalium* (8) with some success. In 1997, Wilkins and Williams (4) conducted a study exploring the degree of conservation required for cross species identification using the combined PMF and amino acid composition method. This study revealed that although amino acid composition was quite well conserved across species, often at an even higher level than actual sequence identity, the conservation of tryptic peptides was very poor finding that below 70% sequence identity the number of conserved peptides in the 700–3,000 Dalton range dropped to virtually zero. The conclusions of this study suggest that PMF can be used to identify proteins across species boundaries but only if the phylogenetic distances between the database and the sample are very small with proteins generally having a sequence identity to their database counterpart of 80% or above and that more than just the spectra is used to search the database. This and other studies at the time (9, 10) conclude that the best way to make cross species identifications is the incorporation of other data such as intact protein mass, isoelectric point and amino acid composition. This lead to the development of MultiIdent (http://expasy.org/tools/multiident/) a web based tool for protein identification using a combination of protein data, peptide composition and PMF (11). Cordwell *et al.* developed this further (12) using the knowledge that highly conserved regions of the protein will have conserved peptide masses and will likely contain a characteristic motif for a particular protein function or domain; potentially highly useful information. However, this technique was found to be limited in scope, as to make the functional motif identifications one has to match very rare highly conserved peptides above a certain size.

In 2000, a study looking at the cross species PMF identification of marsupial proteins, without using amino acid composition information, used tandem MS and *de novo* sequencing as a method of confirming the PMF matches (13). This study successfully identified almost 50% of the analysed proteins via PMF confirming their matches using MS/MS. However, these proteins all had a similarity of over 89% to other mammalian species, and a further theoretical analysis in the study showed that on average, an identity of over 74.6% would have been needed to match marsupial proteins to other mammalian species. They also found that these highly conserved proteins were generally cytoskeletal and "Housekeeping" proteins, suggesting the approach would not be generally applicable genome wide. Another study in 2001 looking at bacterial outer membrane proteins also used PMF without any amino acid data to make cross species protein identifications (14). This study also highlighted the effect of evolutionary divergence on cross species matching, showing a very sharp drop off in the number of bacterial proteins that could be identified against an *E. coli* database as the two species became more divergent.

Although not yet reliable enough for large scale high-throughput experiments, *de novo* sequencing has become a standard method for confirming cross species PMF identifications. This in turn motivated the development of MSBLAST (15), a version of BLAST (16) optimised for peptide sequence tag homology searching. The approach infers candidate peptide sequences based on limited MS data and attempts to reconcile a large set of BLAST hits to find the most likely homologous protein with some success using canine example proteins. Again, the approach is not fully generalisable since one must obtain candidate sequence tags for a large number of peptides to be effective and this is not always possible.

In 2002, Lester and Hubbard conducted a detailed informatics analysis of the cross species proteomics problem, comparing over 30 complete genomes and finding that 70% sequence identity is the minimum required to make an identification, and added further support that PMF remains a useful technique for cross species proteomics, especially when data other than just the peak masses are used in the database search (5). Liska and Shevchenko reviewed the field further in 2003 (17) and described standard experimental workflows for making cross species identifications in proteomics (18, 19). The methods described in this review build on those workflows, using PMF initially to quickly identify well conserved proteins and then moving on to tandem MS/MS methods and automated *de novo* sequencing (20–23), database searching with MSBLAST (15) or one of the other *de novo* identification tools available (24–27). This three staged workflow has been used as the standard set up for conducting cross species proteomics in several recent studies (28–32). Although not discussed in detail in this review, some studies have used high accuracy FT-ICR MS instruments to conduct cross species analysis and perform de novo sequencing (33), but this is a rather specialised approach not available in the majority of labs.

Although a species may not have a sequenced genome, there are quite often alternatives available from samples of the transcriptome (e.g. cDNA and EST data) or a partial or unassembled genome sequence. Recent approaches have exploited this data too to make protein identifications (34–43). Consequently, several tools have been developed to take advantage of and optimise searches against these kinds of databases (44, 45).

Proteomics in general is moving from discovery experiments and plain protein identification towards more targeted and quantitative experiments. Hence, there have also been studies attempting to conduct quantitative proteomics across the species boundary. One recent study has examined the effectiveness of using shared peptides to quantitatively measure relative protein abundances in unsequenced species (46, 47). This showed that it is possible but tricky and requires a good level of homology between the proteins of interest and their database counterparts.

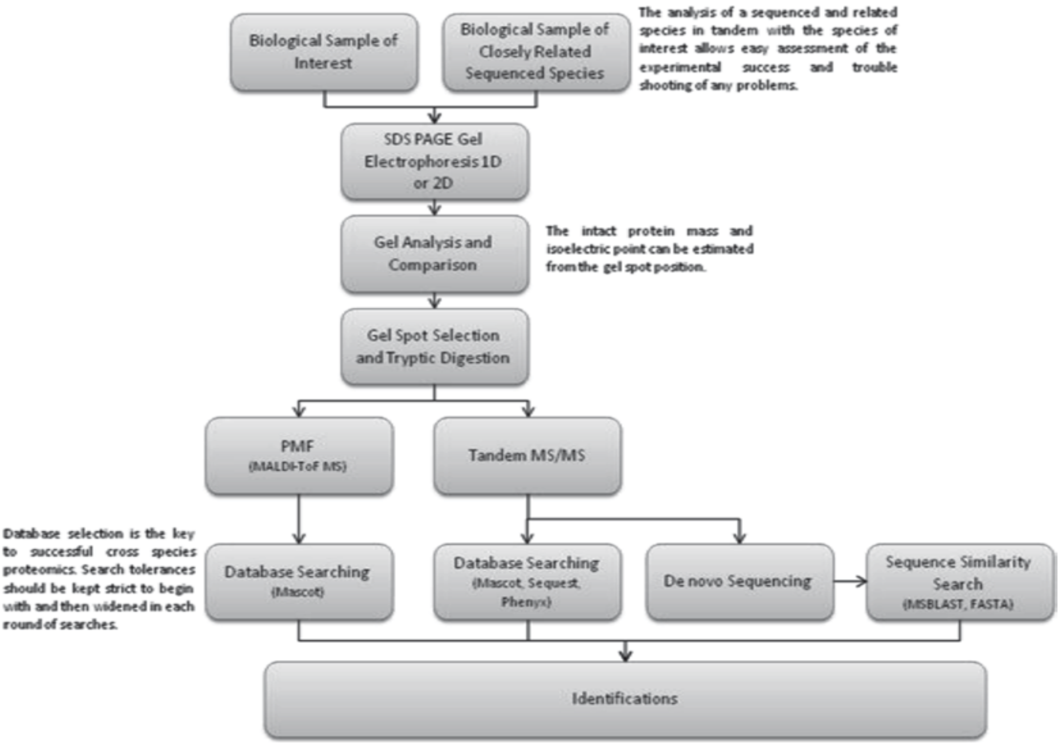
2. Methods for Cross Species Proteomics

The availability of a closely related sequenced genome should be the first consideration of any cross species study. It is worth investigating if any published sequence data on the species to be studied is available (e.g. via a search through UniProt http://www.uniprot.org/ or Genbank http://www.ncbi.nlm.nih.gov/Genbank/). EST and cDNA databases may also be available and can be very useful and are frequently found for public download from the internet (http://www.ncbi.nlm.nih.gov/dbEST/). It is also worth checking that your species is not about to be sequenced at one of the major centres (e.g. Sanger http://www.sanger.ac.uk/Projects/, JGI http://www.jgi.doe.gov/, WashU http://genome.wustl.edu/, TIGR http://www.tigr.org/, etc.).

These kinds of studies across the species boundary can also benefit from being done in parallel with a closely related species if a sample is available. Although it doubles the work, this second experiment provides a control for comparisons and makes it easy to trouble shoot problems in the experiment. Additionally, one might expect any detected post translation modifications to be in common between the proteins being studied.

A generalised approach to cross species proteomics is shown in Fig. 2 showing an increase in the sophistication of the mass spectrometry used from PMF (48) to tandem MS/MS (49, 50) to *de novo* sequencing (51) until each protein is identified. This increase in mass spectrometric complexity usually also implies a reduction in throughput at each stage. Although recent advancements have significantly improved tandem MS/MS throughput, many labs still use PMF on a MALDI-ToF platform as a quick screening method for matching highly conserved proteins and establishing the general quality of the sample. It should also be noted that PMF methods do not separate the tryptic peptides in an LC step. This step effectively removes the peptide-protein connectivity which is retained when analysing protein spots from gels. The tandem MS/MS practitioners have to solve the attendant protein inference problem which PMF does not need to do – this could be a significant advantage in a cross species context.

As mentioned earlier, the key to getting good identifications is to gather as much data on each protein as possible to combine in the database search. Therefore, it is best to initially separate proteins in the sample by two dimensional electrophoresis (52, 53) and then cut out spots for which an estimated isoelectric point


Biological Sample of Interest
Biological Sample of Closely Related Sequenced Species
SDS PAGE Gel Electrophoresis 1D or 2D
Gel Analysis and Comparison
Gel Spot Selection and Tryptic Digestion
PMF (MALDI-ToF MS)
Tandem MS/MS
Database Searching (Mascot)
Database Searching (Mascot, Sequest, Phenyx)
De novo Sequencing
Sequence Similarity Search (MSBLAST, FASTA)
Identifications

The analysis of a sequenced and related species in tandem with the species of interest allows easy assessment of the experimental success and trouble shooting of any problems.

The intact protein mass and isoelectric point can be estimated from the gel spot position.

Database selection is the key to successful cross species proteomics. Search tolerances should be kept strict to begin with and then widened in each round of searches.

Fig. 2. A cross species proteomics workflow and experimental setup – this figure displays a simple workflow for conducting cross species proteomics, initially conducting a gel electrophoresis separation of proteins and then analysing the gel spots using mass spectrometry. Peptide mass fingerprinting works well as a fast initial screening method for identifying well conserved housekeeping and structural proteins. The more difficult proteins which are likely to be those of more interest can then be analysed using tandem MS/MS and from the resulting spectra either be subjected to further non error tolerant database searching or even *de novo* sequencing and sequence similarity searches using short peptide sequence tags

and intact protein mass can be obtained. If this is done in tandem with a related species sequence, many of the protein identities can be estimated through comparison of the gels – although this would not be sufficiently reliable on its own. There is a whole range of software available for this kind of gel analysis (54). Another important concept to keep in mind for this kind of experiment is repetition; having an abundance of sample and running multiple experiments will help you go deeper into the unknown proteome by gathering more high quality spectra.

2.1. Peptide Mass Fingerprinting

The PMF analysis itself should be performed by following the normal best practice and standard protocols for the instrument in question. Once good quality spectra have been obtained, the difficult task of making identifications is faced. Here is where gathering as much information on the protein being analysed and repeated spectra pays off.

1. It is best to begin with a simple database search using a common search tool such as Mascot (55) considering only one or two miss cleavages at a very low mass tolerance and with few modifications considered.
2. If no candidate identifications are made, then add in other data and slowly widen the search parameters repeating the search. A lot of this repeated database searching can be time consuming depending on the database size and computational power available, but it can all be easily automated using simple Perl scripts or distribution tools such as Mascot daemon (www.matrixscience.com/daemon.html).
3. Using a range of different search tools can help boost the number of identifications and increase confidence in the identifications.
4. After identifying as many spectra as possible using PMF, it is worth examining the unidentified spectra and finding those that are of good quality with unmatched peptides and putting those samples forward for tandem MS/MS analysis. Quality can be assessed qualitatively with experience by visualisation. Hopefully, by this point, at least the highly conserved house-keeping and structural proteins in the proteome have been identified, leaving the more "interesting" and less well conserved proteins for further analysis.

2.2. Tandem MS/MS and De Novo Sequencing

Many of the same principles as applied to cross species PMF can be applied when conducting a tandem MS/MS experiment. Standard protocols and best practice for the instruments being used should be followed in order to achieve good quality spectra.

1. If the experiment is being done in tandem with a known sequenced species, it should be clear how good the quality is of the spectra being obtained, by direct comparison of spectra and identifications from the known sequence. This presumes that related proteins from the two species will co-migrate on gels.
2. A faster and more high throughput method would be to skip the gel separation and conduct a MudPIT or shotgun gel-free (56) experiment, digesting the entire proteome and then separating peptides using liquid chromatography. As mentioned, this creates an additional problem where peptides from many different proteins are all mixed up. There are many methods available to resolve the protein inference problem when working with a sequenced species, but for cross species, it is much harder and therefore easier and more accurate to use gel separations. The gel electrophoresis separation of proteins, as mentioned earlier, also contributes extra information that can assist the identification of proteins.

3. It is recommended that to maximise the number of true peptide identifications, multiple search algorithms should be used (57), combined with some variety of decoy database searching (58) to help establish a false discovery rate. This should be followed by further analysis using protein validation software (59) such as that available in the Trans Proteomic Pipeline suite (http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP).
4. Even after this kind of intense MS/MS analysis, it is quite likely that there will be proteins still left unidentified. At this stage, the application of an automated *de novo* sequencing algorithm to generate peptide sequence tags (60), or failing that the more accurate manual sequencing of the spectra, is currently the only way forward. This generates peptide sequence tags which can then be searched in much more error tolerant way against a sequence database. A selection of bioinformatics tools for conducting sequence similarity searches from *de novo* sequenced peptides are available (15, 22–27). Equally, a simple approach of searching candidate peptide tags against a sequence database using BLAST could also be informative if the sufficient sequence has been obtained.

2.3. Using ESTs and Partial Sequence Databases

The *de novo* sequencing of peptides from tandem MS/MS spectra and subsequent searching of peptide sequence tags (PSTs) against a sequence database is a powerful way to make cross species identifications, certainly compared to the more limited capabilities of PMF. However, *de novo* sequencing of peptides can be very time consuming and, despite the advances made with automated sequencing tools, the process is most successfully done manually. Although many species do not have a fully sequenced genome, publically available sequence data in the form of cDNA and expressed sequence tag EST databases are often available. Indeed, since these libraries are usually targeted at a particular tissue or developmental stage, they can be enriched in the genes of interest to a particularly research community. Various studies have compared database searches via EST databases with using a closely related species database (34–43) and most have achieved a higher level of identification, despite inherent sequence inaccuracies in the EST data sets. ESTs are usually single pass sequence tags which are prone to errors such as base changes, frameshifting indels, chimeric splicing with other cDNAs and simple sequencing errors. However, they still represent an enriched set of transcripts corresponding to expressed genes and are clearly worth exploiting if available.

As with other sequence databases used in proteomics, peptide sequence tags from *de novo* sequenced peptides can also be searched against EST databases. Liska et al. have developed an error tolerant search tool, multiTag, specifically for this task (44, 45).

3. Future Developments in Cross Species Proteomics

It is unlikely that PMF will ever be capable of cross species protein identification over large phylogenetic distances, and such challenges are also beyond standard tandem MS/MS methods. Presently, *de novo* sequencing based approaches also remain quite challenging to implement for high throughput work over such evolutionary distances. However, the future holds many developments which will make cross species proteomics somewhat easier. With the development of next generation genome sequencing tools and the ever improving annotation pipelines, it is likely that the next decade will see a further explosion in the number of available genomes covering the world's biodiversity. Also, new technological advancements in proteomics and mass spectrometry are likely to improve the capabilities of tandem MS/MS and make the *de novo* sequencing of peptide a possibility for high throughput experiments (c.f. (61, 62)). Other technologies such as positional proteomics (63) and chemical modification of peptides should also help improve the depth to which the proteome can be explored. Targeted quantification of proteins in cross species proteomics is likely to also become more of an area of interest as identification across the species boundary becomes easier. Spectral libraries (64) and identification methods other than sequence database searches offer an interesting new direction for cross species proteomics. The capability of how well a cross species project can work always comes down to the availability of closely related species and their genomes, but methods of protein function identification without sequence database searching could be the way forward. In this chapter, we have tried to lay down the background to cross species proteomics and some best practices from which others can design their experiments.

Acknowledgments

James Wright acknowledges NERC (NER/S/R/2005/13607) for his PhD funding.

References

1. Mardis E. R. (2008) The Impact of Next Generation Sequencing Technology on Genetics, *Trends in Genetics*, **24**: 133–141
2. May R. M. (1990) How Many Species? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **330**: 293–304
3. Ansong C., Purvine S. O., Adkins J. N., Lipton M. S., Smith R. D. (2008) Proteogenomics: Needs and Roles to be Filled by Proteomics in Genome Annotation, *Briefings in Functional Genomics and Proteomics*, **7**: 50–62

4. Wilkins, M. R., Williams K. L. (1997) Cross-Species Protein Identification using Amino Acid Composition, Peptide Mass Fingerprinting, Isoelectric Point and Molecular Mass: A Theoretical Evaluation, *Journal of Theoretical Biology,* **186**: 7–15
5. Lester P. J., Hubbard S. J. (2002) Comparative Bioinformatic Analysis of Complete Proteomes and Protein Parameters for Cross-Species Identification in Proteomics, *Proteomics,* **2**: 1392–1405
6. Cordwell S. J., Wilkins M. R., Cerpa-Poljak A., Gooley A. A., Duncan M., Williams K. L., Humphery-Smith I. (1995) Cross-Species Identification of Proteins Seperated by Two-Dimensional Gel Electrophoresis using Matrix-Assisted Laser Desorption/Time-of-Flight Mass Spectrometry and Amino Acid Composition, *Electrophoresis,* **16**: 438–443
7. Cordwell S. J., Basseal D. J., Humphery-Smith I. (1997) Proteome Analysis of *Spiroplasma melliferum* (A56) and Protein Characterisation Across Species Boundaries, *Electrophoresis,* **18**: 1335–1346
8. Wasinger V. C., Cordwell S. J., Cerpa-Poljak A., Yan J. X., Gooley A. A., Wilkins M. R., Duncan M. W., Harris R., Williams K. L., Humphery-Smith I. (1995) Progress with Gene-Product Mapping of the Mollicutes: Mycoplasma Genitalium, *Electrophoresis*, **16**: 1090–1094
9. Cordwell S. J., Humphery-Smith I. (1997) Evaluation of Algorithms used for Cross-Species Proteome Characterisation, *Electrophoresis*, **18**: 1410–1417
10. Wilkins M. R., Gasteiger E., Tonella L., Ou K., Tyler M., Sanchez J. C., Gooley A. A., Walsh B. J., Bairoch A., Appel R. D., Williams K. L., Hochstrasser D. F. (1998) Protein Identification with N and C Terminal Sequence Tags in Proteome Projects, *Journal of Molecular Biology*, **278**: 599–608
11. Wilikins M. R., Gasteiger E., Wheeler C. H., Lindskog I., Sanchez J. C., Bairoch A., Appel R. D., Dunn M. J., Hochstrasser D. F. (1998) Multiple Parameter Cross-Species Protein Identification using Multiident – A World-Wide Web Acessible Tool, *Electrophoresis,* **19**: 3199–3206
12. Cordwell S. J., Wasinger V. C., Cerpa-Poljak A., Duncan M. W., Humphery-Smith I. (1997) Conserved Motifs as the Basis for Recognition of Homologous Proteins across the Species Boundaries using Peptide Mass Fingerprinting, Journal of Mass Spectrometry, **32**: 370–378
13. Verrills N. M., Harry J. H., Walsh B. J., Hains P. G., Robinson E. S. (2000) Cross-Matching Marsupial Proteins with Eutherian Mammal Databases: Proteome Analysis of Cells from UV-Induced Skin Tumours of an Opossum (*Monodelphis domestica*), *Electrophoresis,* **21**: 3810–3822
14. Molloy M. P., Phadke N. D., Maddock J. R., Andrews P. C. (2001) Two-Dimensional Electrophoresis and Peptide Mass Fingerprinting of Bacterial Outer Membrane Proteins, *Electrophoresis,* **22**: 1686–1696
15. Shevchenko A., Sunyaev S., Loboda A., Bork P., Ens W., Standing K. G. (2001) Charting the Proteomes of Organisms with Unsequenced Genomes by MALDI-Quadrapole Time-of-Flight Mass Spectrometry and BLAST Homology Searching, Analytical Chemistry, **73**: 1917–1926
16. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs, *Nucleic Acids Res*earch, **25**: 3389–402
17. Liska A. J., Shevchenko A. (2003) Expanding the Organismal Scope of Proteomics: Cross-Species Protein Identification by Mass Spectrometry and its Implications, *Proteomics,* **3**: 19–28
18. Habermann B., Oegema J., Sunyaev S., Shevchenko A. (2004) The Power and Limitations of Cross-Species Identification by Mass Spectrometry-Driven Sequence Similarity Searches, Molecular & Cellular Proteomics, **3**: 238–249
19. Grossmann J., Fischer B., Baerenfaller K., Owiti J., Buhmann J. M., Gruissem W., Baginsky S. (2007) A Workflow to Increase the Detection Rate of Proteins from Unsequenced Organisms in High-Throughput Proteomics Experiments, *Proteomics,* 7: 4245–4254
20. Waridel P., Frank A., Thomas H., Surendranath V., Sunyaev S., Pevzner P., Shevchenko A. (2007) Sequence Similarity-Driven Proteomics in Organisms with Unknown Genomes by LC-MS/MS and Automated De novo Sequencing, , *Proteomics,* 7: 2318–2329
21. Pitzer E., Masselot A., Colinge J. (2007) Assessing Peptide De novo Sequencing Algorithms Performance on Large and Diverse Data Sets, *Proteomics,* 7: 3051–3054
22. Pevtsov S., Fedulova I., Mirzaei H., Buck C., Zhang X. (2006) Performance Evaluation of Existing De novo Sequencing Algorithms, *Journal of Proteome Research*, **5**: 3018–3028
23. Ma B., Zhang K., Hendrie C., Liang C., Li M., Doherty-Kirby A., Lajoie G. (2003) PEAKS: Powerful Software for Peptide De novo Sequencing by Tandem Mass Spectrometry, Rapid Communications in Mass Spectrometry, **17**: 2337–2342

24. Grossmann J., Roos F. F., Cieliebak M., Liptak Z., Mathis L. K., Muller M. (2005) Gruissem W., Baginsky S., AUDENS: A Tool for Automated Peptide De novo Sequencing, *Journal of Proteome Research*, **4**: 1768–1774

25. Halligan B. D., Ruotti V., Twigger S. N., Greene A. S. (2005) DeNovoID: A Web-Based Tool for Identifying Peptides from Sequence and Mass Tags Deduced from De novo Peptide Sequencing by Mass Spectrometry, *Nucleic Acids Research,* **33**: W376–W381

26. Raucci G., Gabrielli M., Novelli S., Picariello G., Collins S. H. (2005) CHASE, a Charge-Assisted Sequencing Algorithm for Automated Homology Based Protein Identification with Matrix Assisted Laser Desorption/Ionization Time of Flight Post-Source Decay Fragmentation Data, *Journal of Mass Spectrometry,* **40**: 475–488

27. Han Y., Ma B., Zhang K. (2004) SPIDER: Software for Protein Identification from Sequence Tags with De novo Sequencing Error, *Proceedings/IEEE Computational Systems Bioinformatics Conference*, 206–215

28. Russeth K. P., Higgins L., Andrews M. T. (2006) Identification of Proteins from Non-Model Organisms Using Mass Spectrometry: Application to a Hibernating Mammal, *Journal of Proteome Research,* **5**: 829–839

29. Ostrowshi M., Fegatella F., Wasinger V., Guilhaus M., Corthals G. L., Cavicchioli R. (2004) Cross-Species Identification of Proteins from Proteome Profiles of the Marine Oligotrophic Ultramircobacterium, Sphingopyxis alaskensis, *Proteomics,* **4**: 1779–1788

30. Kim H. J., Lee D. Y., Lee D. H., Park Y. C., Kweon D. H., Ryu Y. W., Seo J. H. (2004) Strategic Proteome Analysis of *Candida magoliae* with an Unsequenced Genome, *Proteomics,* **4**: 3588–3599

31. Samyn B., Sergeant K., Memmi S., Debyser G., Devreese B., Van Beeumen J. (2006) MALDI-TOF/TOF De novo Sequence Analysis of 2D PAGE Seperated Proteins from *Halorhodospira halophila*, a Bacterium with Unsequenced Genome, *Electrophoresis,* **27**: 2702–2711

32. Savidor A., Donahoo R. S., Hurtado-Ganzales O., Land M. L., Shah M. B., Lamour K. H., McDonald W. H. (2008) Cross Species Global Proteomics Reveals Conserved and Unique Processes in *Phytophthora sojae* and *Phytophthora ramorum*, *Molecular & Cellular Proteomics,* 7: 1501–1516

33. Ahram M., Strittmatter E. F., Monroe M. E., Adkins J. N., Hunter J. C., Miller J. H., Springer D. L. (2005) Identification of Shed Proteins from Chinese Hamster Ovary Cells: Application of Statistical Confidence using Human and Mouse Protein Databases, *Proteomics,* **5**: 1815–1826

34. Nanduri B., Lawrence M. L., Vanguri S., Brugess S. C. (2005), Proteomics Analysis using an Unfinished Bacterial Genome: The Effects of Subminimum Inhibitory Concertrations of Antibiotics on Mannheimia Haemolytica Virulance Factor Expression, *Proteomics,* **5**: 4852–4863

35. Sun J., Wang W., Hundertmark C., Zeng A. P., Jahn D., Deckwer W. D. (2006) A Protein Database Constructed from Low Coverage Genomic Sequence of *Bacillus megaterium* and its use for Accelerated Proteomics Analysis, *Journal of Biotechnology*, **124**, 3, 486–495

36. Huang M., Chen T., Chan Z. (2006) An Evaluation for Cross-Species Proteomics Research by Publicly Available Expressed Sequence Tag Database Search using Tandem Mass Spectral Data, *Rapid Communications in Mass Spectrometry*, **20**: 2635–2640

37. Edwards N. J. (2007) Novel Peptide Identification from Tandem Mass Spectra using ESTs and Sequence Database Compression, *Molecular Systems Biology*, **3**: 102

38. Grimplet J., Gasper J. W., Gancel A., Sauvage F., Romieu C. (2005) Including Mutations from Conceptually Translated Expressed Sequence Tags into Orthologous Proteins Improves the Preliminary Assignment of Peptide Mass Fingerprints on Non-Model Genomes, *Proteomics*, **5**: 2769–2777

39. Kwon K., Kim M., Kim J. Y., Kim K. W., Kim S., Park Y. M., Yoo J. S. (2003) Efficiency Improvement of Peptide Identification for an Organism without Complete Genome Sequence, using Expressed Sequence Tag Database and Tandem Mass Spectral Data, *Proteomics,* **3**: 2305–2309

40. Kim S. I., Kim J. Y., Kim E. A., Kwon K. H., Kim K. W., Cho K., Lee J. H., Nam M. H., Yang D. C., Yoo J. S., Park Y. M. (2003) Proteome Analysis of Hairy Root from Panax Ginseng C.A. Meyer using Peptide Fingerprinting, Internal Sequencing and Expressed Sequence Tag Data, *Proteomics,* **3**, 2379–2392

41. Nam M. H., Heo E. J., Kim J. Y., Kim S. I., Kwon K. H., Seq J. B., Kwon O., Yoo J. S., Park Y. M. (2003) Proteome Analysis of the Responses of Panax Ginseng C. A. Meyer Leaves to High Light: Use of Electrospray Ionization Quadrapole Time of Flight Mass Spectrometry and Expressed Sequence Tag Data, *Proteomics,* **3**: 2351–2367

42. Porubleva L., Vander Velden K., Kothari S., Oliver D. J., Chitnis P. R. (2001) The Proteome of Maize Leaves: Use of Gene

Sequences and Expressed Sequence Tag Data for Identification of Proteins with Peptide Mass Fingerprints, *Electrophoresis*, **22**: 1724–1738

43. Mooney B. P., Krishnan H. B., Thelen J. J. (2004) High Throughput Mass Fingerprinting of Soy Bean Seed Proteins: Automated Workflow and Utility of Unigene Expressed Sequence Tag Databases for Protein Identification, *Phytochemistry*, **65**: 1733–1744
44. Sunyaev S., Liska A. J., Golod A., Schevchenko A. (2003) MultiTag: Multiple Error-Tolerant Sequence Tag Search for the Sequence-Similarity Identification of Proteins by Mass Spectrometry, *Analytical Chemistry*, **75**: 1307–1315
45. Liska A. J., Sunyaev S., Shilov I. N., Schaeffer D. A., Schevchenko A. (2006) Error-tolerant EST Database Searches by Tandem Mass Spectrometry and multiTag Software, *Proteomics,* **5**: 4118–4122
46. Snijders A. P., De Koning B., Wright P. C. (2007) Relative Quantification of Proteins Across the Species Boundary Through the use of Shared Peptides, *Journal of Proteome Research*, **6**: 97–104
47. Pandhal J., Snijders A. P., Wright P. C., Biggs C. A. (2008) A Cross-Species Quantitative Proteomics Study of Salt Adaption in a Halotolerant Enviromental Isolate using 15N Metabolic Labelling, *Proteomics,* **8**: 2266–2284
48. Thiede B., Hohenwarter W., Krah A., Mattow J., Schmid M., Schmidt F., Jungblut P. R. (2005) Peptide Mass Fingerprinting, *Methods*, **35**: 237–247
49. Hernandez P., Muller M., Appel R. D. (2006) Automated Protein Identification by Tandem Mass Spectrometry: Issues and Strategies, *Mass Spectrometry Reviews,* **25**: 235–254
50. Peng J., Gygi S. P. (2001) Proteomics: The Move to Mixtures, *Journal of Mass Spectrometry,* **36**: 1083–1091
51. Standing K. G. (2003) Peptide and Protein De novo Sequencing by Mass Spectrometry, C*urrent Opinion Structural Biology*, **13**: 595–601
52. Issaq H., Veenstra T. (2008) Two-Dimensional Polyacrylamide Gel Electrophoresis (2D-PAGE): Advantages and Perspectives, *Biotechniques*, **44**: 697–700
53. Van den Bergh G., Arckens L. (2005) Recent Advances in 2D Electrophoresis: An Array of Possibilities, *Expert Reviews Proteomics,* **2**: 243–252
54. Marengo E., Robotti E., Bobba M. (2008) 2D-PAGE Maps Analysis, *Methods Molecular Biology,* **428**: 291–325
55. Perkins D. N., Pappin D. J. C., Creasy D. M., Cottrell J. S. (1999) Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data, *Electrophoresis,* **20**: 3551–3567
56. Roe M. R., Griffin T. J. (2006) Gel-free Mass Spectrometry Based High Throughput Proteomics: Tools for Studying Biological Response of Proteins and Proteomes, *Proteomics,* **6**, 17, 4678–4687
57. Palagi P. M., Hernandez P., Walther D., Appel R. D. (2006) Proteome Informatics I: Bioinformatics Tools for Processing Experimental Data, *Proteomics*, **6**, 20, 5435–5444
58. Elias J., Gygi S. (2007) Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry, *Nature Methods,* **4**: 207–214
59. Deutsch E. W., Lam H., Aebersold R. (2008) Data Analysis and Bioinformatics Tools for Tandem Mass Spectrometry in Proteomics, *Physiological Genomics,* **33**, 1, 18–25
60. Mann M., Wilm M. (1994) Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence, *Analytical Chemistry*, **66,** 24: 4390–4399
61. Kim S, Gupta N, Bandeira N, Pevzner PA. (2009) Spectral dictionaries: Integrating De novo Peptide Sequencing with Database Search of Tandem Mass Spectra, *Molecular & Cellular Proteomics,* **8**: 53–69
62. Bandeira N., Pham V., Pevzner P., Arnott D., Lill J.R. (2008) Automated De novo Protein Sequencing of Monoclonal Antibodies, *Nature Biotechnology*, **26**: 1336–1338
63. McDonald L., Beynon R. J. (2006) Positional Proteomics: Preparation of Amino-Terminal Peptides as a Strategy for Proteome Simplification and Characterization, *Nature Protocols,* **1**, 4: 1790–1798
64. Lam H., Deutsch E. W., Eddes J. S., Eng J. K., King N., Stein S. E., Aebersold R. (2007) Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS, *Proteomics*, **7**, 5: 655–667

Chapter 10

Gene Model Detection Using Mass Spectrometry

Bindu Nanduri, Nan Wang, Mark L. Lawrence, Susan M. Bridges, and Shane C. Burgess

Abstract

The utility of a genome sequence in biological research depends entirely on the comprehensive description of all of its functional elements. Analysis of genome sequences is still predominantly gene-centric (i.e., identifying gene models/open reading frames). In this article, we describe a proteomics-based method for identifying open reading frames that are missed by computational algorithms. Mass spectrometry-based identification of peptides and proteins from biological samples provide evidence for the expression of the genome sequence at the protein level. This proteogenomic annotation method combines computationally predicted ORFs and the genome sequence with proteomics to identify novel gene models. We also describe our proteogenomic mapping pipeline – a set of computational tools that automate the proteogenomic annotation work flow. This pipeline is available for download at www.agbase.msstate.edu/tools/.

Key words: Proteogenomic annotation, Peptide validation, PGM, Proteogenomic mapping pipeline, ePST, Expressed protein sequence tag, Gene models

1. Introduction

Rapid advances in genome sequencing technologies and the resulting explosion in the availability of bacterial genome sequences highlight the need for identifying and "annotating" the biological function of all nucleotides in the sequence. The functional elements in bacterial genomes could be protein coding regions, i.e., open reading frames (ORFs), and noncoding RNAs, as well as regulatory elements that are involved the expression of proteins and RNAs (1). Here, we focus on annotating protein coding genes, and for the purposes of this article, "genome annotation" refers to identification, demarcation and delineation of protein coding genes. Genome annotation for predicting open reading frames goes hand in hand with sequencing efforts, but

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_10, © Humana Press, a part of Springer Science+Business Media, LLC 2010

most commonly relies solely on computational algorithms and does not include experimental data, which are often collected for model organisms as EST/cDNA sequencing data (2). Despite improvement in the accuracy of gene prediction programs over the last few years, prediction of short genes still remains challenging (3).

A practical solution for generating accurate gene models for a particular genome is a combinatorial approach that includes computational predictions and experimental methods. When proteome data are used, this approach is called proteogenomic mapping; it combines mass spectrometry-based proteomic workflows with computationally predicted ORFs to confirm expression of predicted proteins, correct gene prediction start and stop codons, identify protein post translational modifications, and identify novel genes missed by initial annotation (4–7). Here, we describe our proteogenomic annotation workflow, which includes a novel method for assigning confidence to peptide identifications and a neural network framework for assigning confidence to newly identified gene models.

2. Materials

The proteogenomic mapping workflow requires a sequenced bacterial genome, the existing protein models for the genome, and a proteomics dataset.

2.1. Peptide Validation and Identification

1. FASTA format genome sequence of the organism of interest should be downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes) or another source. When translated in all six reading frames, the sequences generated will constitute the genome database (gdb).
2. The sequences of all predicted proteins from the organism of interest should be downloaded in FASTA format from NCBI (ftp://ftp.ncbi.nih.gov/genomes) or another source and are used to generate the protein database (prodb).

2.2. Protein Isolations and Mass Spectrometry

1. Total proteins from the organism of interest should be isolated, quantified, trypsin digested and analyzed by tandem mass spectrometry to generate the tandem mass spectra files. Please see our published work in Gram negative and Gram positive bacteria for details about the methods (8–11). We are not providing specific details for this section as the workflows can be diverse, varied, and are beyond the scope of this review, but any mass spectrometry-based proteomics method should be suitable.

2. DBRandomizer tool at AgBase (www.agbase.msstate.edu/epst) is used to generate rprodb and rgdb (randomized decoy databases of prodb and gdb, respectively).
3. Use the Sequest algorithm in Bioworks (ThermoElectron Corp., San Jose, CA) for searching mass spectra after in silico digestion of gdb, prodb, rprodb and rgdb to generate XML output files (see Note 2).
4. Use the PepOut tool (www.agbase.msstate.edu/ePST) with Bioworks XML files to generate a list of identified peptides at a user defined *p*-value threshold.

2.3. Expressed Protein Sequence Tag (ePST) Generation, Feature Collection and Validation

Once the proteomics data has been generated and sets of high quality peptides have been identified using both the genome and proteome databases, this data is used to generate expressed Protein Sequence Tags (ePST), which are DNA sequences that potentially correspond to previously unidentified genes.

1. Download and install the proteogenomic mapping pipeline (PGM) from AgBase (www.agbase.msstate.edu/tools.html).
2. PGM operates under the Windows operating system. The system requirements to run PGM with the download links are listed below:
   (a) Perl and BioPerl (www.cpan.org.; http://www.activestate.com/Products/activeperl/index.mhtml).
   (b) NCBI Local BLAST Package (ftp://ftp.ncbi.nih.gov/blast/).
   (c) Weka machine learning tool kit (http://www.cs.waikato.ac.nz/ml/weka/).
3. For homology searches, a user defined FASTA database of proteins from all closely related organisms.
4. Conserved Domain Database (CDD) (ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd; File: cdd.tar.gz).

3. Methods

Proteogenomic annotation for identifying new, previously unpredicted genes builds upon a typical tandem mass spectrometry-based proteomics analysis subsequent to the isolation of proteins from an organism of interest. Protein identification involves searching tandem mass spectra of peptides against a protein database (digested *in silico* with a protease like trypsin) to identify peptides and their corresponding proteins. We broadly categorize our proteogenomic mapping work flow for identifying new gene models into four steps: peptide identification and validation, ePST

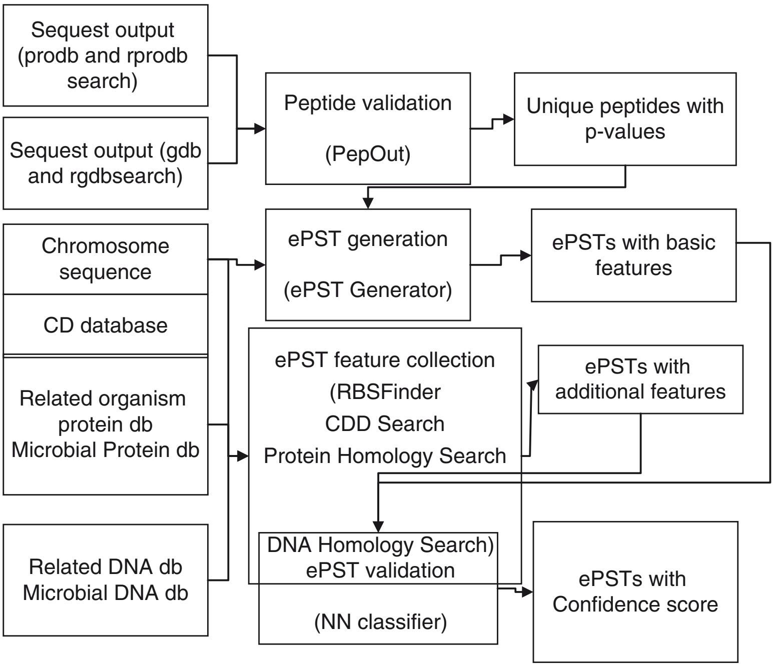

Sequest output (prodb and rprodb search)
Sequest output (gdb and rgdbsearch)
Peptide validation (PepOut)
Unique peptides with p-values
Chromosome sequence
CD database
Related organism protein db Microbial Protein db
Related DNA db Microbial DNA db
ePST generation (ePST Generator)
ePSTs with basic features
ePST feature collection (RBSFinder CDD Search Protein Homology Search DNA Homology Search)
ePSTs with additional features
ePST validation (NN classifier)
ePSTs with Confidence score

Fig. 1. Flowchart of proteogenomic mapping pipeline (PGM)

generation, ePST feature collection, and ePST evaluation as shown in Fig. 1. A brief overview of each of these segments with specific steps is described in the following paragraphs.

3.1. Peptide Identification and Validation

For annotating the genome using mass spectra, we perform searches with the mass spectra against a protein database as well as the corresponding genome sequence translated in all six reading frames. Both databases are *in silico* digested, and the searches are conducted separately. Proteomics-based identification of potential new protein coding ORFs entirely depends on the quality of initial peptide identifications. Therefore, peptides identified with high confidence are critical for proteomics based identification of gene models. We couple a randomized decoy database strategy with distance-based outlier detection for assigning probabilities for peptide identification and validation. The PepOut tool implements an unsupervised machine learning model for distance-based outlier detection to estimate the accuracy of peptide assignments to tandem mass (MS/MS) spectra (12).

1. Generate decoy prodb and gdb databases using the DBRandomizer tool at agbase.msstate.edu/tools/epst.
2. *In silico* trypsin digests prodb, rprodb, and gdb in Bioworks using the *index* function with differential modifications for methionine (single and double oxidations +16 and +32, respectively) and Cysteine carbamidomethylation (+57.2) and perform Sequest search (see Note 2).

3. Export the search results without applying filters for peptide or protein identifications in XML format.
4. Run PepOut with XML files from real prodb,gdb, rprodb and gdb search results to generate a list of peptides (p value ≤ 0.05) in txt format.

3.2. ePST Generation

Peptide lists generated from searches against prodb (proteome database) and gdb (genome database) for a user defined confidence threshold are compared to identify peptides that match only the gdb but not the prodb. The genome-specific peptides represent expressed evidence for potential novel gene models. The sequence coordinates are determined for the genome specific peptides, and the corresponding nucleotide sequence is extended to the first in-frame stop codon in both the 5′ and 3′ directions to generate an expressed Protein Sequence Tag (ePST) that represents a potential novel gene. To find a potential start codon for the ePST, we scan the nucleotide sequence from the first in-frame stop upstream of the peptide nucleotide sequence for the first in-frame start in the 3′ direction. If there is a start codon, it is used to designate the start of the ePST (Fig. 2). In the absence of an in-frame start codon prior to the beginning of the nucleotide sequence corresponding to the genome-specific peptide, the beginning of the peptide itself is used as the start of the ePST.

The process for generating ePSTs, i.e., proteogenomic mapping pipeline (PGM), is completely automated and is implemented in the following steps (see Note 6).

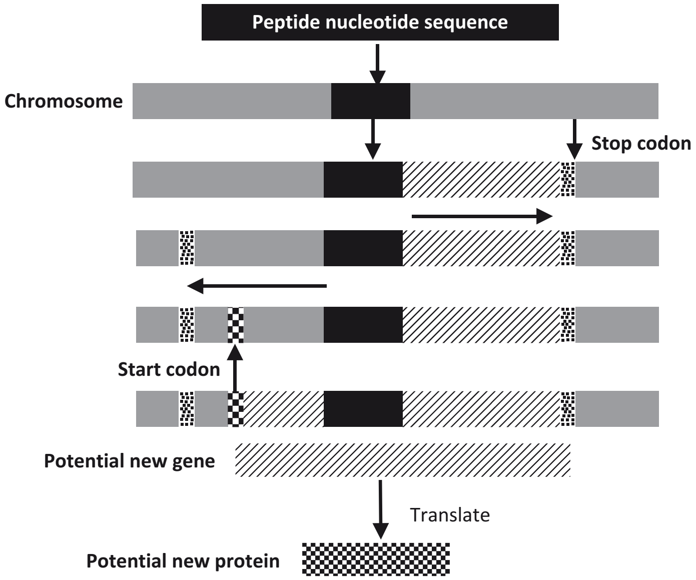


Fig. 2. Generation of ePST

1. Removes the differential modifications from PepOut output (*see* Note 1), compares the sequences of peptides identified in gdb and prodb searches to identify gdb-specific unique peptides.
2. Translates the chromosome in all six reading frames and maps each unique peptide in the appropriate reading frame (see Note 3).
3. Extends the nucleotide sequence that corresponds to the unique peptide in the 5′ and 3′ directions until the first in frame TAG/TGA/TAA stop codon.
4. At the 5′ end, identifies, if any, the first in-frame ATG/GTG/TTG start codon.
5. Designates the sequence between start and stop codon as ePST
6. Translates the nucleotide sequence of ePST to protein.

3.3. ePST Feature Collection

To provide an orthogonal evaluation of the validity of the identified ePSTs as novel genes, PGM compiles peptide level and protein level data for each ePST described in the following steps.

1. Determines if the ePST contains a canonical start codon: ATG/TTG/GTG.
2. Determines the length of the ePST and retains the probability from PepOut of the genome-specific peptide.
3. Determines the number of genome specific peptide matches for each ePST nucleotide sequence.
4. Computes the coverage of ePST by unique peptides.
5. Determines whether the potential novel gene models have a Shine–Dalgarno sequences for protein translation in bacteria using RBSfinder for prediction of ribosomal binding sites.
6. Records the start codon reported by RBSfinder.
7. Conducts BLAST searches with novel ePSTs against closely related species at the nucleotide and protein levels separately.
8. Identifies conserved functional domains, if any, for novel ORF encoded proteins using RPS-BLAST searches against CDD from NCBI.ePST features.

3.4. ePST Evaluation

The proteogenomic annotation workflow typically ends with the identification of the ePST, or novel protein-encoding gene, with various features. However, the end-user still has to identify true protein-encoding genes from potential gene models. Potential novel protein coding ORFs can be validated by RT-PCR. Nevertheless, global genome wide studies often result in large lists of potential ePSTs, and it is not practical to validate each and every potential novel ORF. What is desirable is to generate confidence scores for all potential novel ORFs that can be used by the end user for selecting ORFs for validation by experimental approaches. We address this challenging task of evaluating novel protein-encoding genes and generating confidence scores for potential novel genes (13) in the following steps

1. We generate a training dataset that contains all possible combinations of feature values (*see* Note 4). Scientists with extensive experience in bacterial genomics provide an evaluation of each item in the training set based on the feature values (*see* Note 5).
2. We use our labeled training dataset to build neural network models to predict evaluations based on the feature values.
3. This model is used to generate a confidence score for each of the identified ePSTs. The resulting scores are then used to rank the potential genes for validations using RT-PCR.

4. Notes

1. We routinely use peptides with a *p* value of ≤ 0.05 from prodb and gdb Bioworks searches (PepOut output) for ePST generation described in Subheadings 2.2 and 3.1.
2. We search mass spectra against real and random databases separately using Sequest search algorithm described in Subheading 3.1.
3. We identify the chromosomal location of the genome-specific peptide that is the seed for ePST generation in Bioworks by doing string searches against the chromosome described in Subheading 3.2.
4. End users can use the training dataset that we generated using *Mannheimia haemolytica* when they use PGM for evaluation of novel ePSTs described in Subheading 3.4. However, they can also generate their own training datasets that can be used with our PGM. The training dataset should contain as many possible combinations of features as possible.
5. The range of confidence scores for an ePST generated by neural network model in Subheadings 3.4 is the same as the range of confidence score in the training dataset provided by experts.
6. Published proteogenomic annotation approaches in the literature (14, 15) differ from our method in peptide identification described in Subheadings 3.1.These alternate approaches utilize only the genome sequence database for identifying peptides. However, once a list of peptides that are unique to the noncoding regions of the genome is generated, these methods can utilize PGM.

Acknowledgments

The project was supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number #2006-35600-17688. It was partially funded by a grant from the National Science Foundation (EPS-0556308-06040293).

References

1. Gottesman, S. (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends in Genetics*, *21*, 399–404.
2. Stein, L. (2001) Genome annotation: from sequence to biology. *Nature Reviews*, *2*, 493–503.
3. Guigo, R., Flicek, P., Abril, J. F., Reymond, A., *et al.* (2006) EGASP: the human ENCODE genome annotation assessment project. *Genome Biology*, *7 Suppl 1*, S2.1–31.
4. McCarthy, F. M., Cooksey, A. M., Wang, N., Bridges, S. M., *et al.* (2006) Modeling a whole organ using proteomics: the avian bursa of Fabricius. *Proteomics*, *6*, 2759–2771.
5. Jaffe, J. D., Stange-Thomann, N., Smith, C., DeCaprio, D., *et al.* (2004) The complete genome and proteome of *Mycoplasma mobile*. *Genome Research*, *14*, 1447–1461.
6. Jaffe, J. D., Berg, H. C., Church, G. M. (2004) Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*, *4*, 59–77.
7. McCarthy, F. M., Wang, N., Magee, G. B., Nanduri, B., *et al.* (2006) AgBase: a functional genomics resource for agriculture. *BMC Genomics*, *7*, 229.
8. Nanduri, B., Shah, P., Ramkumar, M., Allen, E. B., *et al.* (2008) Quantitative analysis of *Streptococcus pneumoniae* TIGR4 response to in vitro iron restriction by 2-D LC ESI MS/MS. *Proteomics*, *8*, 2104–2114.
9. Nanduri, B., Lawrence, M. L., Vanguri, S., Burgess, S. C. (2005) Proteomic analysis using an unfinished bacterial genome: the effects of subminimum inhibitory concentrations of antibiotics on *Mannheimia haemolytica* virulence factor expression. *Proteomics*, *5*, 4852–4863.
10. Nanduri, B., Lawrence, M. L., Peddinti, D. S., Burgess, S. C. (2008) Effects of subminimum inhibitory concentrations of antibiotics on the *Pasteurella multocida* proteome: a systems approach. *Comparative and functional genomics*, 254836.
11. Donaldson, J. R., Nanduri, B., Burgess, S. C., Lawrence, M. L: Comparative proteomic analysis of *Listeria monocytogenes* strains F2365 and EGD. *Applied and Environmental Microbiology* 2009, 75(2): 366–373.
12. Wang, N, Yuan C, Wu, D, Burgess, S, Nanduri B, Bridges SM: "PepOut Distance-based Outlier Detection Model for Improving MS/MS Peptide Identification Confidence". MCBIOS 2008, February 23–24, 2008 in Oklahoma City.
13. Wang, N., Yuan, C., Burgess, S. C., Nanduri, B., *et al.* (2008) *WORLDCOMP'08 – The 2008 World Congress in Computer Science, Computer Engineering, and Applied Computing*, Las Vegas, Nevada, USA.
14. Gupta, N., Tanner, S., Jaitly, N., Adkins, J. N., *et al.* (2007) Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Research*, *17*, 1362–1377.
15. Gupta, N., Benhamida, J., Bhargava, V., Goodman, D., *et al.* (2008) Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Research*, *18*, 1133–1142.

Chapter 11

Signal Processing in Proteomics

Rene Hussong and Andreas Hildebrandt

Abstract

Computational proteomics applications are often imagined as a pipeline, where information is processed in each stage before it flows to the next one. Independent of the type of application, the first stage invariably consists of obtaining the raw mass spectrometric data from the spectrometer and preparing it for use in the later stages by enhancing the signal of interest while suppressing spurious components. Numerous approaches for preprocessing MS data have been described in the literature. In this chapter, we will describe both, standard techniques originating from classical signal and image processing, and novel computational approaches specifically tailored to the analysis of MS data sets. We will focus on low level signal processing tasks such as baseline reduction, denoising, and feature detection.

Key words: Mass spectrometry, Proteomics, Signal processing, Base line reduction, Denoising, Peak picking, Feature detection

1. Introduction

Mass spectrometry (MS) coupled with high performance liquid chromatography (HPLC) has become the de-facto experimental standard for analysis in proteomics. Owing to the huge amount of data produced by a single MS experiment and the sheer complexity of the data to be analyzed, computational techniques to interpret the recorded signals are indispensable. Despite recent advantages toward a fully automated analysis of MS data, the early steps in the computational pipeline are still challenging and not satisfactorily solved: although the signal-to-noise ratio (S/N) and the mass accuracy have improved drastically with modern spectrometer generations, low-abundant peptides, chemical noise, and overlapping patterns still hamper the efficient analysis of proteomic data. Consequently, one of the first steps in computational proteomics is the enhancement of the recorded raw spectra by amplifying the

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_11, © Humana Press, a part of Springer Science+Business Media, LLC 2010

signal of interest and suppressing spurious artefacts. This is often accompanied by recasting the signal into a different kind of representation, e.g. by converting the raw spectrum into a set of sticks at the locations of assumed mass peaks. This stage is known as the *signal processing stage* and since it lies at the very root of each computational proteomics pipeline, its quality is crucial for the success of most proteomics techniques. Indeed, the importance of signal processing is continuously increasing: the manual inspection of the spectra, which is arguably the most flexible and often the most accurate analysis technique by far, is only feasible for applications generating modest amounts of data and thus presents an impenetrable bottleneck for high-throughput approaches.

In practice, the simple picture of an analysis pipeline where the signal processing constitutes the first step (or the second, considering the actual experiment as the first stage) is often only a crude approximation to reality. For example, mass spectrometers themselves usually implement some signal processing techniques of their own, and hence, the data that is obtained from the spectrometer is not the real raw data, but an already preprocessed variant. This signal processing is typically interleaved with the experiment (e.g. in the case of MS/MS runs) so that, in a strict view, it cannot be extracted into an individual stage of the pipeline. Similarly, several stages further down the pipeline are often realized in a fashion that mixes signal processing with further analysis, e.g. in the case of differential quantitation, where the quantitation stage often recruits “raw” or merely noise-filtered, but not fully processed data. Nevertheless, a number of processing steps can be identified that are performed in most analysis pipelines, even though they might come in a different ordering or may only be performed implicitly. For the remainder of this work, we will consider baseline reduction, noise filtering, peak detection and fitting, and to some degree higher-level feature detection as parts of the signal processing step. The literature on these topics is far too voluminous to allow for an exhaustive review here. Instead, we will discuss the general problems encountered and some of the more popular approaches to their solution and give a brief outlook on further ongoing challenges.

2. Mass Spectrometry Data

From a computer scientist’s perspective, a mass spectrometric scan $S{:}=\{s_0..s_{N-1}\}$ consists of N data points s_j where each measured point is described by two values: the m/z position p_j and the corresponding intensity I_j where m denotes the mass of the detected molecular fragment and *z\in {1, ..., Z}*, where \in is the “element of symbol its corresponding charge state. Often, the spacing between s_j and s_{j+1} is irregular and therefore prohibits a straight-forward analysis with standard preprocessing tools unless S is resampled by means of interpolation. Unfortunately, the position of a mass peak usually has to be determined within a maximum deviation of a few ppm, and hence, a "simple" resampling can have drastic effects if the sampling rate and/or the concentration of present peptides are low.

A further complication arises as a single molecule M that is subjected to MS does not only trigger a single peak in S, but rather a series of peaks $P := \{p_0..p_L\}$, with characteristic heights and spacings in between, where p_0 is often referred to as *monoisotopic peak*. Fig. 1 illustrates this effect: since every naturally occurring element can have a different number of neutrons, the masses of these isotopes differ by a multiple of the mass of a single neutron m_n, which is roughly 1 Da. When the molecular fragments are peptides – as it is the case in proteomics – the characteristic isotopic pattern, i.e. the sequence of triggered mass peaks, depends on the mass m of the peptide, but is – interestingly – largely independent from the exact amino acid composition[1] (cf., e.g., (1)). Please note that the spacing between neighbouring peaks also depends on the charge state z and is given by m_n/z.

Often, the mass spectrometer is coupled to a liquid chromatographic column (LC) in advance. Due to their physicochemical properties, peptides will leave the column at different time points[2] and therefore enter the mass spectrometer in "small" groups. This separation simplifies the processing and subsequent interpretation of the data drastically: peptides with similar masses will often leave

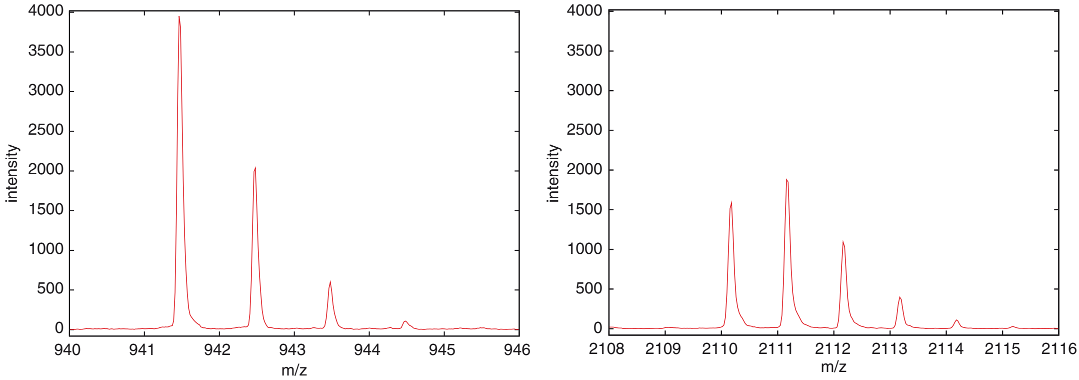


Fig. 1. Two isotope patterns of charge state 1 with a distance between neighbouring peaks of roughly 1 Da. Please note that in the right image the monoisotopic peak is no longer the largest

[1] Of course, this picture is changing as soon as we take posttranslational modifications or labelling techniques into account.

[2] Please note that peptides will usually elute over several subsequent time points and will therefore appear in several neighbouring scans.

the column at different time points, and hence occur in different mass spectra, while without the pre-separation their respective signals would overlap. Consequently, in an LC–MS data set, every data point s_j has an additional value, the so-called *retention time* RT_j. Since LC-MS data is thus two dimensional, we will refer to such data as LC–MS *maps*.

The situation described so far corresponds to an ideal measurement – the signals are optimally resolved and detectable, and only the signal of interest is recorded in the mass spectrum. In reality, though, matters are considerably more involved, since the signal of interest is overlaid with a number of so-called *parasitics*, severely spoiling the quality of the signal.

The first and most obvious of these artefacts seems so natural that it is hardly ever mentioned – the finite width of mass spectrometric peaks. It is easy to see that the finite precision of any experimental measurement process will ultimately "smear out" the mass information instead of providing a clear, sharp stick at the "correct" location. One of the main goals of the signal processing stage hence consists in detecting relevant peaks in the signal to revert the smearing-out-process by converting the signal into a series of sticks at the most probable locations and with height proportionate to the area under the curve in the original signal. Stick conversion usually needs to provide highly accurate mass values, even if in the original signal, two or more peaks were overlapping and have to be separated. The intensity information, on the other hand, is regarded as secondary and has a larger tolerance of error.

In addition to this line-broadening, three main types of parasitics can be distinguished: (a) a slowly varying, smooth baseline term, (b) highly variable electric and shot noise, and (c) chemical noise effects. An additional important distortion that is often relevant is miscalibration of the spectra. This effect and its remedies will not be discussed in this chapter due to space constraints (for an exemplary approach, cf., e.g., (2)).

For simpler intuition, the different parasitics are often imagined as a series of transforms "happening" to the signal: at first, the stick-shaped peaks broaden into peak curves of finite size. Several noise terms are then added to the resulting spectrum, which is finally overlaid with a baseline. With this picture in mind, it makes sense to apply a number of filters, each one approximately reversing one of the parasitic effects, "peeling off" layer after layer of unwanted terms. And indeed, most signal processing methods for computational proteomics that have been described in the literature proceed in this fashion, subsequently improving the quality of the signal before peaks are detected and converted to sticks. In the following sections, we will describe some of the filters proposed for cleansing mass spectrometric data from parasitics and some peak detection techniques relying on these filters before we will turn our attention to some recent techniques for feature detection that completely avoid the signal enhancement steps. The very first filter in most applications, a resampling of the data set to generate equally spaced spectra, is not further described, since it is trivial to achieve, even though problematic in its consequences, as will be discussed later.

3. Signal Processing

3.1. Base Line Reduction

The base line (e.g. Fig. 2) is a systematic error that has been associated with molecular fragments originating e.g. from sample preparation (3). Since these molecules are usually of low mass, the baseline effect diminishes with increasing m/z value.

A classical method for baseline removal is known as the top hat filter (4), a morphological operator that analyses the shape of “objects” in an image or signal. Morphological filters work on so-called structuring elements, i.e. they work locally within a predefined region, which is shifted over the signal. The dilation operator replaces the intensity by its supremum within the structuring element, while the erosion uses the infimum instead. Subtracting the opening, which is defined as erosion followed by dilation, from the original image, leads to the so-called white top hat, which extracts small bright structures and therefore removes slow trends like the baseline.

Another standard technique is the *sliding window* approach (5). In a sense, this method is similar to the previous one in that a certain function (similarly to the structuring element for morphological filters) is shifted over the signal in order to determine the mean or the median within that box. Fitting, e.g., an exponential function to these medians can yield a good approximation of the underlying baseline, which can be easily subtracted from the signal.

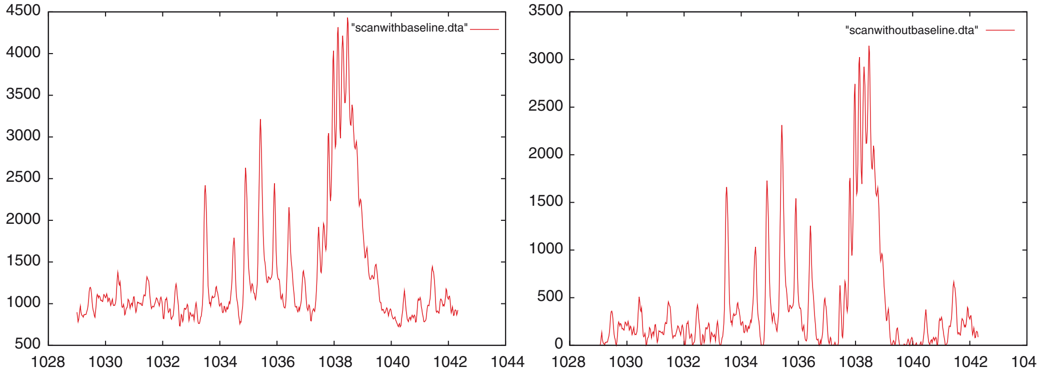


Fig. 2. Part of a scan with significant baseline (*on the left*) and its corrected counterpart (*on the right*) (50) (permission to reproduce this Figure was kindly granted by BMC Bioinformatics)

While the above techniques have been originally developed for the removal of slowly varying baseline terms in non-MS data, the approach presented in (3), based on stochastic *Bernstein* approximation, provides an example of a method specifically designed with MALDI mass spectra in mind.

3.2. Denoising

In proteomics, the term "noise" refers to two very different kinds of artefacts: a highly variable noise component that is present in all real-world experimental data sets and those parts of the signal that were indeed generated by molecular fragments, but not by peptides from the sample, the so-called *chemical noise*. Of course, removing both kinds of noise is necessary at some stage or other during the proteomics pipeline, but both require entirely different denoising techniques. Here, we will focus on the first kind of noise, while chemical denoising will be deferred until the final peak picking and feature detection stage.

Obviously, successful noise removal requires a suitable noise *model,* so that one can determine which components should be retained and which should be eliminated. The subject of noise models in mass spectra, however, has long been nearly overlooked. Consequently, most denoising techniques have traditionally assumed that all the noise in mass spectra (apart from the chemical noise) is electric in origin and thus follows a Johnson–Nyquist behavior, which can be very well approximated by the well-known white noise model with uniform frequency distribution. The most important effect of this kind of noise to proteomics data is the high-frequent "wiggling" of the signal that leads to fast oscillations of the measurements about their correct values. But apart from these electric reasons, at least one further important noise component needs to be taken into account: Poisson-distributed shot noise that is a result of the fundamental discreteness of the events measured in the mass spectrometer (6). The relative importance of the different noise contributions for different kinds of mass spectra was recently investigated by Du et al. (7) with remarkable outcome: in their studies, the noise observed in the data could only be reproduced by a mixture of multinomial and Poisson-distributed effects.

Many of the techniques for noise removal are not specific to mass spectrometry, having their background in fields like image processing or audio analysis, and are hence not specifically adapted to the noise models relevant for proteomics data sets. Of those techniques, the most elementary ones are simple sliding averages. Here, each point in the spectrum is replaced by an (possibly weighted) average over its neighbouring points. A particularly important choice for the weights in the average are normal distributions, leading to the so-called *Gaussian smoothing*, where a Gaussian function is convolved with the data of interest to yield the denoised signal. A naïve application would result in an $O(n^2)$

technique (each point is replaced by the integral over all points in the spectrum), but it is easy to see that a much more efficient algorithm can be constructed: the convolution of two functions can be computed by Fourier-transforming both, multiplying them, and computing the Fourier backtransform. Since the Fourier transform of a Gaussian is again Gaussian, only with a different width, Gaussian smoothing can thus be computed by one Fourier-transform, one function multiplication, and one backtransform, leading to an O(n log(n)) algorithm.

The procedure of computing Gaussian filtering in Fourier space also leads to a different intuitive picture of the process that has proven very useful. In Fourier space, the smoothing operation corresponds to multiplying a Gaussian function with the Fourier transformed signal, effectively suppressing the Fourier components of high frequency. High frequency in Fourier space corresponds to fast variations in the untransformed signal and hence, Gaussian smoothing very directly implements what we want it to do, namely removing the rapidly oscillating noise terms from the data. On the other hand, the Gaussian multiplication changes all frequencies, not only the large ones, and it is by no means clear a-priori which frequencies are only related to noise and which contribute to the real signal of interest. Thus, a smoothing of the data will usually (a) not completely remove the noise components and (b) distort the signal of interest to a certain degree.

An alternative approach to smoothing that is not directly related to a sliding average computation but rather resembles what an experimentalist would do manually to arrive at a denoised spectrum can be found in a seminal publication by Savitzky and Golay (8). Here, the spectrum is replaced by an interpolating, locally polynomial curve that minimizes the least squares error of interpolant to real data points. Similarly, scatterplot smoothers like the popular locally weighted polynomial regression technique LOWESS (9) can be employed.

Finally, a third group of denoising methods that has received considerable attention lately are the so-called "Wavelet shrinkage techniques" (cf., e.g., the famous SURE shrinkage approach by Donoho and Johnstone (10)). Intuitively, the advantage of Wavelet based techniques (cf., Subheading 3.4) for denoising is that the signal is *locally* split into contributions of different length scales, principally allowing to remove noise components where they spoil the data but retaining large frequencies where they are needed to represent the real mass signal.

3.3. Peak Picking, Stick Conversion, and Feature Detection

In recent years, a large number of methods for feature detection or peak picking of proteomics data sets have been described in the literature. The methods proposed are too numerous by far to give a complete overview; for a few select examples, the reader is referred to, e.g., (11–18).

Obviously, before peaks can be converted into sticks or clustered into features (usually isotopic patterns), they need to be detected in the data. In the vast majority of cases, this detection is performed in the mass dimension only, i.e., even in the case of LC–MS, the subsequent mass scans are usually treated as single spectra that are combed for peaks individually instead of two dimensional searches in the whole LC–MS map. Assuming a sufficient prefiltering of the data (i.e. application of the baseline removal and denoising techniques discussed before), peak detection in the mass scans then in principle reduces to a search for local maxima in the scan. Of course, denoising cannot be expected to work perfectly, and hence, not each spike in the signal will correspond to a real mass peak. Thus, peak detection algorithms employ a number of techniques for improving their robustness, e.g. by using a sliding average of first and second derivative of the data to estimate the maxima in noisy data, or by only returning "significant" positions.

But while the simplification of picking peaks in one dimension only has a number of practical advantages, it neglects a wealth of information contained in the map that can be used to greatly improve significance and quality of detection: the finite retention profile of peptides leads to their appearance in several consecutive mass scans, providing some kind of implicit replicate measurement of the same molecule that can be used for noise suppression and signal amplification. Fortunately, this information can be leveraged even for one-dimensional techniques, since every single-scan (1D) method can be extended to whole MS maps (2D), either through the so-called *sweep-line paradigm* (cf., e.g., (19, 20)), through *peak alignment* strategies (cf., e.g., (21)) or through a *correlation*-based similarity measure between neighbouring scans (22, 23).

The peaks that have now been detected in the signal can then be subjected to a stick conversion routine. In the simplest case, stick conversion returns merely the position and intensity of the maximum of each detected peak. This approach is obviously rather unstable with respect to noise and, in addition, neglects the area under the peak in favour of its mere intensity. An alternative and more accurate, albeit more computationally demanding, approach consists in fitting functions describing the expected peak shape to those regions of the signal were peaks were detected. Common functions used to describe mass spectrometric peaks include the Gaussian, Lorentzian, and the *sech*^{2} form (24), but all of these share the same shortcoming for accurately modelling real-world data: they are exactly symmetric, while mass peaks usually feature some more or less pronounced tailing effects – the right shoulder of the peak is broader than the left one. One simple remedy that works well as long as no derivatives are required is to split the symmetric functions at their median and fit a left and a right half separately to the peak of interest. In some applications, though, more intricate methods like the double Gaussian models given in (25, 26), or the inherently asymmetric exponentially modified Gaussian (EMG) that explicitly models the tailing effect (27, 28), can yield much higher accuracy.

The result of peak picking and stick conversion is a list of mass-over-charge and intensity (or area) values, supplemented with the retention times at which the corresponding peak was measured. In most cases, though, this list will be spoiled with a large number of false positives: peak picking techniques can be easily made highly sensitive, but achieving sensible specificity is highly challenging, since quite often, noise and real data are very hard to distinguish. Peak detection techniques are thus prone to return large numbers of spurious mass values for each spike due to electric or chemical noise. Consequently, the resulting peaks are invariably filtered with respect to some likelihood measure, hopefully removing as many false positives while retaining as many true positives as possible. In the simplest case, the likelihood measure employed is merely the height of the peak, which is then compared to a global or local threshold. More intricate schemes combine several descriptors for the peak's quality, like the goodness of fit to the expected peak shape. The trade-off between sensitivity and specificity is by far the most troublesome aspect of signal processing in proteomics, and often leads to a neglect of lowly abundant peptides to achieve manageable false-positive rates. It is also the aspect that is most visible to the end user, since most program packages require him to set up a number of different parameters in order to tune the receiver operating characteristics to his demands, often with hardly foreseeable consequences.

Finally, after this filtering stage has been completed, individual peaks that have survived so far can be assigned into isotopic patterns, using, e.g., standard clustering techniques or dynamic programming algorithms based on the mass differences found in the list of sticks. Similarly, potential monoisotopic peaks with a mass position that violates the so-called *peptide-mass-rule* (cf., e.g., (1, 29, 30)), can be excluded since the decimal places of their mass indicate an atomic composition untypical for peptides. Finally, a number of goodness-of-fit measures to theoretically expected distributions can be used to estimate the likelihood that a given isotopic pattern was generated by a real peptide or whether it should rather be seen as the result of chemical noise.[3]

[3] This whole "de-isotoping" step is often seen as part of a later stage of the proteomics pipeline – the identification stage – since it usually operates not on the raw data, but on the list of sticks. However, as we will show in a later section, integrating de-isotoping (feature detection) into the signal processing can improve prediction performance by extracting further valuable information from the data that would otherwise be neglected.

3.4. Feature Detection with Implicit Parasitic Removal

The popular scheme of subsequent application of filtering steps before peak picking, stick conversion, and feature detection as described in the previous section has a number of clear practical advantages, the most important of which probably is its inherent modularity. In principle, each filter step can be implemented as an independent module with a standardized interface, so that pipelines of signal processing techniques can be freely combined and exchanged to yield programs specifically adapted to particular applications. If the algorithm designer knows up front, e.g., that the data will be highly resolved with only low degree of noise, he might opt for simpler denoising and peak detection schemes with lower computational load while for worse data sets, he might combine several intricate filters into a slow albeit more accurate setup.

But the very notion of repeatedly transforming the signal to prepare it for the genuine peak detection that leads to this modularity necessarily introduces a severe problem: each filtering step will, by definition, alter the "real" underlying signal, sometimes very significantly so. For example, most denoising techniques will distort the peak shapes and shift the maximum positions away from their real values. Hence, the more preprocessing steps are combined, the larger the risk that low abundant peptides are lost due to artificially introduced artefacts.

Recent years have seen the advent of a number of adapted methods especially designed for the analysis for mass spectrometric data that do not require explicit application of a series of filters before feature detection, but are rather capable of handling the parasitics implicitly.

The most important tools for this goal are the so-called *Wavelets*, which are – as suggested by their name – small wave-like functions. The Wavelet transform naturally generalizes the Fourier transform in that a signal is *locally* split into components of different frequencies. Since high-frequent noise and low-frequent baseline artefacts live on different frequency ranges than the signal of interest, a correctly designed Wavelet will automatically ignore spurious parasitics and rather focus on the "correct" frequency bands.

In (24) e.g., Lange et al. used the well-known Marr/Mexican-Hat Wavelet to pick peaks in MS data sets. While this approach has been shown to handle single peaks very accurately, the technique still requires an additional clustering step that combines picked peaks into isotopic patterns (cf. in contrast the Isotope Wavelet, Subheading 3.5). The Marr Wavelet has also been applied in (31), where – in contrast to Lange et al. – several scales of the wavelet transform are computed, leading to "ridges" at points of interest. Combining several Wavelet scales, this approach has the potential to be inherently more robust, but at larger computational cost. A different function was used by Carlson et al. (32), who employed an area-under-the-curve (AUC) filter function which is a hybrid between the simple *Haar* and the Marr Wavelet.

While the before-mentioned Wavelet-based methods use a *Continuous Wavelet Transform* (CWT), the technique presented in (33) features a multiresolution analysis (MRA) using *discrete Wavelet* functions. Here, a series of Wavelet transforms based on orthogonal basis functions is created.

A different approach for feature detection in LC–MS data sets is given by Andreev et al. (34). Using matched filter theory, the authors are able to design a so-called transfer function which maximizes the signal-to-noise ratio (S/N) in the chromatographic domain. A particularly nice feature of this method is the automatic adaption to the noise characteristics present in a particular data set. Mantini et al. (35) also make direct use of the second dimension by an independent component analysis. The method produces nearly no false positives, but often suffers from a low recall rate.

3.5. The Isotope Wavelet

As a general rule, signal processing techniques employ some kind of knowledge about the structure of the signal component of interest in order to separate information from background. In the case of MS, this knowledge has classically been the decomposition of the spectra into three components living at separate length scales or frequencies: peaks, baseline, and noise. But in the case of *proteomics* data, it turns out that we possess additional information about the structure of the wanted signal components: the characteristic pattern of the isotopic distributions for each peptide. While this information is usually employed in later stages of the proteomics pipeline in order to reduce the list of sticks to classes belonging to the same molecular fragments, it is only seldom used in signal processing frameworks. Some interesting examples demonstrating the use of isotopic information can be found in (36–38), where the *Poisson*[4] distributed envelope of isotope patterns is explicitly modelled using a template to simplify identification of the monoisotopic position.

Recently, the authors of this chapter have contributed a novel Wavelet-based technique that is specifically designed to exploit the full isotopic information while being "as simple as possible" for the end user: the combination of different filters and techniques into signal processing pipelines as outlined before often leads to a large number of parameters a user has to set to adapt the technique to his data. Careful usage of the isotope information, though, combined with a suitable Wavelet technique provides us with a method that is nearly self-adapting to the spectra, with only one easily interpretable free parameter in the one dimensional case. In addition, we find that using the isotope information allows to strongly increase sensitivity without sacrificing specificity: each peak greatly facilitates the detection of each other peak in the same pattern.

[4] In reality, the peak intensities rather follow a binomial distribution, but can be approximated by a Poisson distribution.

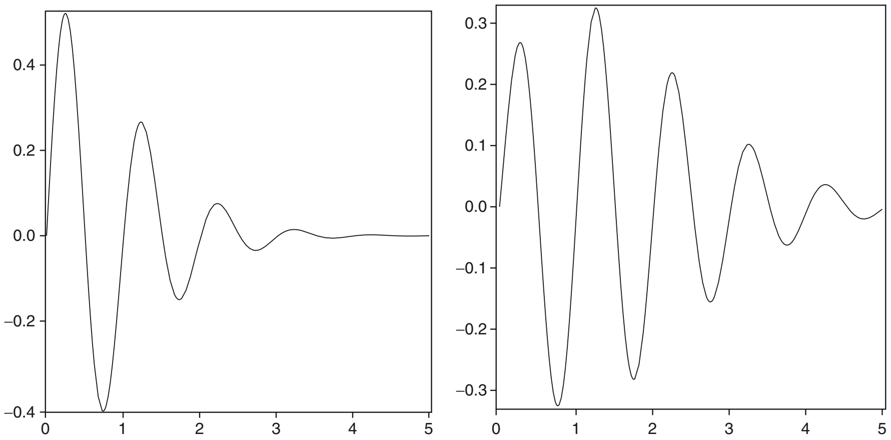


Fig. 3. The Isotope Wavelets corresponding to the isotopic patterns depicted in Fig. 1

This *Isotope Wavelet* (39) method hinges on a newly developed (adaptive[5]) Wavelet function that approximately transfers a signal into the space of isotopic patterns. The function is – by construction – robust against baseline as well as electrical and chemical noise artefacts. Thus, while standard Wavelets like the Marr Wavelet (*see* Subheading 3.4) have great value in being general tools in today's signal processing and can be used nearly independent of the type of the data, the Isotope Wavelet has been tailored to mass spectrometry and models explicitly the height distribution of peptidic patterns (cf. Figs. 1 and 3).

In this sense, the Isotope Wavelet works as a matched filter, shifting an adapted wavelet for each charge variant over the spectrum to compute its correlation with the underlying data. If the Wavelet thus traverses an isotopic pattern, the resulting correlation has a *sinc*-like[6] structure. Exploiting this structure, we have a robust criterion to separate noise (in particular chemical noise) from "real" data, which is supported by its superior feature detection capabilities in experiments on real-world-data (19, 20, 39).

3.6. Advanced Challenges

Since mass spectrometric signal processing is located at the borderline of techniques aiming to improve the quality of the raw data and of methods performing feature detection and characterization, there are many more challenges to be solved in the future.

Particularly important, in our opinion, is the question of statistically sound, stringent validation of signal processing techniques. This is a highly nontrivial question, since usually, we do not know the "correct" answers – how does a "correctly" denoised

[5] The adaptive Wavelet transform is a slight generalization of the classical Wavelet transform in that the Wavelet kernel can vary with position; hence, the transform does not correspond to a simple convolution, but rather to a more complicated integral transform.

[6] The *sinc*-function is defined by $sinc(x) := sin(x)/x$.

spectrum look like? What is the "real" area beneath a peak? Was a given spike in the signal due to a molecular fragment, or rather a noise effect? Hence, evaluation can currently only hinge on secondary properties, derived from the signal-processed spectra. A popular measure for the quality of peak picking techniques, e.g., is the coverage or score achieved when submitting the results to a database search engine like Mascot (40) or SEQUEST (41). But this can only provide crude estimates of the real error distribution. Significant progress in this field might be expected with the advent of highly accurate spectral simulators where the "true" answers are known.

A further problem not discussed in this chapter so far is the determination of the charge state, which is often done by a simple "look-around" in order to find the characteristic distances between peaks of the same isotopic cluster. More advanced methods use Fourier transforms, machine learning and additional information from MS/MS spectra to solve this problem (42–47).

Also, signal processing for MS/MS spectra warrants further studies, since it is essentially even harder than for single MS: the fragmentation patterns here show a less regular and less predictable isotopic distribution. In addition, many peaks that correspond to "real" fragments cannot be annotated, since the underlying fragmentation process is unknown.

Possibly, the most formidable challenge today, though, is the automatic detection and quantification of posttranslational modifications (PTMs) and of different labelling techniques (e.g. isotope labelling and stable isotope labelling with amino acids (SILAC) (48)). Traditionally, these modifications have been entirely associated with later stages of the proteomics pipeline, i.e., detection of modifications has been attempted on the stick-converted data sets. But a simple real-world example easily shows the relevance of PTMs (and analogously, labelling techniques) for the signal processing step: one well-known PTM often found in the data is the so-called *deamidation* that shifts the monoisotopic mass of the triggered signals by approximately 1 Dalton, and hence creates perfectly overlapping signals in the case that both versions of the peptide (deamidated and amidated) are present in the sample. Separating overlapping signals in both dimensions (m/z and RT) can be solved by pattern recognition or machine learning techniques (cf., e.g., (49)), but still poses a hard computational problem especially for low-abundant peptides, whose signals are often affected by noise artefacts.

4. Conclusions

In this chapter, we have presented the main goals and challenges of signal processing in computational proteomics and sketched some of the currently available solutions, standard image processing techniques applied to proteomics data as well as hand-tailored methods. So far, the problems described have not been solved in their entirety, and each method features its own advantages and shortcomings. In particular, the arguably most flexible strategy of using specific techniques for each individual task of the whole signal processing pipeline comes at a cost: signals of low-abundant peptides can be shifted or damaged by smoothing or denoising operations, especially if resampling techniques are involved with low-resolution spectra. Here, methods such as Wavelets, for example, which are inherently capable of implicit removal of parasitics, can help overcome these problems.

From a purely practical point of view, a problem that is often underestimated in the more theoretical research but that severely restricts the applicability of many advanced techniques in real-world situations, is the number and interpretation of the free parameters introduced by the algorithms. Faced with a large number of possibly undocumented switches and knobs, the end user will often decide to leave everything at its default value which might be entirely unsuited to the kind of data at hand.

In our opinion, it will thus be one of the main challenges of the future development of signal processing in computational proteomics to provide algorithms for feature detection that are not only highly accurate, able to cope with the diverse kinds of labelling and modifications, and efficient enough to support large-scale high-throughput experiments, but that are also easily applicable, preferably self-adapting to the data at hand. While significant progress along this direction has been made in recent years, the gap between what is possible in principle and what is used in practice (in the wet-lab, e.g.) is still opening further, and we are convinced that harvesting the theoretical advances in signal processing for real-world applications will push the boundaries of proteomics far beyond the current state of the art.

Acknowledgments

The authors would like to express their gratitude to Mrs. Anna Katharina Dehof, Mrs. Sophie Weggler, and Mrs. Linda Wolters for critical reading of the manuscript.

References

1. Gay, S., Binz, P. A., Hochstrasser, D. F., Appel, R. D. (1999) Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis* **20**, 3527–34.
2. Bocker, S., Makinen, V. (2008) Combinatorial approaches for mass spectra recalibration. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5**, 91–100.

3. Kolibal, J., Howard, D. (2006) MALDI-TOF baseline drift removal using stochastic Bernstein approximation. *Eurasip Journal on Applied Signal Processing* **1**, 61.
4. Sauve, A. C., Speed, T. P. (2004) Normalization, baseline correction and alignment of high-throughput mass spectrometry data. In: *Proceedings of the Genomic Signal Processing and Statistics workshop*; 26–7.
5. Williams, B., Cornett, S., Crecelius, A., Caprioli, R., Dawant, B., Bodenheimer, B. (2005) An algorithm for baseline correction of MALDI mass spectra. In: *ACM Southeast Regional Conference: ACM Proceedings.*
6. Shin, H., Koomen, J., Baggerly, K., Markey, M. (2004) Towards a Noise Model of MALDI TOF Spectra. In: *American Association for Cancer Research (AACR) Advances in Proteomics in Cancer Research, Waikoloa.*
7. Du, P. C., Stolovitzky, G., Horvatovich, P., Bischoff, R., Lim, J., Suits, F. (2008) A noise model for mass spectrometry based proteomics. *Bioinformatics* **24**, 1070–7.
8. Savitzky, A., Golay, M. J. E. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **36**, 1627.
9. Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–36.
10. Donoho, D. L., Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–24.
11. Wehofsky, M., Hoffmann, R. (2002) Automated deconvolution and deisotoping of electrospray mass spectra. *Journal of Mass Spectrometry* **37**, 223–9.
12. Hoopmann, M. R., Finney, G. L., MacCoss, M. J. (2007) High-speed data reduction, feature detection and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Analytical Chemistry* **79**, 5620–32.
13. Gambin, A., Dutkowski, J., Karczmarski, J., Kluge, B., Kowalczyk, K., Ostrowski, J., Poznanski, J., Tiuryn, J., Bakun, M., Dadlez, M. (2007) Automated reduction and interpretation of multidimensional mass spectra for analysis of complex peptide mixtures. *International Journal of Mass Spectrometry* **260**, 20–30.
14. Kaur, P., O'Connor, P. B. (2006) Algorithms for automatic interpretation of high resolution mass spectra. *Journal of the American Society for Mass Spectrometry* **17**, 459–68.
15. Horn, D. M., Zubarev, R. A., McLafferty, F. W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry* **11**, 320–32.
16. Mantini, D., Petrucci, F., Pieragostino, D., Del Boccio, P., Di Nicola, M., Di Ilio, C., Federici, G., Sacchetta, P., Comani, S., Urbani, A. (2007) LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *Bmc Bioinformatics* **8**, 101.
17. Noy, K., Fasulo, D. (2007) Improved model-based, platform-independent feature extraction for mass spectrometry. *Bioinformatics* **23**, 2528–35.
18. Samuelsson, J., Dalevi, D., Levander, F., Rognvaldsson, T. (2004) Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics* **20**, 3628–35.
19. Schulz-Trieglaff, O., Hussong, R., Gröpl, C., Hildebrandt, A., Reinert, K. (2007) A fast and accurate algorithm for the quantification of peptides from mass spectrometry data. In: *Research in computational molecular biology*, Springer; 473–87.
20. Schulz-Trieglaff, O., Hussong, R., Gropl, C., Leinenbach, A., Hildebrandt, A., Huber, C., Reinert, K. (2008) Computational quantification of peptides from LC-MS data. *Journal of Computational Biology* **15**, 685–704.
21. Yu, W. C., He, Z. Y., Liu, J. F., Zhao, H. Y. (2008) Improving mass spectrometry peak detection using multiple peak alignment results. *Journal of Proteome Research* **7**, 123–9.
22. Muddiman, D. C., Rockwood, A. L., Gao, Q., Severs, J. C., Udseth, H. R., Smith, R. D., Proctor, A. (1995) Application of sequential paired covariance to capillary electrophoresis electrospray-ionization time-of-flight mass-spectrometry – unraveling the signal from the noise in the electropherogram. *Analytical Chemistry* **67**, 4371–5.
23. Fleming, C. M., Kowalski, B. R., Apffel, A., Hancock, W. S. (1999) Windowed mass selection method: a new data processing algorithm for liquid chromatography-mass spectrometry data. *Journal of Chromatography A* **849**, 71–85.
24. Lange, E., Gröpl, C., Reinert, K., Kohlbacher, O., Hildebrandt, A. (2006) High-accuracy peak picking of proteomics data using wavelet techniques. In: *Pac Symp Biocomput*; 243–54.
25. Strittmatter, E. F., Rodriguez, N., Smith, R. D. (2003) High mass measurement accuracy determination for proteomics using multivariate regression fitting: application to electrospray

ionization time-of-flight mass spectrometry. *Analytical Chemistry* **75**, 460–8.

26. Kempka, M., Sjodahl, J., Bjork, A., Roeraade, J. (2004) Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **18**, 1208–12.
27. Di Marco, V. B., Bombi, G. G. (2001) Mathematical functions for the representation of chromatographic peaks. *Journal of Chromatography A* **931**, 1–30.
28. Zubarev, R. A., Hakansson, P., Sundqvist, B. (1996) Accurate monoisotopic mass measurements of peptides: possibilities and limitations of high resolution time-of-flight particle desorption mass spectrometry. *Rapid Communications in Mass Spectrometry* **10**, 1386–92.
29. Wool, A., Smilansky, Z. (2002) Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting. *Proteomics* **2**, 1365–73.
30. Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., Yates, J. R., 3rd (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Analytical Chemistry* **75**, 2470–7.
31. Du, P., Kibbe, W. A., Lin, S. M. (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* **22,** 2059–65.
32. Carlson, S. M., Najmi, A., Whitin, J. C., Cohen, H. J. (2005) Improving feature detection and analysis of surface-enhanced laser desorption/ionization-time of flight mass spectra. *Proteomics* **5**, 2778–88.
33. Randolph, T. W., Yasui, Y. (2006) Multiscale processing of mass spectrometry data. *Biometrics* **62**, 589–97.
34. Andreev, V. P., Rejtar, T., Chen, H. S., Moskovets, E. V., Ivanov, A. R., Karger, B. L. (2003) A universal denoising and peak picking algorithm for LC–MS based on matched filtration in the chromatographic time domain. *Analytical Chemistry* **75**, 6314–26.
35. Mantini, D., Petrucci, F., Del Boccio, P., Pieragostino, D., Di Nicola, M., Lugaresi, A., Federici, G., Sacchetta, P., Di Ilio, C., Urbani, A. (2008) Independent component analysis for the extraction of reliable protein signal profiles from MALDI-TOF mass spectra. *Bioinformatics* **24**, 63–70.
36. Gras, R., Muller, M., Gasteiger, E., Gay, S., Binz, P. A., Bienvenut, W., Hoogland, C., Sanchez, J. C., Bairoch, A., Hochstrasser, D. F., Appel, R. D. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* **20**, 3535–50.
37. Breen, E. J., Hopwood, F. G., Williams, K. L., Wilkins, M. R. (2000) Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis* **21**, 2243–51.
38. McIlwain, S., Page, D., Huttlin, E. L., Sussman, M. R. (2007) Using dynamic programming to create isotopic distribution maps from mass spectra. *Bioinformatics* **23**, I328–I36.
39. Hussong, R., Tholey, A., Hildebrandt, A. (2007) Efficient analysis of mass spectrometry data using the isotope wavelet. In: Arno, P. J. M. S., Michael, R. B., Robert, C. G., Ad, J. F., editors *CompLife*. (AIP) American Institute of Physics http://proceedings.aip.org/proceedings/Melville, NY; 139–49.
40. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–67.
41. Eng, J. K., McCormack, A. L., Yates, J. R. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**, 976–89.
42. Tabb, D. L., Shah, M. B., Strader, M. B., Connelly, H. M., Hettich, R. L., Hurst, G. B. (2006) Determination of peptide and protein ion charge states by Fourier transformation of isotope-resolved mass spectra. *Journal of the American Society for Mass Spectrometry* **17**, 903–15.
43. Sadygov, R. G., Hao, Z., Huhmer, A. F. R. (2008) Charger: combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra. *Analytical Chemistry* **80**, 376–86.
44. Na, S., Paek, E., Lee, C. (2008) CIFTER: automated charge-state determination for peptide tandem mass spectra. *Analytical Chemistry* **80**, 1520–8.
45. Colinge, J., Magnin, J., Dessingy, T., Giron, M., Masselot, A. (2003) Improved peptide charge state assignment. *Proteomics* **3**, 1434–40.
46. Chen, L., Yap, Y. L. (2008) Automated charge state determination of complex isotope-resolved mass spectra by peak-target Fourier transform. *Journal of the American Society for Mass Spectrometry* **19**, 46–54.
47. Klammer, A. A., Wu, C. C., MacCoss, M. J., Noble, W. S. (2005) Peptide charge state determination for low-resolution tandem mass spectra. In: *IEEE Computational Systems Bioinformatics Conference*. IEEE Computer Society.

48. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics* **1**, 376–86.

49. Du, P. C., Angeletti, R. H. (2006) Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. *Analytical Chemistry* **78**, 3385–92.

50. Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., Kohlbacher, O. (2008) OpenMS – an open-source software framework for mass spectrometry. *BMC Bioinformatics* **9**, 163.

Chapter 12

A High-Performance Reconfigurable Computing Solution for Peptide Mass Fingerprinting

Daniel Coca, Istvan Bogdan, and Robert J. Beynon

Abstract

High-throughput, MS-based proteomics studies are generating very large volumes of biologically relevant data. Given the central role of proteomics in emerging fields such as system/synthetic biology and biomarker discovery, the amount of proteomic data is expected to grow at unprecedented rates over the next decades. At the moment, there is pressing need for high-performance computational solutions to accelerate the analysis and interpretation of this data.

Performance gains achieved by grid computing in this area are not spectacular, especially given the significant power consumption, maintenance costs and floor space required by large server farms.

This paper introduces an alternative, cost-effective high-performance bioinformatics solution for peptide mass fingerprinting based on Field Programmable Gate Array (FPGA) devices. At the heart of this approach stands the concept of mapping algorithms on custom digital hardware that can be programmed to run on FPGA. Specifically in this case, the entire computational flow associated with peptide mass fingerprinting, namely raw mass spectra processing and database searching, has been mapped on custom hardware processors that are programmed to run on a multi-FPGA system coupled with a conventional PC server. The system achieves an almost 2,000-fold speed-up when compared with a conventional implementation of the algorithms in software running on a 3.06 GHz Xeon PC server.

Key words: Field programmable gate array (FPGA), Peptide mass fingerprinting (PMF), Mass spectrometry, Reconfigurable computing, Proteomics

1. Introduction

Following significant advances in mass spectrometry instrumentation, modern mass spectrometers can carry out routine analysis of minute amounts (femtomoles) of complex peptide mixtures and generate mass spectrometric data at speeds far greater than the time required for processing it. As current mass spectrometers boast acquisition rates of up to 200 spectra per second, post-instrument data processing is the major bottleneck in proteomics workflow.

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_12, © Humana Press, a part of Springer Science+Business Media, LLC 2010

As experimental design, instrument performance and user skills increase, it is now feasible to contemplate the analysis of a substantial proteome in a single experiment that can generate many gigabytes of raw mass spectrometric data.

In this context, high-performance computing resources are essential to address the current analysis bottlenecks, thus allowing scientists to deal efficiently with the large amounts of proteomics data that are being generated.

Although computational grid resources could be useful for solving computational tasks in proteomics, the performance gains recently reported in the literature (1) for the grid implementation of a BLAST search algorithm (60-fold speed increase using 600 CPU's) are far from spectacular, given the resources allocated for the task.

At the same time, processing of mass spectrometric data is ideally performed "near-instrument", wherein the end user has the option of adjusting the search strategy according to results obtained in real-time. However, computation platforms based on low number of processors are unlikely to deliver the speed that will be required.

In this situation, relying on grid computing or dedicated computer clusters does not make sense unless there are batches of hundred or thousand of mass spectra to process. Moreover, dedicated high-performance computer clusters require a significant amount of infrastructure to deal with interconnectivity and power dissipation. It has been argued (2) that more efficient high-performance computing solutions are necessary to mitigate the costs of housing and powering the next generation petascale and larger high performance computer systems, which are expected to be prohibitive for many institutions and programs.

Although in recent years manufacturers have attempted to address limitations of the conventional microprocessor architecture by placing several multiprocessor cores on the same die, it is difficult to envisage a fixed "general-purpose" multi-core architecture that would deliver the required performance gains required across the entire range of computationally challenging problems in proteomics and other fields.

This article advocates reconfigurable computing as an alternative approach to conventional high-performance computing, focusing on a specific bioinformatics problem in proteomics namely protein mass fingerprinting.

Reconfigurable computers consist of a standard microprocessor system coupled with hardware processors whose circuitry can be programmed (and re-programmed) according to the algorithm that is being run.

Historically, the idea of reconfigurable computing originated in the 1960's when Estrin in, a landmark paper (3, 4), proposed the concept of a computer made of a standard processor and an array of "reconfigurable" hardware.

Reconfigurable computing became a reality in the 1990s with the advent of high-density Field Programmable Gate Arrays

(FPGAs), which are de-facto the reconfigurable processors in almost all current reconfigurable computing platforms. Modern FPGA's can be programmed to implement and run a custom digital hardware design with the same flexibility and ease as a conventional computer program (see Note 1).

To fully understand the significant advances made in this area, the reader is referred to the excellent books (5, 6) which are amongst the first comprehensive surveys and tutorials in the field of FPGA-based reconfigurable computing.

In biocomputation, early applications of FPGA devices addressed the gene sequence analysis problem (7) and have been successfully employed to speed-up DNA sequencing algorithms (8–13). FPGAs were also used in the attempt to accelerate search of substrings similar to a template in a proteome (14). A multiple sequence alignment solution implemented in FPGA hardware is also available (15).

FPGAs have been used to accelerate sequence database searches with MS/MS-derived query peptides (16). FPGA-accelerated BLAST search algorithms are available and have been used, for example, to perform EST sequencing (17). More recently, the Aho-Corasick string set matching algorithm was implemented in FPGA hardware and used for matching peptide sequences against a genome translated in six reading frames (18).

This paper describes the development of a reconfigurable computing solution for peptide mass fingerprinting. It provides an introduction to the field of reconfigurable computing highlighting the major concepts and issues that the designer has to master in order to fully exploit the power of this technology. The results further demonstrate the applicability of reconfigurable computing to computational bottlenecks faced by MS-based proteomics.

2. Materials

2.1. FPGA Devices

Reconfigurable computers are based almost exclusively on reprogrammable SRAM-based FPGA devices. These devices are credited for re-igniting and enabling the reconfigurable computing revolution.

Since their introduction in 1985, field-programmable gate arrays have continuously expanded their use from being the ultimate prototyping platform and providing basic "glue logic" functionality to being at the heart of complex digital systems in a wide range of application areas ranging from telecommunication, automotive, aerospace and defence to biomedical and high-performance computing.

The remarkable success of these devices is attributed to the inherent advantages offered by the parallel programmable architecture, which allows designers exploit algorithm and instruction-level parallelism to accelerate computations and to add or modify features and functionality provided by an existing FPGA-based system by reconfiguring the device. The tremendous increase in gate densities and lowering of unit costs and the development of more sophisticated and user-friendly design tools have also been determining factors to skyrocketing demand for FPGAs in recent years.

An FPGA device is an integrated circuit that contains an array of programmable resources, the most important of which are programmable logic blocks, input/output (I/O) blocks and interconnects. For the majority of current FPGA devices, the logic, interconnect and I/O configuration (which determines the algorithm that is being implemented by a device) is stored in an on-chip SRAM and therefore these FPGA devices are re-programmable.

The elementary programmable logic block (or slice) contains programmable combinatorial logic structures (function generators), typically implemented using look-up tables (LUT), sequential logic elements i.e. flip-flops or latches and dedicated carry logic for fast implementation of arithmetic addition and subtraction. Modern FPGA devices provide significant flexibility in the configuration of the logic block. For example, each 4-input function generator in the Xilinx Virtex II devices can be configured to implement either a 4-input LUT, 16 bits of distributed synchronous RAM or a 16-bit shift register. To reduce the need for programmable routing most FPGA architectures, combine two or more logic slices into a cluster using fast interconnections. As shown in Fig. 1, each slice is connected to a switch matrix that provides access to the general routing matrix. While Xilinx calls such a cluster a Configurable Logic Blocks (CLBs), Altera uses the term Logic Array Block (LAB).

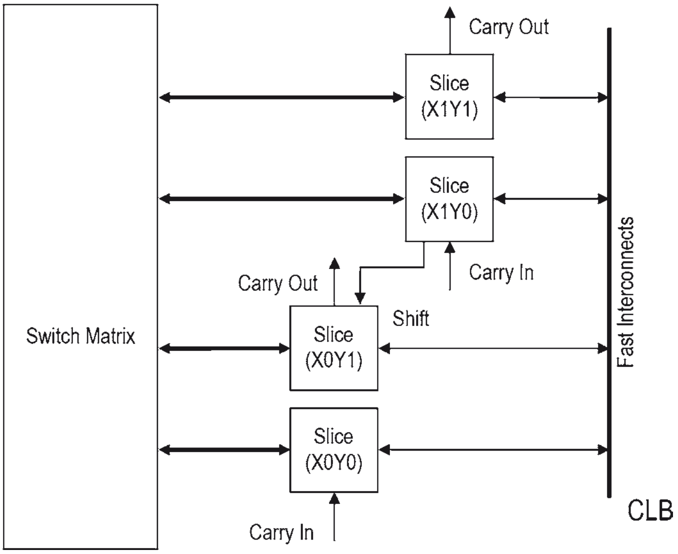


Fig. 1. Virtex-II configurable CLB (source: adapted from (27), Figure 14, Module 2, p.12)

Another class of programmable resources are the Input/Ouput Blocks (IOBs) that provide the interface between the package pins and the internal configurable logic. The IOBs of modern FPGA devices offer a wide range of features including support for most common signalling standards, digitally controlled impedance to prevent reflections and maintain signal integrity, double-data-rate (DDR) transfer (data is transferred on the rising as well as on the falling edge of the clock signal) and optional delay elements.

Finally, programmable routing resources allow interconnecting all the other programmable elements on an FPGA such as logic clusters and I/O blocks. In a Xilinx FPGA device, each programmable element is connected to a switch matrix allowing multiple connections with the general routing matrix. Essentially, as illustrated in Fig. 2, the FPGA device can be viewed as an array of switch matrices each attached to a logic block. Similar to all the

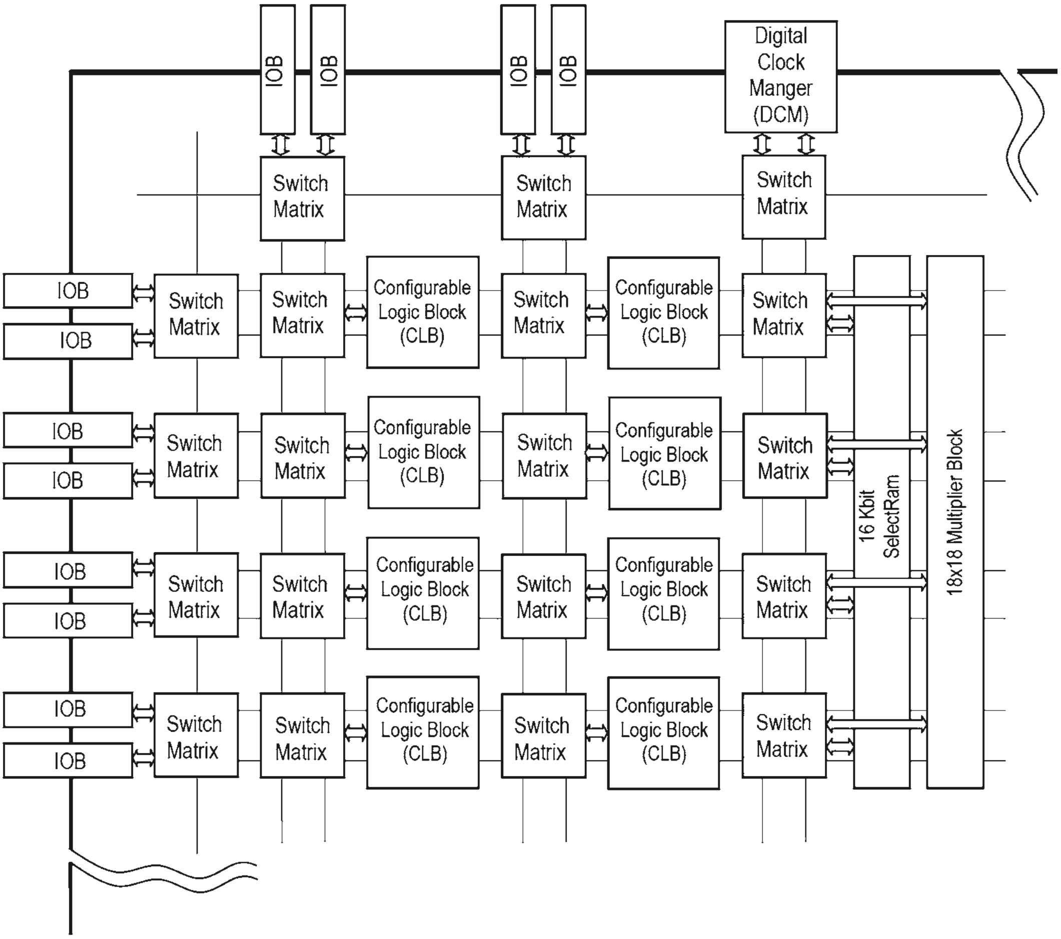


Fig. 2. Virtex-II routing resources (source: adapted from (27), Figure 48, Module 2, p.32)

other programmable elements, the routing configuration is defined by values stored in static memory cells during initialization.

The programmable logic elements described so far are the basic building blocks used to implement an algorithm in hardware. Modern FPGA's, however, have evolved to include specialized programmable blocks such as embedded RAM, dedicated DSP structures, embedded microprocessors, system monitoring functions, digital clock managers and fast serial transceivers. Because of the widening spectrum of applications for FPGA devices and the difficulty in creating a single FPGA architecture which satisfies all users, in recent years, device manufacturers such as Xilinx have followed a new strategic route of creating a family of FPGA platforms that have been optimized for particular application domains. For example, Xilinx Virtex-5 family includes four product types that are optimized for high-performance general logic applications, high-performance logic with advanced serial connectivity, high-performance signal processing applications with advanced serial connectivity and high-performance embedded systems with advanced serial connectivity. This enables users to select the FPGA device that incorporates the most appropriate collection of programmable resources for a particular design.

As the SRAM-based FPGA manufacturers are amongst the earlier adopters of new digital-CMOS manufacturing processes, in recent years, FPGAs have advanced at a faster pace than microprocessors, the latest devices offering unprecedented performance and density gains with speeds on average 30% faster and a logic capacity 65% greater than previous generations. The latest devices have as many as 1.2 billion transistors and allow the implementation of few thousand conventional microcontrollers on a single FPGA chip.

2.2. Reconfigurable Computing Platform

The hardware processors described in this paper were implemented on a commercial off-the-shelf (COTS) multi-FPGA reconfigurable hardware platform, consisting of a BenNuey FPGA motherboard from Nallatech Ltd. (www.nallatech.com) communicating with the host PC server (Xeon 3.06 MHz processor and 4 GB RAM) via a PCI interface (32 bits, 33 MHz) (see Note 2). The motherboard is equipped with one FPGA for user designs (Xilinx Virtex-II XC2V8000) and 4 Mbytes on-board RAM. In our application, this FPGA has been used to implement the mass spectra processor.

A second, smaller FPGA (Xilinx Spartan-II) on the motherboard is programmed to handle the communication between the PC server and the FPGA system.

The motherboard can be configured to hold three additional FPGA modules that can be plugged into dedicated motherboard

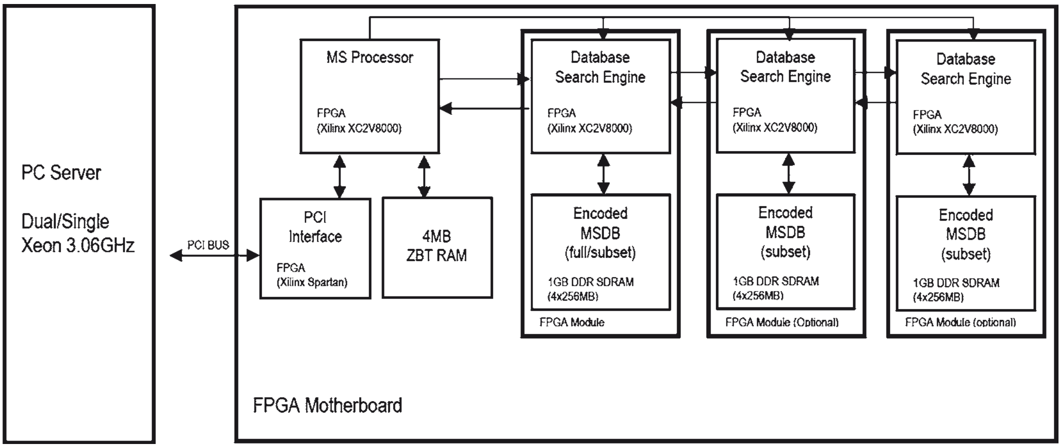


Fig. 3. Block diagram of the Nallatech FPGA system

slots. At present, only one additional module (BenData DD) has been used to implement the database search. The FPGA module has one user FPGA device (Virtex-II XC2V8000) and 1 GB of DDR SDRAM memory that can hold the entire encoded MSDB protein database. The current version of the encoded database occupies approximately 680 MB. Each module is connected with the motherboard FPGA and with the other two modules via a 64 bit, 66 MHz local bus. This architecture enables the implementation of parallel searches at FPGA level as well as across modules. The block diagram of the FPGA system is shown in Fig. 3.

The FPGA board was installed and tested on Single and Dual 3.06 GHz Xeon processor servers with 4 GB RAM under Windows XP Professional.

2.3. Development Tools

As the demand for System-on-Chip solutions has rocketed, in recent years the high-level design tools have evolved and diversified dramatically, effectively dismantling most of the barriers between algorithm development and hardware implementation (see Note 3). For those familiar with Matlab and Simulink, Xilinx System Generator Toolbox provides an ideal graphical design environment. This was the main design tool adopted for this application. This design tool generates synthesizable hardware (see Note 4) and allows clock-accurate simulation of the designs as well as real-time verification on the actual FPGA hardware using hardware-in-the-loop simulation. Because the software comes with a large library of highly-optimized Intellectual Property (IP) blocks, System Generator generate highly efficient designs capable of running at high speeds. The software generates highly optimized HDL code including testbench and test vectors (see Note 5) as well as required project files for the most popular logic synthesis tools. For this application, we used Xilinx Synthesis Technology (XST) tool which is part of Xilinx's ISE Foundation software.

2.4. MS Data Generation

To create the raw data used to evaluate the FPGA implementation, single proteins were diluted with 50 mM ammonium bicarbonate and digested with trypsin at a ratio of protein: enzyme of 50:1. Digestion was carried out at 37°C for 24 h after which time, 1 μl digested material was spotted onto a MALDI target. This was mixed with 1 μl α-cyano hydroxycinnamic acid matrix and analyzed using a Micromass M@LDI mass spectrometer (Waters, Manchester, UK) typically over the *m/z* range 800–4000.

3. Methods

Peptide Mass Fingerprinting (PMF) is an established technique to identify proteins on the basis of the analysis of a subset of constituent peptide fragments, generated by proteolytic digestion. The approach is predicated on the assumption that a pattern of proteolytic peptide masses provide a quasi-unique signature for every protein in the database.

The peptide mixtures resulting from proteolysis are typically analysed by MALDI-TOF (matrix-assisted laser desorption/ionization time-of-flight) mass spectrometers. Specific algorithms are required to convert the raw mass spectrometric data into a "peak list" consisting of monoisotopic peptide masses and intensities. A subset of the experimentally generated "peak list", so called "peptide mass fingerprint", is then compared with theoretical proteolytic peptide maps derived from protein sequence database by *in silico* digestion.

The computations associated with the PMF approach can therefore be divided into two distinct processing stages: (a) processing the raw MALDI-TOF mass spectrometric data to extract a peptide mass fingerprint and (b) using the protein signature to search for a match in a comprehensive database of known proteins.

Both stages of computation have been implemented (see Note 6) as dedicated hardware processors (19, 20). The complete FPGA-hardware solution for peptide mass fingerprinting, which incorporates a raw mass spectra processor and a parallel search engine, is presented in the following sections.

3.1. Mass Spectra Processing Algorithms

Following specific protein digestion, a MALDI-TOF mass spectrometer generates pairs of mass-to-charge (*m/z*) and abundance values $d_k = (x_k, y_k)$, $(k = 1, 2, \ldots, N)$. Typically, the number of points (N) in the spectrum ranges from a few thousand to a few hundred thousand. The determination of experimental monoistotopic peptide masses requires processing of the raw mass spectrum in order to discriminate between spectral peaks that correspond to digested peptides and associated isotopes and the spurious peaks caused by noise and sample contamination.

The spectra processing algorithm adopted for FPGA implementation is based largely on the algorithm proposed in (21). The major difference is the method used to implement in hardware the aggregation of natural isotopomers (due primarily to the natural abundance of ^{13}C and ^{15}N), which is based on a Poisson distribution approximation of the isotopic patterns for every peptide (22). The FPGA algorithm also implements an additional smoothing function (Savitzky-Golay) which is optional.

It should be emphasized that the algorithm adopted is not the only one available to perform peak extraction. The algorithm was chosen because it is computationally efficient and well suited for implementation as a deep computational pipeline.

In practice, the tractability of the algorithm to parallelization or pipelining is crucial to maximizing performance gains through implementation in FPGA hardware.

Specifically, the processing steps can be summarized as follows:

1. *Smoothing (Optional)*. Performing Savitzky–Golay smoothing (23) over the raw input spectrum can reduce the effect of instrumentation noise. The algorithm is based on performing a least squares linear regression fit of a polynomial of degree M over at least $M+1$ data points around each point in the spectrum. The main advantage of this procedure is that it tends to preserve the shape of the signal peaks. The smoothing operation is implemented as a standard FIR filter

$$ys_i = \sum_{j=1}^{F} b_j^M \, y_{1-j+1}$$

Where F is the size of the smoothing window ($F=2q+1$), y is the input data stream, ys is the FIR output and b_j^M are the time-varying filter coefficients. For a given filter of order M and (odd) frame size F ($F>M+1$), all the coefficients needed to implement the smoothing operation form a $F\times F$ matrix $[b_{i,j}^M]_{i,j}^F=1$.

The Savitzky–Golay smoothing operation can be represented in matrix form as follows:

$$\begin{bmatrix} ys_1 \\ ys_2 \\ \cdots \\ ys_q \end{bmatrix}^T = \begin{bmatrix} b_{2q+1,1} & b_{2q+1,2} & \cdots & b_{2q+1,2q+1} \\ b_{2q,1} & b_{2q,2} & \cdots & b_{2q,2q+1} \\ \cdots & \cdots & \cdots & \cdots \\ b_{q+2,1} & b_{q+2,2} & \cdots & b_{q+2,2q+1} \end{bmatrix} \begin{bmatrix} y_{2q+1} \\ y_{2q} \\ \cdots \\ y_1 \end{bmatrix}$$

$$\begin{bmatrix} ys_{q+1} \\ ys_{q+2} \\ \cdots \\ ys_{N-q} \end{bmatrix}^T = \begin{bmatrix} b_{q+1,2} \\ b_{q+1,2} \\ \cdots \\ b_{q+1,2q+1} \end{bmatrix}^T \begin{bmatrix} y_{2q+1} & y_{2q+2} & \cdots & y_N \\ y_{2q} & y_{2q+1} & \cdots & y_{N-1} \\ \cdots & \cdots & \cdots & \cdots \\ y_1 & y_2 & \cdots & y_{N-2q} \end{bmatrix}$$

$$\begin{bmatrix} ys_{N-q+1} \\ ys_{N-q+2} \\ \cdots \\ ys_N \end{bmatrix}^T = \begin{bmatrix} b_{q,1} & b_{q,2} & \cdots & b_{q,2q+1} \\ b_{q-1,1} & b_{q-1,2} & \cdots & b_{q-1,2q+1} \\ \cdots & \cdots & \cdots & \cdots \\ b_{1,1} & b_{1,2} & \cdots & b_{1,2q+1} \end{bmatrix} \begin{bmatrix} y_N \\ y_{N-1} \\ \cdots \\ y_{N-2q} \end{bmatrix}$$

The smoothing window F can be chosen according to instrument resolution setting (number of data points recoded per 1 m/z unit). For processing the raw spectra used in this paper, filters of order $M = 11$ and a frame length $F = 23$ were found to produce best results.

2. *Baseline and Noise Detection*. Baseline estimation is performed in order to correctly evaluate the magnitude of the peaks. Peak magnitude data is essential for performing de-isotoping. An estimate of the noise (which can be chemical or instrument noise) is also computed and used to perform segmentation.

   The raw spectrum mass list $\{x_k | k = 1, 2, \ldots, N\}$ is divided into small intervals (m_i) of width ω. For each interval, the local minimum (Z_i) and maximum (Υ_i) abundances, and their differences (W_i) are computed as follows

   $$Z_i = \min(\Upsilon x \in m_i)\ \Upsilon_i = \max(\Upsilon x \in m_i)\ W_i = \Upsilon_i - Z_i,\ i = 1, \ldots, [\mathrm{N}/\omega]$$

   where (x, y) are m/z, abundance pairs.

   For each integer mass in the spectrum (x_j, $j = 1, 2, \ldots, k$), a symmetric window (M_j) of width Ω (Ω >> ω) is placed around and baseline and noise levels are estimated as follows:

   $$\Upsilon_{\mathrm{base}}(j) = \sum_{i=j(-)}^{j(+)} w_i Z_i;\ \Upsilon_{\mathrm{noise}}(j) = \sum_{i=j(-)}^{j(+)} w_i \Upsilon_i;\ w_i = \frac{\frac{1}{W_i^2}}{\sum_{p=j(-)}^{j(+)} \frac{1}{W_p^2}}$$

   where $j(-)$ is the index of the leftmost sub-interval (m_i) covered by M_j, $j(+)$ is the index of the rightmost sub-interval (m_i) covered by M_j (21). Subsequently, the signal to noise ratio (s_k) is computed for each spectral point d_k, $k = 1, 2, \ldots, N$

   $$S_k = \frac{y_k - \Upsilon_{\mathrm{base}}(j_k)}{\Upsilon_{\mathrm{noise}}(j_k) - \Upsilon_{\mathrm{base}}(j_k)},\ j_k \lfloor m_k \rfloor.$$

3. *Spectrum Segmentation*. Using the signal-to-noise information, the spectrum is classified into three sets: noise ($d_k \in$

D_{noise}, $s_k < 1$), support ($d_k \in D_{\text{support}}$, $1 \le s_k < \text{SN}$) and signal ($d_k \in D_{\text{signal}}$, $\text{SN} \le s_k$), where SN is a user adjustable threshold.

4. *Peak Detection*. Peaks are constructed from data points that are signal or support points and are bounded by noise.
   A peak is defined as a set of data points

   $P = \{d_j \mid x_j < x_{j+1}, j = 1,\ldots,p, d_0 \in D_{\text{noise}}, d_{p+1} \in D_{\text{noise}}, d_j \in D_{\text{signal}}$ or $d_j \in D_{\text{support}}\}$.

For every peak, center of the mass (m_p) and abundance (a_p) are computed as follows:

$$mp = \frac{\sum_{j=1}^{p} x_j y_j}{\sum_{j=1}^{p} y_j} \quad ap = \max_{j \in P}(y_j - Y_{base}(j)).$$

5. *Clustering*. Involves grouping valid peaks that are about 1 m/z apart, into clusters. Two peaks p and q are in the same cluster if $1 - \tau < |m_p - m_q| < 1 + \tau$, where τ is a user-defined parameter (typically $\tau = 0.2$). A valid cluster has at least one peak with a data point d_k in D_{signal}.
6. *Deisotoping* involves determining the monoisotopic peptide masses from overlapping isotopic distributions (due primarily to the natural abundance of ^{13}C and ^{15}N). The algorithm uses Poisson distributions to approximate the isotopic patterns for every peptide (22).

   The expected proportional abundance of the heavier isotopes $m_p(i)$, $i = 1,2,\ldots$ with respect to the monoisotopic peaks $m_p = m_p(0)$ are computed as follows

$$E(i, m_P) = \frac{a_P F(m_P)^i}{i!}, i = 1, 2, \ldots; E(i, m_P) > 0,$$
$$F(m_P) = 0.000594 m_P - 0.03091,$$

where $E(1, m_p)$ is the theoretical abundance of the first isotope of m_p, $E(2, m_p)$ is the effect of the second isotopic contribution etc. The best-fit Poisson models of isotopic distributions are shown to match those of theoretical distributions (22).

Deisotoping a cluster of R consecutive peaks is summarized below:

- The first peak of the cluster is always considered a monoisotope.
- Compute the expected abundance of the heavier isotopes $E(1,m_1)$, $E(2,m_1)$, ...,$E(R-1,m_1)$.
- Subtract these higher contributions from the actual abundances of the next $R-1$ peaks: $a_2 - E(1, m_1)$, $a_3 - E(2, m_1)$,...,

$a_R - E(R-1, m_1)$. Only the results higher than a threshold (ISOTR) are retained for further processing.

These sub-steps are recursively repeated for the residual cluster until all the residual peaks are less than the threshold (19). For example, if $a_2 - E(1, m_1) > \text{ISOTR}$, $(m_2, a_2 - E(1, m_1))$ becomes the next monoisotope and the steps are repeated.

3.2. FPGA Implementation of the Mass Spectra Processor

The algorithm outlined above has several properties that make it suitable for hardware implementation (see Note 5). The calculations involved in steps 1–4 are performed on a long mass/abundance data stream so can be implemented as deep computational pipeline involving mainly complex combinatorial logic. State machines are used for data sorting in step 2 and peak construction in step 4. Pipelining allows multiple sequential calculations to be performed in parallel on the data stream so that, once the pipeline is full, results are produced every clock cycle. Pipelining also allows a relatively high clock frequency (see Note 7).

Clustering and deisotoping are sequential operations performed on the significant centroided peaks in the spectrum (m_p, a_p pairs) which represent only a small subset of the original mass spectrum data set. As a consequence, in practice, these operations take only a fraction of the time required to process the entire spectrum in steps 1–4. Because clustering and deisotoping operations cannot be performed in a pipelined manner, the set of identified peaks have to be stored in a memory and dedicated state machines are required to implement clustering and deisotoping operations. Since the amount of memory required is relatively small, the best solution is to use the fast and wide data-width on-chip memory (embedded RAM) integrated on the FPGA chip. The arithmetic operations involved are well suited for fixed-point arithmetic resulting in significant savings of resources on the FPGA. For this implementation, it was found that a wordlength of 32 bits with 12 bits allocated for the fractional part introduced negligible errors when compared with a floating point PC implementation of the same algorithm.

The hardware solution described here has two major functional blocks: a peak detection unit, which identifies all significant spectral peaks (Implements steps 1–4 and the cluster flag generation) and a peptide identification unit that generates the final list of peptide masses and associated abundances (steps 5 and 6). The block diagram of the hardware processor is depicted on Fig. 4.

The first block is a Savitzky–Golay smoothing filter that implements the equations from the first algorithmic step. The smoothing operation is optional; the user can specify if the data is to be pre-processed or not.

The Savitzky–Golay smoothing filter is implemented as a 43 tap FIR filter with coefficients that can be loaded as user parameters in an LUT. The size of the smoothing window F can be set

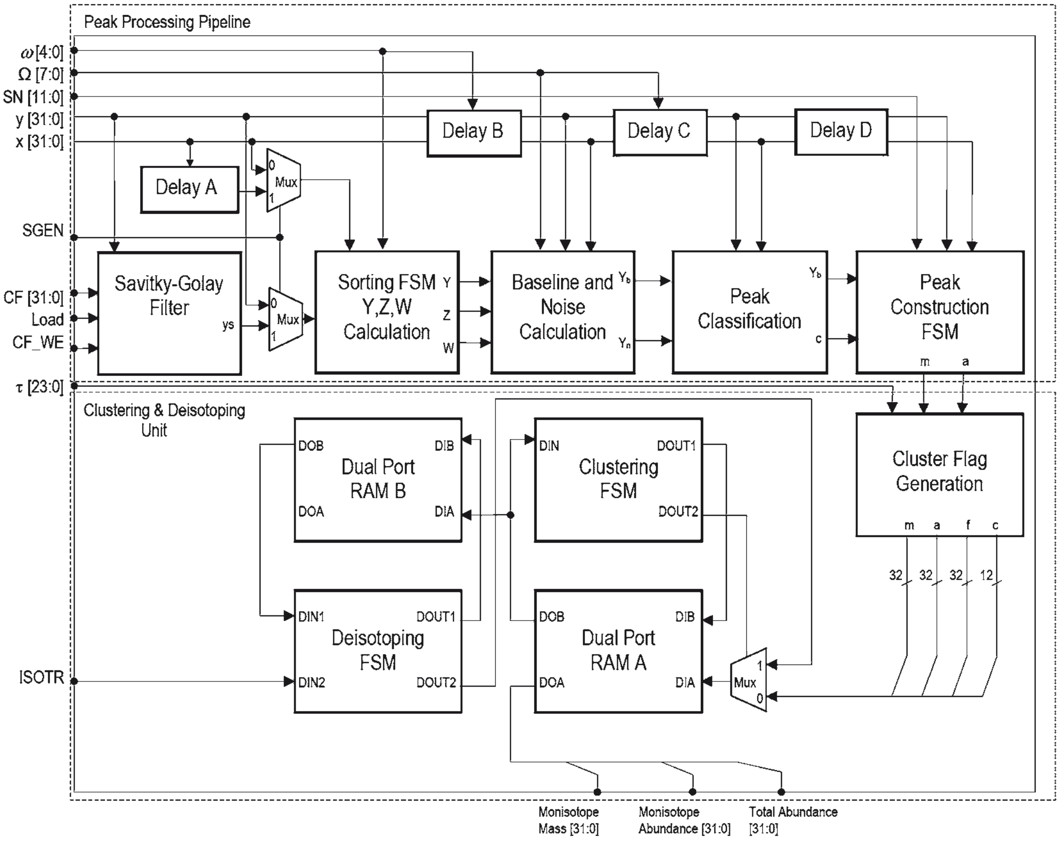


Fig. 4. Block diagram of the mass spectrum processor

by the user according to the instrument resolution setting (number of data points recorded per 1 m/z unit).

The raw mass spectrum is processed in three stages: (a) the first F spectral abundance values ($y_1,\ldots,y_F$) are loaded as filter coefficients and the and the last (F-1)/2 rows of the filter coefficient matrix are loaded on the FIR data input, (b) the middle row of the filter coefficients is loaded as coefficients and the following abundance values ($y_{F+1},\ldots,y_{N-F}$) is provided as the input to the FIR filter and finally (c) the last F data points ($y_{N-F+1},\ldots,y_N$) are loaded as FIR coefficients and the last (F-1)/2 rows of the coefficient matrix are loaded as data input. The filter is implemented as a single channel, highly parallel filter using a Xilinx Logicore block (24).

The peak construction pipeline depicted on Fig. 4 implements steps 2–4 of the algorithm. The sorting block computes the minimum (Z) and maximum (Y) abundances and their difference (W) over a sliding window of length ω. It is implemented using a structure that sorts in ascending order input data stream over the data window before the maximum and minimum values are found.

Baseline and noise are computed over a bigger spectral interval of Ω small windows of length ω using 32 bit pipelined dividers and accumulators.

Spectrum segmentation is performed by computing a signal to noise ratio for each delayed spectrum data point (x, y) according to step 3. The result is then compared with a user selectable threshold (SNTHR) and, for each data point, a 2 bit classification flag c is set to 1, 2 or 3 depending on whether the respective data point is classified as noise, support or signal respectively. The original spectral points (x, y), their associated classification flag c and the base-line (Υ_{base}) are aligned and fed into the peak construction state machine. Here, the centered mass m_k and base-line subtracted abundance a_k of spectral peaks are computed according to step 4 of the algorithm and passed to the cluster flag generator block for further processing.

This block is used to detect possible peak candidates that are isotopes of one or more singly charged chemical compounds, separated by the mass of a neutron. The circuit is a delay line for the input data peaks with a maximum length of p. It is assumed that the spectrum is sorted by increasing mass. Distances between the masses of all consecutive signal peaks starting with the lowest mass value m_1 are computed. If the distance between two successive peaks is between $1 - \tau$ and $1 + \tau$ (τ is a user selectable value typically set to 0.2) the peaks belong to the same cluster.

To speed up computations, there are p circuits that compute mass differences $m_1 - m_2, \cdots, m_{p+1} - m_1$ between m_1 and the following p consecutive mass values $m_2 < m_3 < \cdots < m_p < m_{p+1}$ in parallel. In our design, p is an adjustable parameter, which is selected according to mass spectrometer resolution, to be larger than the maximum number of signal peaks that are registered within a window of $1 + \tau\, m/z$. Typically about 50–100 samples/(m/z) are taken, so $p = (60\text{–}120)/3 = 20\text{–}40$ or less.

The output of the circuit is a cluster flag f_k of p-bits, which generated for each peak (m_k, a_k). For example, the k-th bit of the flag f_1 associated with m_1 is set to 1 if $m_1 - m_k \in [1-\tau, 1+\tau]$. If all the bits in the flag f_1 are zero, this indicates that m_1 has no isotopes.

The mass, abundance and cluster flag associated with each peak (m_k, a_k, f_k) are stored in RAM (A) at consecutive addresses, starting from zero as a $32 + 32 + 32 + 12$-bit word (32 bits for mass, abundance and flag each and 12 bits for the cluster index-not set initially). After all peaks are stored in the memory, the results are processed further by the clustering state machine.

Clustering is implemented as a state machine that sequentially reads the first dual port RAM (A) using the cluster flag to calculate the memory location of the peaks belonging to the same cluster. The identified peaks in a cluster are stored at consecutive memory locations in another RAM block (B). In the process, all peaks identified to belong to the same cluster are assigned the same cluster index, an integer that uniquely identifies a cluster.

The final processing step required to generate the peptide fingerprint is deisotoping. The deisotoping unit processes previously computed clusters from the dual port RAM (B), writes back partial results in RAM (B) and the final peak list in RAM (A) as illustrated in Fig. 4.

A cluster may contain more than one peptide and to each peptide may correspond a set of peaks separated by approximately 1 *m/z*. Effectively, the peaks in a cluster can be viewed as a superposition of isotopic distributions of two or more peptides. The deisotoping unit identifies all peptides (the mass of the monoisotope) within a cluster and calculates the total abundance of the peptide corresponding to each monoisotope using an approximation of isotopic patterns by Poisson distribution as described in step 6. Starting with the first peak in a cluster (assumed to be a monoisotope), the algorithm generates the theoretical isotopic distribution based on peak height (abundance) and mass value. The computed abundance values are then subtracted from the original peaks at the corresponding *m/z* values. Following subtraction, any abundance less than a specified threshold is set to zero. The next (non-zero) peak in the residual cluster is then considered a monoisotope, and the process is repeated until there are no residual peaks left. At each step, the monoisotopic mass value and original and total abundance (the sum of the monoisotopic peak and its theoretical isotopic abundances) are recorded in the final peak list. A subset of the final peaks will be used to perform database searching.

3.3. Database Search Engine

The database search engine (20) traces the peptide fingerprint back to the originating peptide by matching it against the expected (theoretical) peptide masses obtained by digesting *in silico* – on the fly – all protein sequences in the database. The protein sequence databases (MSDB for example) are in fact flat text files. A block diagram of the FPGA search engine is illustrated in Fig. 5.

In order to fully exploit the benefits of FPGA acceleration, the entire MSDB database was encoded, using a 5-bits code for each of the 20 constituent aminoacids, and stored in the local onboard memory of the FPGA module in a format that facilitates fast parallel searches. Additional codes were used to represent the end of a protein sequence and the end of the database.

By encoding the database more efficiently using only 5-bit "characters", the database size was reduced by about 40%. The resulting "shrunk" database currently occupies only 60% of the total 1 GB DDR SDRAM memory installed on the FPGA module.

One effect of storing the database in the local module memory is that it eliminates a significant memory access bottleneck, because of the PCI interface, that would otherwise be present if the protein database were stored in the computer memory. However, the most significant reason for encoding and storing the protein database in the local memory is that it enables parallel processing of protein sequences (see Note 8). In the current implementation, there are 48

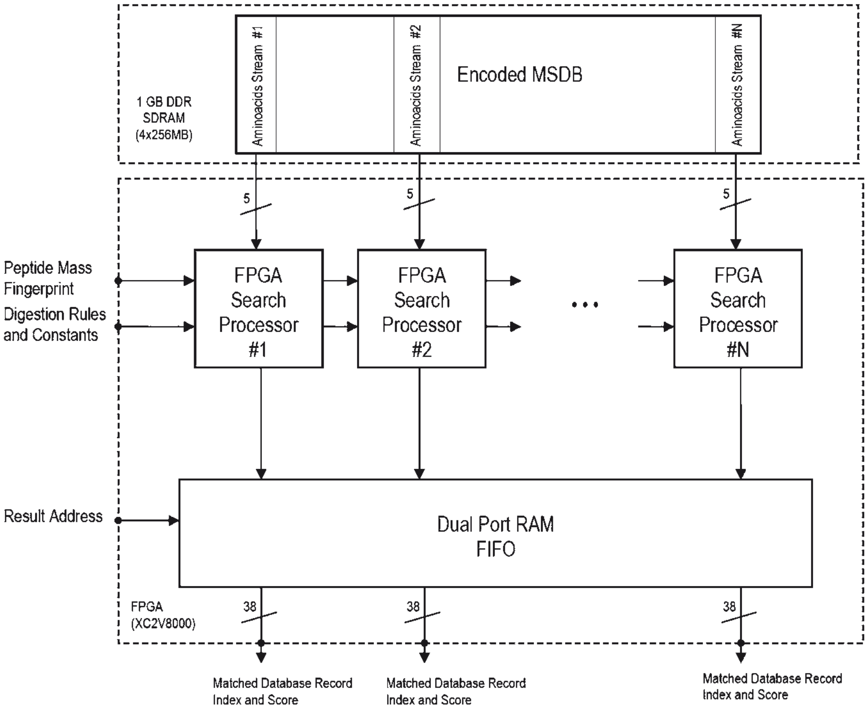


Fig. 5. Block diagram of the database search engine

protein sequences that are streamed out, in parallel, from the memory, as shown in Fig. 5. Each protein sequence is processed sequentially by a search processor implemented in the module's FPGA.

Each search processor performs two basic operations:

(a) the computation of theoretical peptide masses for every protein in the database by *in-silico* digestion and

(b) the subsequent computation, for each protein, of a matching score that indicates the likelihood that the peptide mass fingerprint generated by the mass spectrum processor, on the basis of experimental data, belongs to that particular protein.

Consequently, the search processor has two major functional blocks: an *in-silico* protein digestion unit and a scoring module.

Each search engine is connected to a 5-bit data stream. It reads one code every clock cycle from the corresponding memory column and passes it to the digestion unit. The digestion unit is responsible for calculating the peptide masses according to the specified digestion rule/parameter. The digestion unit calculates the cumulative mass of the aminoacids received until it encounters a cleavage site, protein record delimiter or the end of database marker. The masses of individual aminoacids, used to compute the peptide masses, are stored into a look-up table as 32 bits fixed-point numbers.

When calculating the peptide mass, the user can specify additional, post-translational modifications (PTM) rules. The system described here implements only fixed modifications. However, a modified version of the search processor, which can deal with variable modifications has been designed recently and is currently being tested. The new design also incorporates new functionality to allow searches to take into account peptides with missed cleavage sites.

The scoring unit calculates the number of peptide masses in the peptide mass fingerprint that are matched for every digested protein in the database. A user defined mass matching tolerance can be programmed by the user to account for MS instrument precision and other sources of errors that may affect the accuracy of the peptide mass fingerprint.

The precision of the comparison between the m/z values to be searched (m_1, m_2, ..., m_n) and the in-silico computed throretical peptide fragments m/z values (mt_i) can be specified as an user-defined parameter that holds the desired error tolerance (as ppm) which may be dictated by the MS instrument accuracy.

When a theoretical peptide fragment mass mt_i is computed, it is compared in parallel to the peptide mass fingerprint (m_1, m_2, ..., m_n) generated by the mass spectra processor. The result of the comparison is used to generate the basic cumulative score for every processed protein. In the current implementation, $n = 13$ but the number of m/z values that are used in the search can be increased at the expense of increasing the complexity of the design of individual search processors that will use more space on the FPGA device. This normally means that the number of processors that can be allocated on current FPGA device (Xilinx XC2V8000) may need to be reduced. As mentioned earlier, however, the logic capacity of the latest FPGA devices (Xilinx Virtex 5 family for example) has increased dramatically when compared with the device used in this implementation, so it is feasible to expect that a larger number of complex search processors could be fitted on the latest devices. Moreover, because of the parallel nature of the computations, the entire database can be divided into distinct subsets and loaded on separate FPGA modules each with its own parallel database search engine so that maintaining or even increasing the computational performance of the solution should not be an issue.

If a match is found, the score counter is incremented by one. The position of a match is also recorded in an n-bit match index word. When the end of a record is found, the record index counter, the score counter and the match index register outputs are stored in intermediate registers.

Each search processor has three outputs: a processing end flag that remains set after processing ends until the search processor is reset; an output index that remains set to the last available FIFO address where the total the number of matches is stored and a 39 bits output that contains the results (database index).

Results of the 48 search engines are collected in dual port RAM devices organized as FIFO structures of 64 words of 39 bits each. The user can specify a score threshold $\tau_{s,}$ so that only the matches that are above a threshold are saved i.e. if the score of a given match is higher than a programmable threshold τ_s, the corresponding record index and match index are stored in the output FIFO. The basic matching score can be used to implement more sensitive scoring schemes which account for peptide frequency distributions such as MOWSE (25), PIUMS (21) or more comprehensive Bayesian scoring approaches which also account for the individual properties of the proteins analyzed such as ProFound (26). Because of the low speed-up gain expected from a hardware implementation, these scoring methods are run on the PC server. The externalization of the scoring statistics means that the output of the search can be rapidly evaluated using different scores, and even developed into a consensus score validation scheme.

The database search engine occupies 99% of the FPGA's logic resources, 99% of the FPGA's internal RAM resources and 53% of the FPGA's I/O resources (see Note 9). The design has a clock frequency of 100 MHz which is dictated by clock frequency of the 1 GB on-board DDR SDRAM.

The design includes all necessary control and FIFO structures that implement a 64-bit wide data transfer between the FPGA devices at a rate of 320 Mbytes/s (see Note 10).

All arithmetic operations on the *m/z* values were performed using 32-bit unsigned fixed-point binary number representation of mass and abundance values, with 12 bits after the radix point.

3.4. Results

In order to validate the accuracy of the hardware implementation, reference C programs were developed for all the algorithms implemented in hardware to process the raw mass spectra and perform database searching and matching. The first validation step involved checking that both the hardware and software implementation generate identical results. The second validation step involved, in the case of the mass spectra processor, comparing the results with commercial software solutions such as MassLynx and assessing the ability to correctly resolve isotopomer distributions derived from asparagine containing peptides and the deamidated cognate peptide (19). The database search engine was also tested using theoretical peptide mass fingerprints derived from randomly selected proteins in the database.

The performance gains of the hardware implementation relative to the conventional software solution were evaluated by measuring the execution time of the main computational loops of the reference C implementations and comparing this with the FPGA processing time. It should be noted that the time elapsed for initializations of memory locations before the effective processing of data and disk access time are not included in the software processing time. Only the processing of effective algorithmic steps was measured. The reference design was compiled in C and was simulated on a dual 3.06 GHz Xeon processor server. Each C simulation was repeated 30 times and the average processing time was used for comparison.

The speed gains corresponding to the mass spectra processor, for different lengths of the analysed mass spectrum are summarized in Table 1.

In the case of the database search engine the performance gains compared with the C implementation of the algorithms are illustrated in Fig. 6 over 50 runs.

Table 1
Speed gains of FPGA vs. C implementation of mass spectra processing

Spectrum length	Processing time [ms] (Dual Xeon 3.06 GHz)	Processing time [ms] (XC2V8000,180 MHz)	Speed gain
25,488	20.27	0.1632	124.20
50,448	31.23	0.3105	100.56
75,168	47.33	0.4557	103.86
101,040	62.50	0.5607	111.46
125,184	79.17	0.7557	104.76
150,144	114.33	0.8547	133.76
175,104	130.20	1.0024	129.88
200,976	188.63	1.1219	168.13

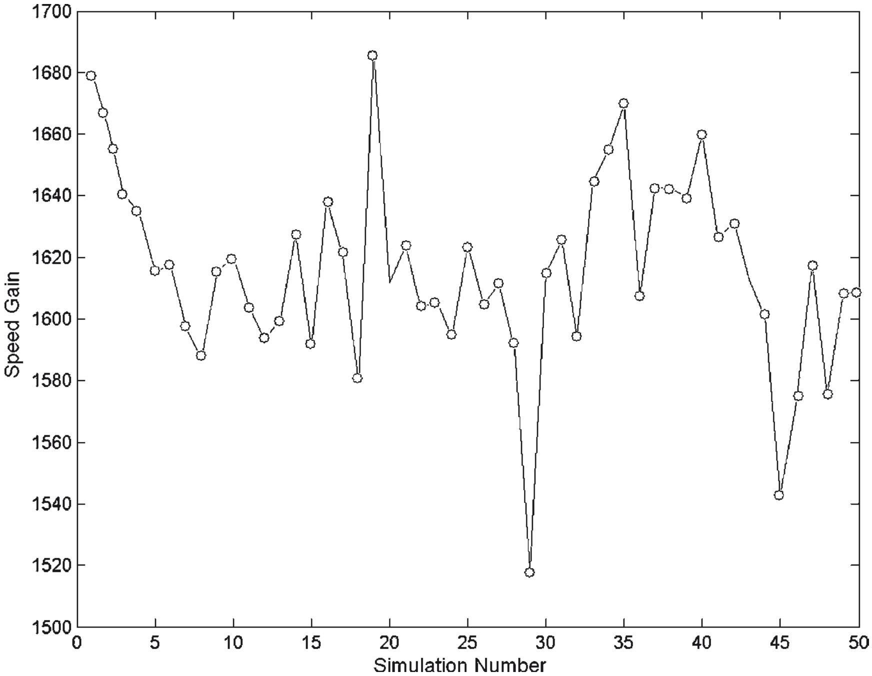


Fig. 6. Speed gains of FPGA vs. C implementation of database search engine

The C implementation was run on a Dual Xeon 3.06 GHz server – only the main computational loop was timed. In each run, a peptide mass fingerprint was generated for a randomly selected protein in the database. The "synthetic" peptide fingerprint was then used to perform a complete database search.

If only fixed modifications are considered and assuming no missed clevages, the FPGA system performs a complete database search in 240 ms (±0.02) irrespective of the peptide mass fingerprint. As seen from Fig. 6, the speed gain of the FPGA implementation when compared with the C software implementation ranges from 1,520 to 1,680-fold average speedup.

4. Notes

1. It is said that Reconfigurable Computers merge the benefits of application-specific digital hardware devices with the flexibility of software programmability. However, not every computational task is well suited for FPGA implementation. Algorithms that involve processing large streams or data and/or are inherently parallel are the best candidates for hardware acceleration. Whilst the flexibility of software programmability is evident, it is easy to overlook the difficulties of mapping an algorithm in hardware, which are still considerable despite the availability of high-level design tools, such as System Generator Toolbox for Matlab or Impulse C, the C-to-FPGA compiler. Conventional FPGA development tools were designed for electronics hardware engineers and require in-depth knowledge of hardware design languages (VHDL, Verilog) and digital electronics. The emerging high-level design tools, whilst offering a great level of abstraction still require a fair amount of manual optimization hence low-level design knowledge is still essential. Moreover, because there is no standard RC architecture, most common design tools do not target specific FPGA boards and as a result designs have to be mapped manually on the chosen RC platform. This cannot be achieved without a detailed understanding of the architecture of the hardware system.
2. An important factor that has to be considered is the communication overhead associated with data transfer between PC and device, which should represent only a fraction of the actual execution time. For a known reconfigurable computing platform, it is possible to evaluate at this stage the actual communication costs incurred by transferring data between hardware and software. This aspect is a major decision factor in the selection of the FPGA system best suited for an application.

3. There are tools available that allow automatic conversion of an algorithm implemented in C into a hardware application engine, such as Impulse C (www.impulsec.com). This is perhaps the easiest route to develop a hardware design. These tools typically provide integrated support for profiling and partitioning the algorithm and have design space exploration capabilities for evaluating different implementations to determine the best trade-off between resource utilization and performance gain for a particular algorithm.

4. Different high-level tools will typically generate hardware designs that differ in terms of both area utilization and speed. Moreover, the quality of compiler-generated designs is said to be significantly inferior to that of manually generated ones. In some cases, developing a hardware implementation badly may be enough to cancel the potential speed-up that can be achieved.

5. "Debugging" a computer program is an unpleasant but familiar task to any software developer. For hardware implementation, this process is intrinsically more difficult as the correctness of the result can be affected by subtle timing issues that cannot be detected easily, especially when dealing with a complex parallel design. To help debug a design, some tools provide clock-accurate hardware simulation models others require that the simulation code is written by hand. Consequently, debugging a hardware implementation is still largely the realm of digital electronics engineers rather than software developers, at least for now.

6. The first step in developing a RC bioinformatics solution is to identify computational kernels most likely to benefit from hardware acceleration and try to estimate the potential speed-up achievable. Profiling tools such as the gprof utility provided by most Unix systems (also available under the Cygwin Linux-like environment for Windows) provide detailed timing information which can be used to identify the computational intensive kernels of an application. Sections of the code or subroutines that are executed most of the time are potential candidates for hardware implementation.

7. In our application, processing a mass spectrum involved a fixed sequence of operations performed on every data point, which consumed over 90% of processing time, hence a computational pipeline provided an ideal implementation architecture. Such pipeline, once full, provides a result every clock cycle so that execution time and hence speed-up could be estimated if the clock frequency of the hardware design is known. The clock frequency however depends on the final implementation as well on the characteristics the targeted FPGA device. Although some automatic profiling methodologies that attempt to quantify the potential benefit of hardware

acceleration have started to emerge, it is still difficult to obtain a good estimate of achievable speed-up without designing and implementing the digital hardware realization of a computationally intensive kernel. A rough estimate of the processing time can be derived though, using a conservative value for the clock frequency.

8. The code selected to run in hardware should ideally have low data dependency, to facilitate parallel implementation. Fingerprint matching for example can be performed in parallel on groups of proteins or in the extreme case on individual proteins. Since the number of proteins in the database is very large, the potential for speed up is huge given the right FPGA platform. Specifically, FPGA module used to implement the database search engine (BenData DD) provided sufficient on-board memory to hold the entire database. Crucially, the architecture of the module allowed data transfers between memory and FPGA on a 256 bit-wide channel at 100 MHz so that multiple proteins could be streamed out from the memory and processed in parallel by individual search processors programmed on the FPGA fabric.
9. The initial designs had to be optimized manually, at HDL level, to reduce the area used and to increase the clock frequency of the synthesized design.
10. Developing the complete hardware solution involves significant low-level programming for designing control and synchronisation modules to manage data transfers between the hardware processors running on different FPGAs, between the FPGA system and the host PC and between FPGAs and the on-board memory modules.

References

1. Andrade, T., Berglund, L., Uhlén, M., and Odeberg, J. (2006) Using Grid technology for computationally intensive applied bioinformatics analyses. *In Silico Biol.* **6**, 495–504.
2. El-Ghazawi, T., Bennett, D., Poznanovic, D., Cantle, A., Underwood, K., Pennington, R., Buell, D., George, A., and Kindratenko, V. (2006) Is high-performance reconfigurable computing the next supercomputing paradigm? *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*,Tampa, Florida. doi:10.1109/SC.2006.38.
3. Estrin, G. (1960) Organization of computer systems – the fixed plus variable structure computer. *Proceeding of western joint computer conference*, New York, pp. 33–40.
4. Estrin, G. (2002) Reconfigurable computer origins: the UCLA fixed-plus-variable (F+V) structure computer. *IEEE Ann. Hist. Comput.* **24**, 3–9.
5. Gokhale, M.B., and Graham, P.S. (2005) *Reconfigurable computing: accelerating computation with Field-Programmable Gate arrays*, Springer.
6. Hauck, S., and Dehon, A. (2008) *Reconfigurable Computing: The theory and practice of FPGA-based computation*, Elsevier.
7. Fagin, B., Watt, J.G., and Gross, R. (1993) A special-purpose processor for gene sequence analysis *Comput. Appl. BioSci.* **9**, 221–226.
8. Hughey, R. (1996) Parallel hardware for sequence comparison and alignment. *Comput. Appl. BioSci.* **12**, 473–479.
9. Wozniak, A. (1997) Using video-oriented instructions to speed up sequence comparison. *Comput. Appl. BioSci.* **13**, 145–150.

10. Guerdoux-Jamet, P., and Lavenier, D. (1997) SAMBA: hardware accelerator for biological sequence comparison *Comput. Appl. BioSci.* **13**, 609–615.
11. Lavenier, D. (1998) Speeding up genome computations with systolic accelerator. *SIAM News* **31**, 1–8.
12. Guccione, A.S., and Keller, E. (2002) Gene matching using Jbits. *Proceedings of the reconfigurable computing is going mainstream, 12th international conference on field-programmable logic and applications*, pp. 1168–1171.
13. Simmler, H., Singpiel, H., and Männer, R. (2004) Real-time primer design for DNA chips. *Interscience Concurr. Comput. Pract. Exper.* **16**, 855–872.
14. Marongiu, A., Palazzari, P., and Rosato, V. (2003) Designing hardware for protein sequence analysis. *Bioinformatics* **19**, 1739–1740.
15. Oliver, T., Schmidt, B., Nathan, D., Clemens, R., and Maskell, D. (2005) Using reconfigurable hardware to accelerate multiple sequence alignment with ClustaIW. *Bioinformatics* **21**, 3431–3432.
16. Anish, T.A., Dumontier, M., Rose, J.S., and Hogue, C.W.V. (2005) Hardware-accelerated protein identification for mass spectrometry. *Rapid Commun. Mass Spectr.* **19**, 833–837.
17. Panitz, F. *et al.* (2007) SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation. *Bioinformatics* **23**, 387–391.
18. Dandass, Y.S., Burgess, S.C., Lawrence, M., and Bridges, S.M. (2008) Accelerating string set matching in FPGA hardware for bioinformatics research. *BMC Bioinformatics* **9**, doi: 10.1186/1471-2105-9-197.
19. Bogdan, I.A, Coca, D., Rivers J., and Beynon J.R. (2007) Hardware acceleration of processing of mass spectrometric data for proteomics. *Bioinformatics* **23**, 724–731.
20. Bogdan, I.A., Rivers, J., Beynon, J.R., and Coca, D., (2008) High-performance hardware implementation of a parallel database search engine for real-time peptide mass fingerprinting. *Bioinformatics* **24**, 1498–1502.
21. Samuelsson, J., Dalevi, D., Levander, F., and Rögnvaldsson, T. (2004) Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics* **20**, 3628–3635.
22. Breen, E.J., Hopwood, F.G., Williams, K.L., and Wilkins, M.R. (2000) Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis* **21**, 2243–2251.
23. Savitzky, A., and Golay, M.J.E. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639.
24. Xilinx. (2004) Distributed arithmetic FIR filter V9.0, DS240, Xilinx Inc.
25. Pappin, D.J., Hojrup, P., and Bleasby, A.J., (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3**, 327–332.
26. Zhang, W., and Chait, B.T., (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* **72**, 2482–2489.
27. Xilinx. (2007) Virtex II platform FPGAs: complete data sheet. DS031, Xilinx Inc.

Chapter 13

Mining Proteomic MS/MS Data for MRM Transitions

Jennifer A. Chem (Mead), Luca Bianco, and Conrad Bessant

Abstract

Multiple reaction monitoring (MRM) of peptides is a popular proteomics technique that employs tandem mass spectrometry to quantify selected proteins of interest, such as those previously identified in differential protein identification studies. Using this technique, the specificity of precursor to product transitions is exploited to determine the absolute quantity of multiple proteins in a single sample. Selection of suitable transitions is critical for the success of MRM experiments, but accurate theoretical prediction of fragmentation patterns and peptide signal intensity is currently not possible. A recently proposed solution to this problem is to combine knowledge of the preferred properties of transitions for MRM, taken from expert practitioners, with MS/MS evidence extracted from a proteomics data repository. In addition, by predicting retention time for each peptide candidate, it allows selection of several compatible transition candidates that can be monitored simultaneously, permitting MRM. In this chapter, we explain how to go about designing transitions using the web-based transition design tool, MRMaid, which leverages high quality MS/MS evidence from the Genome Annotating Proteomic Pipeline (GAPP).

Key words: GAPP, Mass spectrometry, MRMaid, Multiple reaction monitoring, Quantitative proteomics

1. Introduction

1.1. Introduction to MRM and Principles of Transition Design

Multiple reaction monitoring (MRM) is a technique that uses mass spectrometry (MS) to determine the quantities of specific proteins of interest. MS and MS/MS are performed as reverse phase HPLC separation is underway. Each tryptic peptide is analyzed by the selection of a specific mass over charge ratio (*m/z*) using a quadrupole MS (Q1). Once separated, it undergoes fragmentation in the collision cell, generating product ions exclusive to the precursor, which are selected for monitoring by a third quadrupole (Q3). By filtering by mass at two stages, background may be overcome through improved signal to noise ratio, and

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_13, © Humana Press, a part of Springer Science+Business Media, LLC 2010

several transitions monitored quickly. MRM becomes a quantitative approach when a known quantity of labeled synthetic peptide is spiked into the sample. This surrogate peptide is identical in sequence to the expected target peptide, so elutes at the same moment, but demonstrates a known mass shift value in MS/MS (1). The pair of observed *m/z* ratio of a peptide, and its corresponding product ion *m/z* ratio, is referred to as a "transition." Accordingly, to monitor a protein of interest, it must be known in advance which transitions are most suitable for the protein targets. In simple protein mixtures, a single transition may be adequate to monitor a particular protein of interest, but in complex samples, such as serum, several transitions are generally required due to noise and proteins of very high abundance affecting the signal (2, 3).

The major challenge for MRM is selecting which peptides are most appropriate for monitoring because usually each protein has several tryptic cleavage sites. Traditionally, a lab-based discovery phase is carried out prior to MRM, allowing the observed MS information to direct the selection of transitions (4–6). This empirical method is usually successful, however it consumes time and costly laboratory resources. Furthermore, the experimental approach represents repetition of effort since, in many cases, other groups have already acquired MS/MS data for the protein of interest and have deposited that data in a publicly accessible database.

To increase the efficiency of transition design, we have developed MRMaid (7), a transition design tool that utilizes a combination of expert knowledge about what makes a good MRM transition, and MS/MS data from the genome annotating proteomic pipeline (GAPP) database.

1.2. Overview of GAPP

GAPP is an automated pipeline that assigns peptide and protein identifications to proteomic MS/MS data (8). It is publicly available web-based service at http://www.gapp.info, and includes both public and private datasets, the latter accessible exclusively via user authentication. MS/MS peak lists (in .pkl, .mgf or mzXML format) and corresponding metadata (such as cleavage agent, mass tolerances, species, and number of missed cleavages) are submitted via the website. GAPP performs peptide identification using an X!Tandem-based database search engine (9), with peptide identifications being validated and protein identifications inferred using the advanced average peptide score (APS) algorithm (10). Peptides and proteins identified in data submissions are automatically stored in a database, along with related confidence scores. The pipeline is capable of finding any peptides expected, including those that cross intron–exon boundaries, and those due to SNPs, alternate splicing and those that contain post-translational modifications (PTMs).

Although GAPP's initial focus was the re-annotation of genomes with protein data, the database of peptides identified by

GAPP has proved useful as a repository of high quality identifications. All identifications in the GAPP database are derived from MS/MS data by the GAPP pipeline, which is particularly rigorous in ensuring whether identifications are valid using a reverse decoy database search.

Since anyone can upload data to GAPP for processing via the web interface, the GAPP database continues to grow, providing increasingly more evidence for transition design. The web interface also allows the content of the database to be browsed and searched, which can give an indication as to whether MRMaid is going to be useful in your research (e.g. Is there any data for the species you are working with? Is there a large representation of the biological sample type you are using?). Such questions can be answered using the "Mine the GAPP" section of the website.

1.3. MRMaid Overview

While evidence for transition design is already present in GAPP, MRMaid performs the crucial role of selectively mining the data and contextualizing it using expert knowledge, to help ensure that transitions are fit for purpose. For example, a peptide must be unique to the protein of interest (proteotypic (11)) if it is to be useful in MRM. Furthermore, MRMaid calculates the hydrophobicity and retention time of each peptide candidate so suitable transitions may be selected, such that multiple peptides can be monitored simultaneously without overlapping elution times in reverse-phase HPLC. Information on how MRMaid works, and how you can use it for transition design, are in the following sections.

2. Materials

MRMaid is accessible at http://www.mrmaid.info (note that MRM is not capitalized in this URL). MRMaid does not require any software to be installed locally, and it is designed to work in all major web browsers.

To use MRMaid, you must have the following:

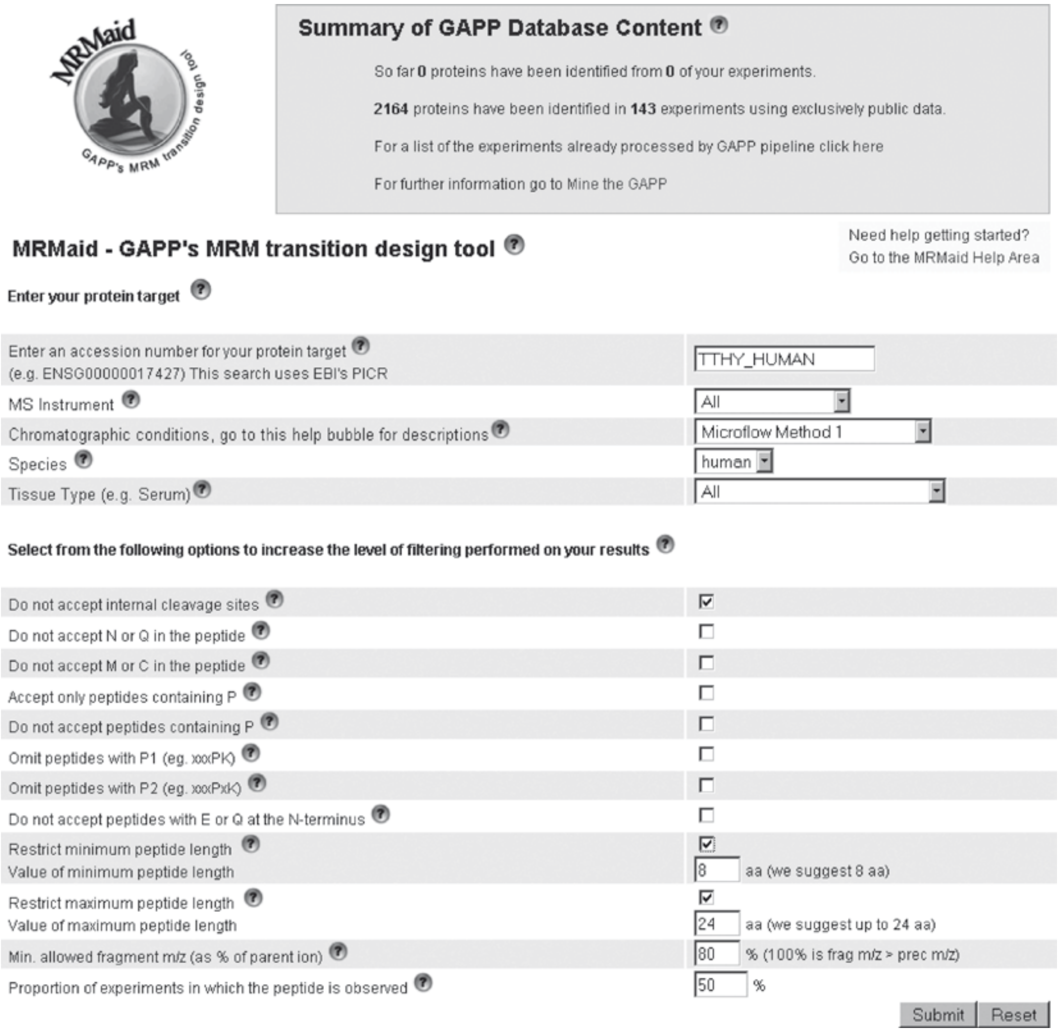
1. A protein, or several proteins, that you would like to monitor using MRM. For example, a protein identified as a biomarker of disease by differential studies.
2. The accession number for the protein(s) that you would like to monitor. Since MRMaid uses the EBI's Protein Identifier Cross-Reference (PICR) service (12), an identifier from virtually any major database can be used to start the search (e.g. Swissprot, PDB, Ensembl, IPI).

   Optional: The MS instrument type, the reverse phase chromatographic conditions and the type of tissue or sample in which your protein is present. This information can be used to ensure the transitions predicted by MRMaid are suitable for your individual requirements, by only considering data that fits your monitoring conditions.

3. Methods

3.1. Data Input

The front page of the MRMaid tool is where all the information necessary to design a transition is entered (Fig. 1). On this page, and throughout the rest of MRMaid, access to interactive help is provided in the form of circles containing question marks. Clicking on a question mark will open up a new small (but resizable) window containing an explanation of the relevant topic.



Summary of GAPP Database Content

So far **0** proteins have been identified from **0** of your experiments.

2164 proteins have been identified in **143** experiments using exclusively public data.

For a list of the experiments already processed by GAPP pipeline click here

For further information go to Mine the GAPP

MRMaid - GAPP's MRM transition design tool

Need help getting started?
Go to the MRMaid Help Area

Enter your protein target

Enter an accession number for your protein target (e.g. ENSG00000017427) This search uses EBI's PICR	TTHY_HUMAN
MS Instrument	All
Chromatographic conditions, go to this help bubble for descriptions	Microflow Method 1
Species	human
Tissue Type (e.g. Serum)	All

Select from the following options to increase the level of filtering performed on your results

Do not accept internal cleavage sites	☑
Do not accept N or Q in the peptide	□
Do not accept M or C in the peptide	□
Accept only peptides containing P	□
Do not accept peptides containing P	□
Omit peptides with P1 (eg. xxxPK)	□
Omit peptides with P2 (eg. xxxPxK)	□
Do not accept peptides with E or Q at the N-terminus	□
Restrict minimum peptide length Value of minimum peptide length	☑ 8 aa (we suggest 8 aa)
Restrict maximum peptide length Value of maximum peptide length	☑ 24 aa (we suggest up to 24 aa)
Min. allowed fragment m/z (as % of parent ion)	80 % (100% is frag m/z > prec m/z)
Proportion of experiments in which the peptide is observed	50 %

Fig. 1. The MRMaid homepage, where all the information for a transition design is entered. This example shows the Swissprot identifier for transthyretin protein (TTHY_HUMAN)

In addition, a glossary of terms plus other related material, including tutorial videos and demonstrations, can be found in the MRMaid "help area" accessible via the toolbar or homepage.

As mentioned, the only mandatory field is the first one, the protein identifier. The other options on this page allow additional constraints to be placed on the transition design. These options represent the most commonly used design constraints and are explained below.

Internal cleavage sites: As a default setting, peptides with internal cleavage sites (namely peptides with K or R, unless followed by P) are omitted to prevent selection of peptides that may be irregularly cleaved.

MS Instrument: Each type of instrument is known to have a different set of preferred proteotypic peptides (11). This option accounts for this phenomenon by allowing you to select the type of instrument that you will be using for MRM.

Chromatographic conditions: Similarly, it makes sense to specify the chromatographic conditions that you are planning to use for your MRM experiments because MRMaid applies this information to predict peptide retention times using a published model (13). There are eight different options, including nano-, micro-, and normal-flow setups. For a full description of each, refer to the help documentation.

Tissue type: Since protein expression can vary significantly between tissue types, the type of biological sample can be specified as a filter in the search for transitions. In complex samples, such as serum, this is particularly useful since the wide dynamic range of protein abundance, and inherent sample complexity, can present unique challenges for transition design. The ability to choose transitions based upon experiments performed on your particular sample type increases the likelihood of successful candidate selection.

Peptide sequence features: In addition to constraining experiment-specific factors, some practitioners use sequence information when selecting peptides for MRM. For example, peptides containing N and Q may be avoided because these residues can be deamidated, which can result in irregularities and problems with reproducibility of fragment ion *m/z* values. Opting to omit peptides containing Q or E at the N-terminus may be desirable because these residues can spontaneously cyclize to form pyroglutamate, and residues that are frequently modified, such as M and C, may also be omitted to avoid unreliable mass shifts.

Proline presence and location: Proline is a particularly important residue to consider when designing transitions; peptides containing P may be considered favorable because they generally produce high intensity MS/MS peaks. However, this can result in a single highly abundant fragment ion that swamps the remainder of the tandem MS spectrum, so users may opt to omit peptides containing P. If peptides including P are to be considered, the location of P in the peptide primary sequence can also be important. Having P adjacent to the C-terminus (P_1) or in the second position from the C terminus (P_2) is generally not desirable in MRM because a very short, non-specific product y-ion will result. The option to omit P_1 and P_2-containing peptides is therefore provided.

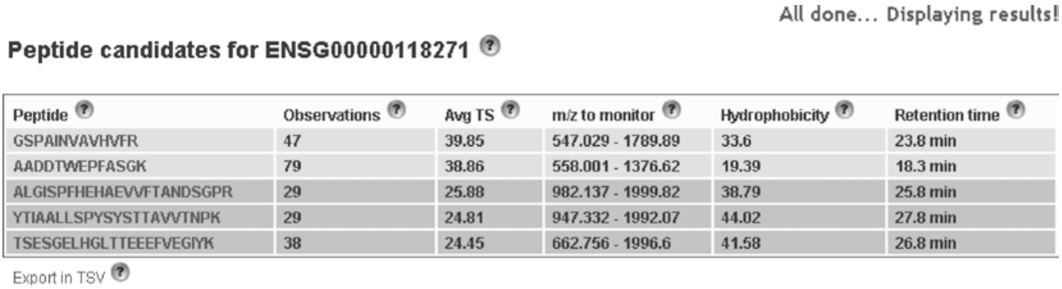
Peptide length: It makes sense to restrict peptides length to within a certain range because short peptides (<8 residues) are unlikely to be unique, and peptides longer than approximately 20–25 residues are likely to exceed the acceptable mass range: usually within *m/z* of 500–1600.

Proportion of experiments in which the peptide should be seen: Using this option, frequency of peptide identifications in the database may be used as a measure of transition reliability for the protein. For example, if you enter “50,” then this means that the peptide candidate(s) presented in green in the results table were assigned in at least 50% of occasions when your protein target was successfully identified. The higher the value you enter, the more stringent the prediction will be.

Once all the desired options have been selected, click the “submit” button; this initiates the transition design process. The peptide level results are then printed to screen, providing a portal to the product ion information.

3.2. Interpreting the Results

Using the example of protein transthyretin (TTHY_HUMAN), with the chromatographic conditions set to Microflow method 1 (default), peptide length set to between 8 and 24 amino acids and with the remaining constraints set to default, five peptide candidates are found. These are shown on the peptide candidate results page, a portion of which is shown in Fig. 2. Note that the results presented in this chapter are those acquired at the time of writing



All done... Displaying results!

Peptide candidates for ENSG00000118271

Peptide	Observations	Avg TS	m/z to monitor	Hydrophobicity	Retention time
GSPAINVAVHVFR	47	39.85	547.029 - 1789.89	33.6	23.8 min
AADDTWEPFASGK	79	38.86	558.001 - 1376.62	19.39	18.3 min
ALGISPFHEHAEVVFTANDSGPR	29	25.88	982.137 - 1999.82	38.79	25.8 min
YTIAALLSPYSYSTTAVVTNPK	29	24.81	947.332 - 1992.07	44.02	27.8 min
TSESGELHGLTTEEEFVEGIYK	38	24.45	662.756 - 1996.6	41.58	26.8 min

Fig. 2. The peptide candidates results page for transthyretin (TTHY_HUMAN), processed on November 24, 2008. *Darker shading* represents rows colored *red*, and *lighter shading* for *green* rows

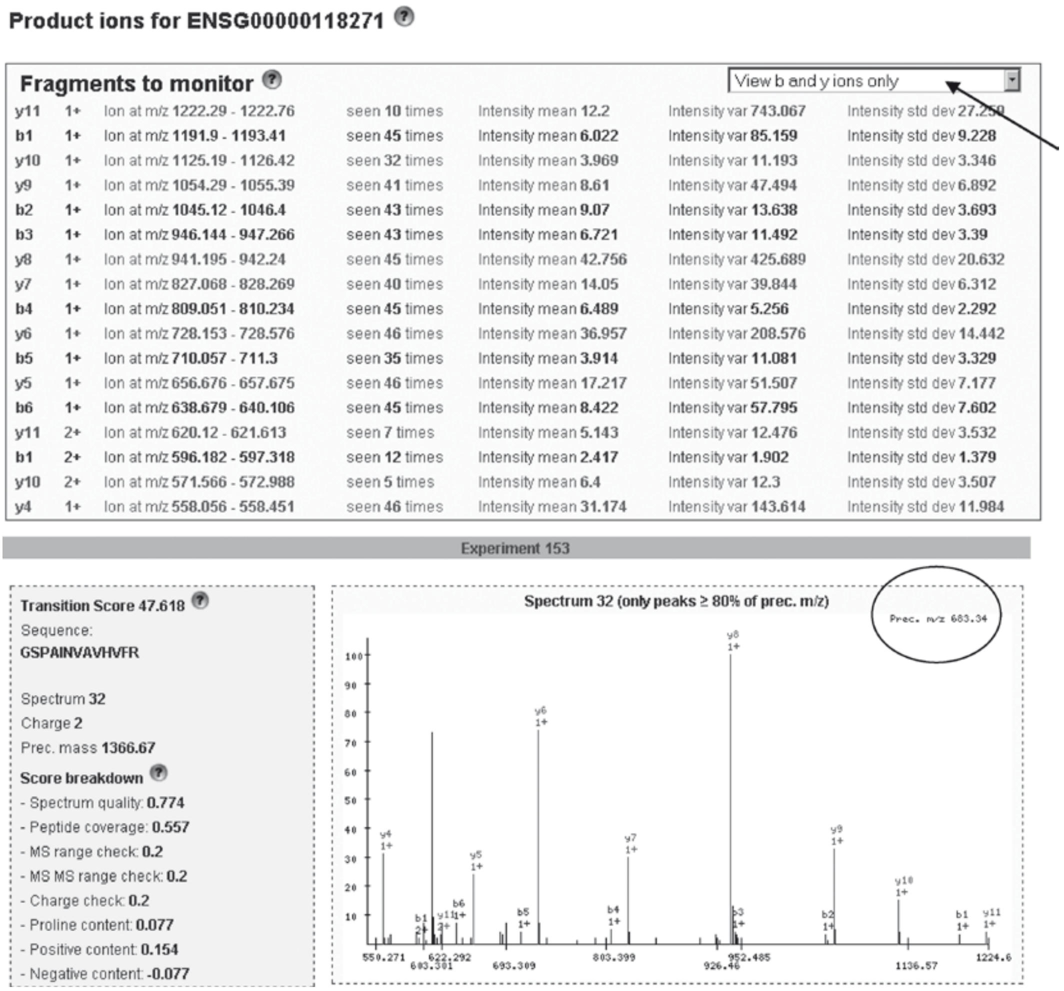
– as data continues to be submitted to GAPP, these results will most likely change.

On the screen, this table is displayed in color, with green rows showing peptide sequences that met the (default) threshold of 50% of observations. The rows of peptides that do not meet this particular criterion are shown in red. Key features of this table include the number of observations, which is the number of discrete datasets in GAPP to which this peptide was assigned for transthyretin, and the transition score (TS), which is a weighted sum of several key characteristics associated with each peptide. TS gives a quantitative measure of expected performance of that peptide in MRM; it is an average because it accounts for all TS values, across all observations of the peptide in GAPP database.

The components included in the TS calculation include coefficients to reflect both the peptide sequence, as well as the nature of the MS/MS spectral evidence for the peptide. For example, peptides with residues that are often post-transitionally modified, such as C and M are negatively weighted, and peptides with a 2+ or 3+ charge state are positively weighted over those with 1+. The finer details on how TS is derived, and the rationale for each coefficient can be found in the documentation on the MRMaid web site and in the accompanying paper (7). For the purposes of using MRMaid, a TS value above approximately 28 is good, and above 35 is excellent. As shown in Fig. 2, the peptide candidates are ranked according to average TS in the results table.

From the results for transthyretin, we see that GSPAINVAVHVFR looks like a good candidate, with the highest number of observations and highest transition score. Its retention time is predicted to be 23.8 min, given the chromatographic conditions that were entered at time of search. For the purpose of this example, we therefore select GSPAINVAVHVFR for further analysis. By clicking on this sequence, you can retrieve the product ion information, shown in Fig. 3.

In this particular example, MRMaid has predicted that y8, y6, and y4 should be suitable for monitoring. Of these, y8 in the *m/z* range 941.195 – 942.24 looks the best because it has a high number of observations *and* has a high mean signal intensity, with a relatively low signal standard deviation. A range of *m/z* values is given for y8, and the other ions in Fig. 2 because the fragment mass tolerance window is applied when assigning ion types. When the data was originally submitted to the GAPP pipeline, the resolution of the MS/MS instrument was taken into account by applying mass tolerances to the peptide identification process. In this way, if a peak falls anywhere within the specified window (above and below the expected absolute value given the masses of the amino acids), then it is successfully assigned. Tolerances are an issue because the peak separation achieved is dependent on the type of mass spectrometer.



Product ions for ENSG00000118271 ?

Fragments to monitor ?

View b and y ions only

y11	1+	Ion at m/z 1222.29 - 1222.76	seen 10 times	Intensity mean 12.2	Intensity var 743.067	Intensity std dev 27.259
b1	1+	Ion at m/z 1191.9 - 1193.41	seen 45 times	Intensity mean 6.022	Intensity var 85.159	Intensity std dev 9.228
y10	1+	Ion at m/z 1125.19 - 1126.42	seen 32 times	Intensity mean 3.969	Intensity var 11.193	Intensity std dev 3.346
y9	1+	Ion at m/z 1054.29 - 1055.39	seen 41 times	Intensity mean 8.61	Intensity var 47.494	Intensity std dev 6.892
b2	1+	Ion at m/z 1045.12 - 1046.4	seen 43 times	Intensity mean 9.07	Intensity var 13.638	Intensity std dev 3.693
b3	1+	Ion at m/z 946.144 - 947.266	seen 43 times	Intensity mean 6.721	Intensity var 11.492	Intensity std dev 3.39
y8	1+	Ion at m/z 941.195 - 942.24	seen 45 times	Intensity mean 42.756	Intensity var 425.689	Intensity std dev 20.632
y7	1+	Ion at m/z 827.068 - 828.269	seen 40 times	Intensity mean 14.05	Intensity var 39.844	Intensity std dev 6.312
b4	1+	Ion at m/z 809.051 - 810.234	seen 45 times	Intensity mean 6.489	Intensity var 5.256	Intensity std dev 2.292
y6	1+	Ion at m/z 728.153 - 728.576	seen 46 times	Intensity mean 36.957	Intensity var 208.576	Intensity std dev 14.442
b5	1+	Ion at m/z 710.057 - 711.3	seen 35 times	Intensity mean 3.914	Intensity var 11.081	Intensity std dev 3.329
y5	1+	Ion at m/z 656.676 - 657.675	seen 46 times	Intensity mean 17.217	Intensity var 51.507	Intensity std dev 7.177
b6	1+	Ion at m/z 638.679 - 640.106	seen 45 times	Intensity mean 8.422	Intensity var 57.795	Intensity std dev 7.602
y11	2+	Ion at m/z 620.12 - 621.613	seen 7 times	Intensity mean 5.143	Intensity var 12.476	Intensity std dev 3.532
b1	2+	Ion at m/z 596.182 - 597.318	seen 12 times	Intensity mean 2.417	Intensity var 1.902	Intensity std dev 1.379
y10	2+	Ion at m/z 571.566 - 572.988	seen 5 times	Intensity mean 6.4	Intensity var 12.3	Intensity std dev 3.507
y4	1+	Ion at m/z 558.056 - 558.451	seen 46 times	Intensity mean 31.174	Intensity var 143.614	Intensity std dev 11.984

Experiment 153

Transition Score 47.618 ?
Sequence:
GSPAINVAVHVFR

Spectrum 32
Charge 2
Prec. mass 1366.67

Score breakdown ?
- Spectrum quality: 0.774
- Peptide coverage: 0.557
- MS range check: 0.2
- MS MS range check: 0.2
- Charge check: 0.2
- Proline content: 0.077
- Positive content: 0.154
- Negative content: -0.077

Spectrum 32 (only peaks ≥ 80% of prec. m/z)

Fig. 3. Product ion information for peptide GSPAINVAVHVFR assigned to TTHY_HUMAN. By default, only b- and y- ions are shown in the fragment table, however the table can display all ion information by selecting from the options in the drop down menu in the top *right* corner of the table (indicated by an *arrow*). y-ions are colored *red* and b-ions in *blue* for both the fragment summary table (*top*) and the spectrum schematic (*bottom*). All other ions are shown in *black*. The *circle* highlights the place where you can read off the observed precursor (peptide) ion *m/z* for individual experiments processed by GAPP

In general, y-ions with a single positive charge are selected in preference to all other ion types for quantitative monitoring, particularly y8–y10. This trend is reflected in the many studies where experimentally validated transitions have been published; Anderson and Hunter's paper is a good example (14). y-ions are chosen because they generally have higher signal intensity, so they can be distinguished easily from background noise, and being 1+ charge state, they are most suitable for the mass filtering in the triple quadrupole instrument – the MS instrument routinely used for MRM studies.

3.3. Transition Validation

Based on the output of MRMaid, the recommended transition for TTHY_HUMAN comprises the peptide ion (GSPAINVAVHVFR) with *m/z* of approximately 683.4, which is shown in MRMaid's schematic spectra, and the fragment ion *m/z* between 941.195 and 942.24 for the y8 ion, shown in the table (Fig. 3). Formally, the transition is written as "precursor *m/z*"/"product ion *m/z*" followed by the retention time, so if we take a mid-range product ion value for y8, the transition for TTHY_HUMAN is: 683.4/941.5 at 23.8 min.

However, it is important to remember that this is a transition predicted computationally, so it should be validated before commencing quantitation. Indeed, it is good practice to validate two or three of the transitions predicted by MRMaid per protein of interest. Nevertheless, using MRMaid or a similar tool is still a much more efficient way to choose transitions when compared with testing tens or hundreds of transitions manually as would be the case in the absence of computational assistance.

3.4. Using Fragment Ion Information for Selection of Multiple Transitions

It is worth noting that MRMaid can be used for multiple transition design. Multiple transitions are transitions that require more than one fragment ion to be monitored to confirm the presence of the protein target in a sample. For example, in a complex biological sample, such as whole serum, it may be necessary to monitor three or four product ions to be sure the target protein is present; this may be due to excessive noise in the spectrum. By looking down the list of fragments in the product ion view of MRMaid – as in Fig. 3 – you may select several suitable ions for the peptide, without requiring additional searches to be performed.

3.5. Exporting Results as a Spreadsheet

All the peptide and product ion information for each protein target may be downloaded for viewing as a spreadsheet. This is useful for keeping a record of the predicted candidates for performing validative MS, and for use when designing peptide surrogates. To download the results, click on "Export in TSV" below the peptide table (Fig. 2). Your browser will then prompt you to save the file. The files are in TSV (tab separated values) format, which is a text-based format accepted by all major spreadsheets and data analysis programs. A video demonstrating the process for downloading results is available in the help area of the MRMaid website.

3.6. MRMaid for MRM Transition Design

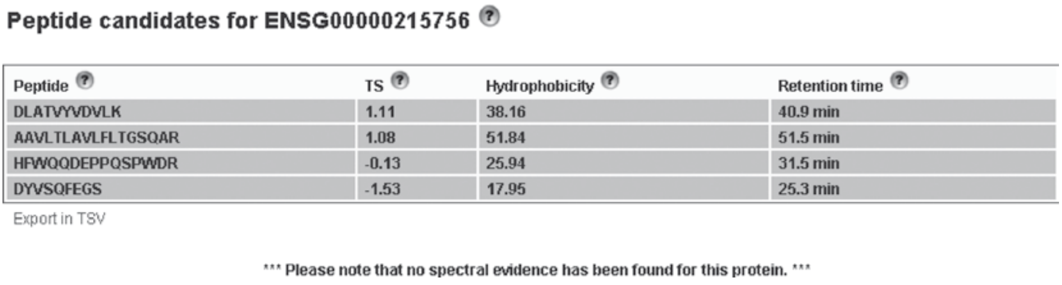
As the name suggests, MRMaid can be used to produce a shortlist of candidates for several protein targets to be monitored in an MRM experiment. However, MRMaid is essentially a single reaction monitoring (SRM) design tool which can be used for MRM design by combining the results of several rounds of transition design. The key to this is the retention time associated with each peptide, which is calculated using an established algorithm (13) and is valid for peptides up to approximately 20 amino acids in length. Because these retention times are predicted, these may not be totally accurate, but by selecting peptides whose retention times are notably different makes designing an MRM experiment possible using MRMaid. We suggest downloading the results for each protein target, and combining the best candidates by cutting and pasting into a single spreadsheet. Retention time of each peptide can then be used to order this shortlist and identify overlaps, before finally validating them in MS.

4. Notes

4.1. No Transitions Found

A frustrating experience when using MRMaid is when no transition candidates can be found for the protein of interest. A transition can only be predicted if at least one suitable peptide candidate can be found for the given protein. In some cases, there is no such peptide due to the nature of the protein sequence. Lack of suitable candidates can also occur if the MRMaid search criteria are set to a very high level of stringency – with many of the options selected. In this case, it is less likely that a suitable candidate will be found because the peptide must pass many criteria – one or more of which may rule it out. To avoid this happening, searches must be performed with different search settings to find the optimum for the protein target.

MRMaid's transition score (TS) is reliant on the availability of MS/MS spectral evidence in the GAPP database. Some proteins may not have been identified by the GAPP pipeline at the time of search and as a result, no spectral evidence is available at that time. In this situation, MRMaid can still generate a shortlist of peptide candidates, but cannot predict product ion candidates. When this scenario occurs, the peptide results are shown as a table with blue rows, with a warning message, as shown in Fig 4. As one would expect, the TS values are considerably lower than those with a spectral evidence component available (compare to Fig. 2, for example), but may still be informative when used as a



Peptide candidates for ENSG00000215756 ?

Peptide ?	TS ?	Hydrophobicity ?	Retention time ?
DLATVYVDVLK	1.11	38.16	40.9 min
AAVLTLAVLFLTGSQAR	1.08	51.84	51.5 min
HFWQQDEPPQSPWDR	-0.13	25.94	31.5 min
DYVSQFEGS	-1.53	17.95	25.3 min

Export in TSV

*** Please note that no spectral evidence has been found for this protein. ***

Fig. 4. List of theoretical peptide candidates returned when insufficient experimental evidence is available in GAPP database. The shaded rows are colored *blue* on the website. These predictions are based on the default chromatography condition (Microflow method 1)

relative comparison between the available candidates. The ultimate solution to this problem is to upload data to GAPP from your own experiments to increase the available evidence for the prediction. A link on the left toolbar leads you through the process of registering as a GAPP data submitter should you wish to do this.

4.2. Other Transition Design Tools

Proteome bioinformatics is a growing field, so the range and quality of tools, such as those for MRM and quantitative proteomics in general, are increasing rapidly. As such, MRMaid is not the only tool for designing transitions for MRM. Other include vendor-specific software packages, such as Applied Biosystem's MIDAS™ (MRM-initiated detection and sequencing) Workflow Designer software (ABI, Foster City, CA), which calculates theoretical peptides and corresponding MRM transitions, then builds the MIDAS acquisition method (15) using a Q-TRAP (ABI) that iteratively cycles through scans to select suitable peptides.

TIQAM (Targeted Identification for Quantitative Analysis by MRM) (16) is one early example of a freely available tool for designing transitions, which applies a similar approach to MRMaid although TIQAM uses evidence from the PeptideAtlas repository (17), a system which is described elsewhere in this book. When no data is available, all theoretically possible proteotypic peptides are computed using physiochemical properties alone. Based on user preferences, transitions are generated for the peptide list. In a final step, MS/MS is performed and the results mapped on to the list of transition candidates.

TIQAM fundamentally differs from MRMaid because it usually includes a step where the user experimentally acquires their own MS/MS data to select suitable candidates from all possible transitions. In most cases, MRMaid is able to provide a shortlist of transitions without this step as transitions mined from existing MS/MS evidence are ranked using the transition scoring algorithm. MRMaid is also distinct from TIQAM because it is a web-based service that does not need to be downloaded and installed locally.

Since developing MRMaid, there have been further additions to the field of automated transition design; newly available tools include: MRMer (18), skyline (19), MRM worksheet (20), and MaRiMba (21), each having a slightly different approach for predicting transition candidates. These tools are reviewed elsewhere (22)

4.3. Ensembl Mappings

One of the drawbacks of using the Ensembl database (23) as the framework for the GAPP pipeline and database is that it contains multiple gene entries for some protein sequences. This can occasionally cause a problem when using MRMaid, because in some rare cases both identifiers are picked up by the search; take Apolipoprotein-C-III (APOC3_HUMAN), for example, which can be mapped to both ENSG00000215755 and ENSG00000110245.

MRMaid detects both Ensembl gene entries for this protein because PICR maps the single Swissprot id to two entries in Ensembl. In these cases, it is recommended that you view MRMaid's predictions for both accession numbers to make sure all available results are accounted for.

References

1. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., Gygi, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *PNAS.* **100**, 6940–6945.
2. Kay, R. G., Gregory, B., Grace, P. B., Pleasance, S. (2007) The application of ultra-performance liquid chromatography/tandem mass spectrometry to the detection and quantitation of apolipoproteins in human serum. *Rapid Comm Mas Spectrom.* **21**, 2585–2593.
3. Keshishian, H., Addona, T., Burgess, M., Kuhn, E., Carr, S. A. (2007) Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Mol Cell Proteomics.* **6**, 2212–2229.
4. Barnidge, D. R., Dratz, E. A., Martin, T., Bonilla, L. E., *et al.* (2003) Absolute quantification of the G protein-coupled receptor rhodopsin by LC/MS/MS using proteolysis product peptides and synthetic peptide standards. *Anal Chem.* **75**, 445–451.
5. Zhang, F., Bartels, M. J., Stott, W. T. (2004) Quantitation of human glutathione S-transferases in complex matrices by liquid chromatography/tandem mass spectrometry with signature peptides. *Rapid Comm Mass Spec.* **18**, 491–498.
6. Wolf-Yadlin, A., Hautanleml, S., Lauffenburger, D. A., White, F. M. (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *PNAS.* **104**, 5860–5865.
7. Mead, J. A., Bianco, L., Ottone, V., Barton, C., *et al.* (2008) MRMaid: the web-based tool for designing multiple reaction monitoring (MRM) transitions. *Mol Cell Proteomics.* E-pub 15th Nov, M800192-MCP800200.
8. Shadforth, I., Xu, W., Crowther, D., Bessant, C. (2006) GAPP: a fully automated software for the confident identification of human peptides from tandem mass spectra. *J. Proteome Res.* **5**, 2849–2852.
9. Craig, R., Beavis, R. C, (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry.* **17**, 2310–2316.
10. Shadforth, I., Dunkley, T., Lilley, K., Crowther, D., Bessant, C. (2005) Confident protein identification using the average peptide score method coupled with search-specific, ab initio thresholds. *Rapid Communications in Mass Spectrometry.* **19**, 3363–3368.
11. Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., *et al.* (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotech.* **25**, 125–131.
12. Côté, R. G., Jones, P., Martens, L., Kerrien, S., *et al.* (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics.* **18**, 401.
13. Krokhin, O. V., Craig, R., Spicer, V., Ens, W., *et al.* (2004) An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: Its application to protein peptide mapping by off-line HPLC-MALDI MS. *Mol Cell Proteomics.* **3**, 908–919.
14. Anderson, L., Hunter, C. L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol Cell Proteomics.* **5**, 573–588.
15. Unwin, R. D., Griffiths, J. R., Leverentz, M. K., Grallert, A., *et al.* (2005) Multiple reaction monitoring to identify sites of protein phosphorylation with high sensitivity. *Mol. Cell. Proteomics.* **4**, 1134–1144.
16. Lange, V., Malmstrom, J., Didion, J., King, N. L., *et al.* (2008) Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. *Mol Cell Proteomics.* **7**, 1489–1500.
17. Deutsch, E. W., Lam, H., Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows *EMBO Rep.* **9**, 429–434.
18. Martin, P. B, Holzman, T., May, D., Peterson, A. et al. (2008) MRMet: An interactive open-source and cross-platform system for data

extraction and visualization of multiple reaction monitoring experiments. Mol Cell. Proteomics **7**, 2270–2278

19. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B. et al. (2009) Expediting the development of target SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res **8**, 2733–2739.

20. Walsh, G.M., Lin, S., Evans, D.M., Khosrovi-Eghbal, A., et al. (2009) J. Proteomies **72**, 838–852

21. Sherwood, C., Eastham, A., Peterson, A., Eng, J.K. et al. (2009) MaRiMba: a software application for spectral library-based MRM transition list assembly J. Proteome Res. **8**, 4396–4405

22. Cham, J.A., Bianco, L., Bessant, C. (2010) Free computational resources for designing selected monitoring (SRM) transitions Proteomics (in press)

23. Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., *et al.* (2004) An Overview of Ensembl. *NAR*. **14**, 925–928.

Chapter 14

OpenMS and TOPP: Open Source Software for LC-MS Data Analysis

Knut Reinert and Oliver Kohlbacher

Abstract

The automatic analysis of mass spectrometry data is becoming more and more important since increasingly larger datasets are readily available that cannot be evaluated manually. This has triggered the development of several open-source software libraries for the automatic analysis of such data. Among those is OpenMS together with TOPP (The OpenMS Proteomics Pipeline). OpenMS is a C++ library for rapid prototyping of complex algorithms for the analysis of mass spectrometry data. Based on the OpenMS library, TOPP provides a collection of tools for the most important tasks in proteomics analysis. The tight coupling of OpenMS and TOPP makes it easy to extend TOPP by adding new tools to the OpenMS library. We describe the overall concepts behind the software and illustrate its use with several examples.

Key words: Bioinformatics, Data analysis, Proteomics, Open-source software, Workflows

1. Introduction

Mass spectrometry is an essential analytical technique for high-throughput analysis in proteomics and metabolomics, both of which produce large amounts of data usually not amenable for manual inspection. The development of new separation techniques, precise mass analyzers, and novel experimental protocols is a very active field of research which leads to new, complex experimental setups. Consequently, analysis of the data is currently often the bottleneck for experimental studies. Although software tools for many data analysis tasks are available today, they are often difficult to combine with each other or not flexible enough to allow for rapid prototyping of a new analysis workflow. Hence, there is a need for software systems that (a) allow developers to quickly

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_14, © Humana Press, a part of Springer Science+Business Media, LLC 2010

implement novel algorithms and (b) to allow bioinformaticians and experimentalists to construct complex workflows using algorithmic components with welldescribed interfaces and functionality.

In this chapter, we describe software that addresses both points. First, OpenMS, which is a software framework for rapid application development in mass spectrometry. OpenMS has been designed to be portable, easy-to-use, and robust while offering a rich functionality ranging from basic data structures to sophisticated algorithms for data analysis. This has already been demonstrated in several studies. Second, we describe a set of tools for proteomics data analysis – TOPP, The OpenMS Proteomics Pipeline. TOPP provides a set of computational tools which can be easily combined into analysis pipelines even by nonexperts and can be used in proteomics workflows. These applications range from small utilities (file format conversion, peak picking) to wrapper applications for known applications (e.g. Mascot) to completely new algorithmic techniques for data reduction and analysis.

2. Concepts

OpenMS is a C++ framework that includes basic data types (peak lists, two-dimensional HPLC-MS maps) as well as methods to manipulate and search in the data structures (e.g. iterate over the spectra in a map). In addition, it contains algorithmic components to process the data, including – among other things – algorithms for signal processing quantitation, and protein identification. It is easy for a C++ programmer to add functionality to OpenMS because it is well documented, adheres to coding conventions, and provides guidelines and tutorials for the programmer not acquainted with the framework.

However, for users not fluent in C++ or without a more formal software engineering background, this might pose a considerable hurdle possibly too high to clear. Here, TOPP comes into play. TOPP provides standalone tools for the most important algorithmic components. Hence, it is a tool for end users and developers of analysis pipelines in the lab. TOPP has standardized input and output formats using XML, and hence they can be easily linked. On top of this, OpenMS is designed to easily allow addition of new TOPP tools.

As a simple example, consider the problem of subtracting a baseline from an MS/MS spectrum. Using the algorithmic components in OpenMS, you can write the C++ program depicted in Fig. 1. The program loads the data in mzData format (1), then it defines a morphological filter (TopHatFilter),

```
RawMap exp_raw;
RawMap exp_filtered;

MzDataFile mzdata_file;

mzdata_file.load("../TEST/data.PeakPicker_test.mzData");

TopHatFilter th;
Param param;
param.setValue("struc_elem_length",1.0);
th.setParameters(param);
```

Fig. 1. C++ code piece to subtract a baseline from an MS/MS spectrum

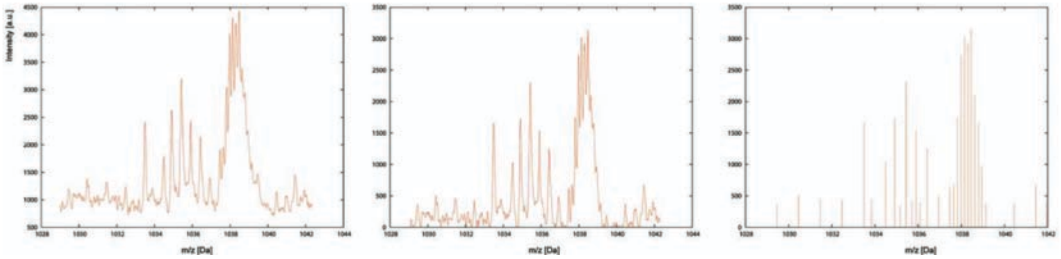


Fig. 2. The figure shows a raw spectrum on the left, the result of baseline filtering in the middle, and peaks resulting from peak picking on the right (from (13)).

sets a parameter and applies it to the data in exp_raw. The result is stored in exp_filtered.

The functionality of this code is wrapped into a standalone TOPP tool BaselineFilter. This tool – as well as all other TOPP tools – can be configured via an XML file. This illustrates the basic idea of the TOPP/OpenMS combination. The result of the baseline filter is shown in Fig. 2. On the left, one can see a spectrum with a baseline, in the middle the same spectrum is depicted with the baseline subtracted.

Apart from the baseline filter, there are numerous other daily tasks that have been programmed in OpenMS and subsequently offered as a TOPP tools. Fig. 3 shows the current state of the TOPP package. Please refer to the OpenMS website (www.OpenMS.de) for the most current release and the full list of tools in OpenMS.

TOPP also offers a convenient editor for parameters, which is called INIFileEditor. This allows convenient editing and storing of the XML based parameter sets. Taking all this into account, TOPP is ideal to quickly string together an analysis pipeline based on standard data formats and XML based configuration files.

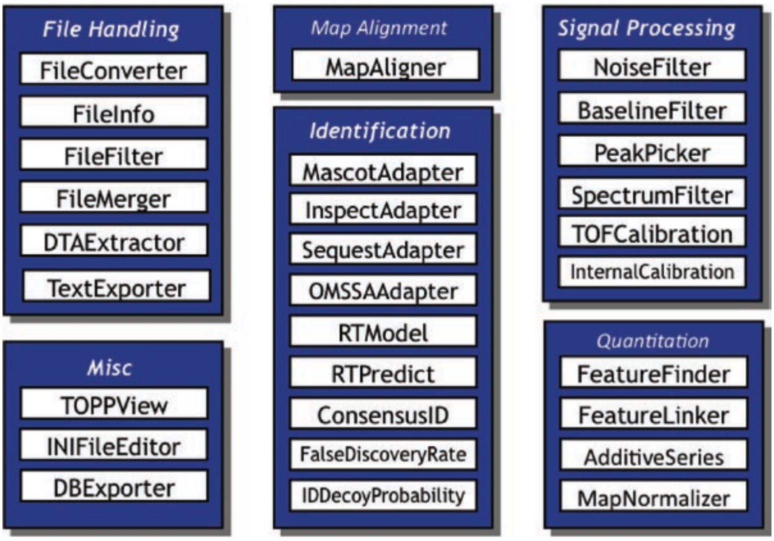


Fig. 3. Some of the available TOPP modules. The list is being constantly expanded

3. Methods

In the following, we will go into more detail concerning some of the main aspects of TOPP.

3.1. Data Management

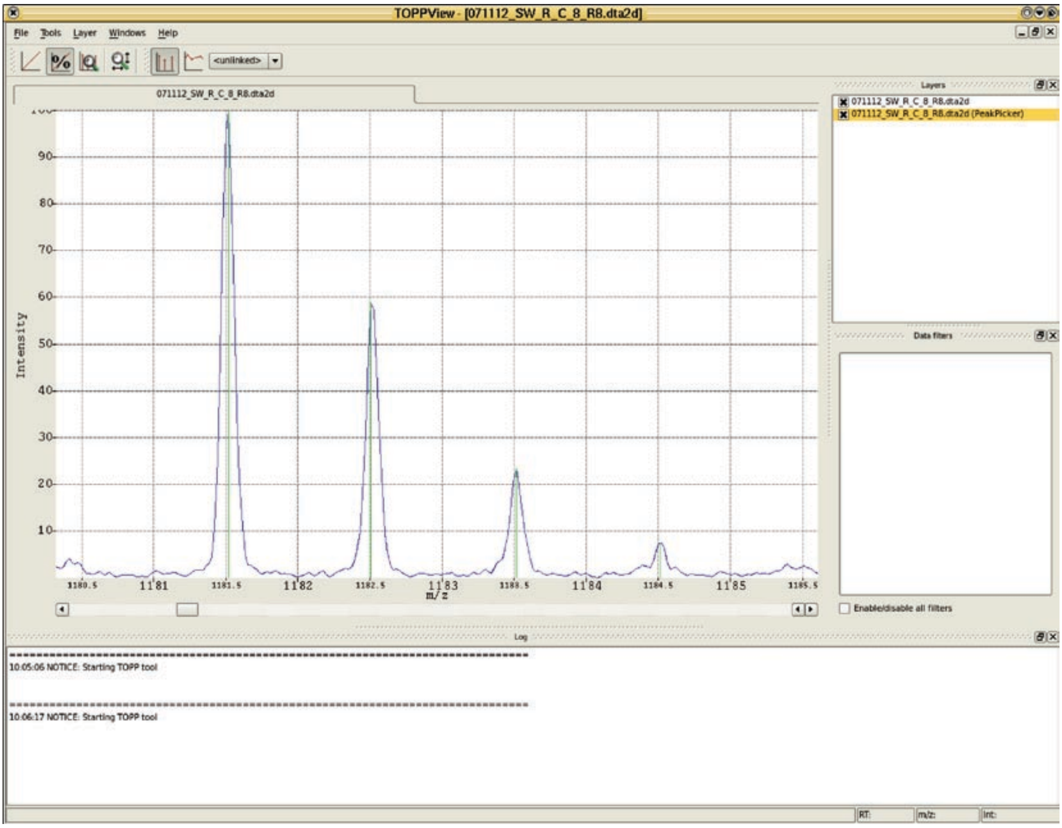
While analyzing mass spectrometry data, a lot of problems arise early on, especially when handling the input. Different vendors still have proprietary data formats and also store different kinds of information. OpenMS and TOPP, therefore, put a lot of emphasis on standard data formats developed by the Human Proteome Organization's Proteomics Standard Initiative (HUPO PSI, see http://www.psidev.info). Vendor-specific formats are not supported due to license restrictions; however, for all instrument vendors, software is readily available to convert the data into PSI formats. Fig. 4 shows the OpenMS code for converting between two standard formats. The TOPP tool FileConverter allows the user to do various conversions between formats.

3.2. Visualization and Data Analysis

While OpenMS and TOPP aim at automatic, high-throughput data analysis, it is still indispensable for researchers to visualize data and computational results. For this, TOPP offers the viewer/ editor TOPPView, which has two main goals. The first goal is to allow the user to browse through one- and two-dimensional MS data. For example, Fig. 5 shows a zoomed-in part of a one-dimensional spectrum. The window on the right shows the two layers of data that can be individually selected for display, one layer depicting the raw data, the other picked peaks.

```
 Int main(){

      MzXMLFile mzxml;
      MzDataFile mzdata;

      // temporary data storage
      MSExperiment<RawDataPoint1D> map;

      // convert MzXML to MzData
      mzxml.load("Tutorial_FileIO.mzXML",map);
      mzdata.store("Tutorial_FileIO.mzData",map);

      return 0;
} //end of main
```

Fig. 4. Code example for converting an mzXML file into an mzData file



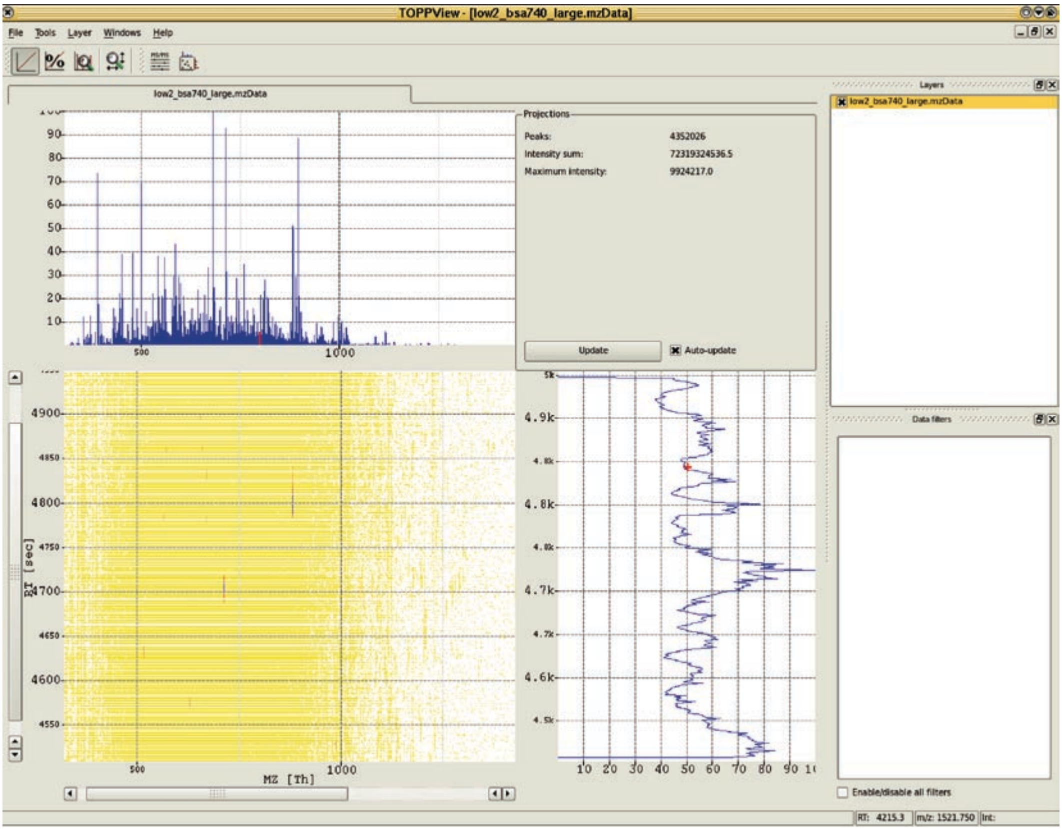
☒ 071112_SW_R_C_8_R8.dta2d
☒ 071112_SW_R_C_8_R8.dta2d (PeakPicker)

□ Enable/disable all filters

```
================================================================================
10:05:06 NOTICE: Starting TOPP tool

================================================================================
10:06:17 NOTICE: Starting TOPP tool
```

Fig. 5. Screenshot of one-dimensional MS data.

In Fig. 6, you can see a screenshot of a two-dimensional map; TOPPView allows zooming in and out, to navigate in the map, and – if needed – to project the two-dimensional information onto



Projections

Peaks: 4352026

Intensity sum: 72319324536.5

Maximum intensity: 9924217.0

☒ low2_bsa740_large.mzData

Fig. 6. Visualization of a two-dimensional map. In the lower left part is the plot in m/z and retention time dimension. To the right and on the top, the cumulative projections are depicted

the corresponding one-dimensional axes (retention time and mass over charge).

A second use of TOPPView is to actually apply TOPP components to the data visualized in the program. For example, the user can read a scan from a file, visualize it, and apply the TopHatFilter tool to it. He can then examine the result and afterward call the PeakPicker TOPP tool. In this way, the user can, for example, interactively explore optimal parameter settings which he can in turn store in a file for subsequent automatic analysis.

3.3. Peptide Identification

Peptide identification (either de novo or database driven) algorithms are probably among the best-known and most widely researched algorithms in the field of proteomics. The current commercial standard tools for this task are MASCOT (2) and SEQUEST (3). Over recent years, they have been supplemented by various tools of comparable performance that are free for academic use. Among these tools are X!Tandem (4), OMSSA (5), InsPecT (6), and Phenyx (7). A key problem with using these tools is the very different interface. Each tool accepts different input formats,

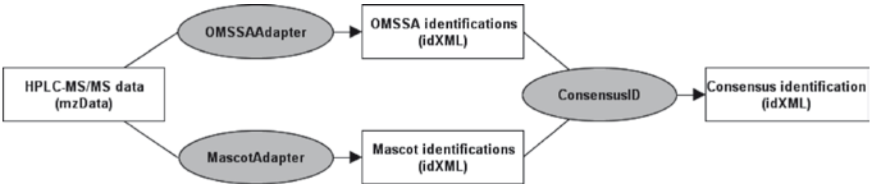

OMSSAAdapter
OMSSA identifications (idXML)
HPLC-MS/MS data (mzData)
ConsensusID
Consensus identification (idXML)
MascotAdapter
Mascot identifications (idXML)

Fig. 7. Dataflow of an mzData file through two ID engines with subsequent combination of the results to increase confidence in the identification

requires different parameters, and has different output formats. This requires a considerable effort when replacing one identification tool with another or adding an identification tool when more than one is necessary.

To simplify this task, TOPP implements unified adapters for various database search tools. Currently, TOPP offers adapters for MASCOT, SEQUEST, InsPecT, X!Tandem, OMSSA, and PepNovo. The adapter translates the query spectra into the tool-specific input formats, runs the search engine (which is not part of TOPP and needs to be installed and licensed separately), retrieves the output, and translates it into a common file format. Search engines are thus encapsulated in these adapters and can be easily exchanged.

In addition, TOPP offers a consensus identification engine that can take the results of different search engines and combine them, thus increasing the identification rate. The following figure (Fig. 7) shows how this could be set up using standard XML data formats.

3.4. Quantification

Another central part of OpenMS and TOPP is the differential analysis of protein expression. Most differential analysis methods today share the following steps: first, a *separation* of the proteins or peptides, then the *quantification* of single charge variants and a *normalization* of the samples with respect to intensity. This is usually followed by *relative quantification* using mass spectrometry, which requires a preceding *matching* of corresponding peptides as well as their *identification*. Fig. 8 shows a three-dimensional view of parts of two MS measurements. The peptides are separated through different retention time and mass over charge. The task is to detect the individual peptides – which we call *features* (four are circled in each map) – then assign the corresponding ones in each map (depicted by the arrows between the circles), and finally to compare their normalized intensities.

Feature detection is a central concept in OpenMS. As noted earlier, a feature is a signal in an HPLC-MS map which is caused by a peptide ion. OpenMS includes several algorithms for the detection of peptide features in HPLC-MS data, tailored for datasets of different mass resolutions and measured on various instrument types.

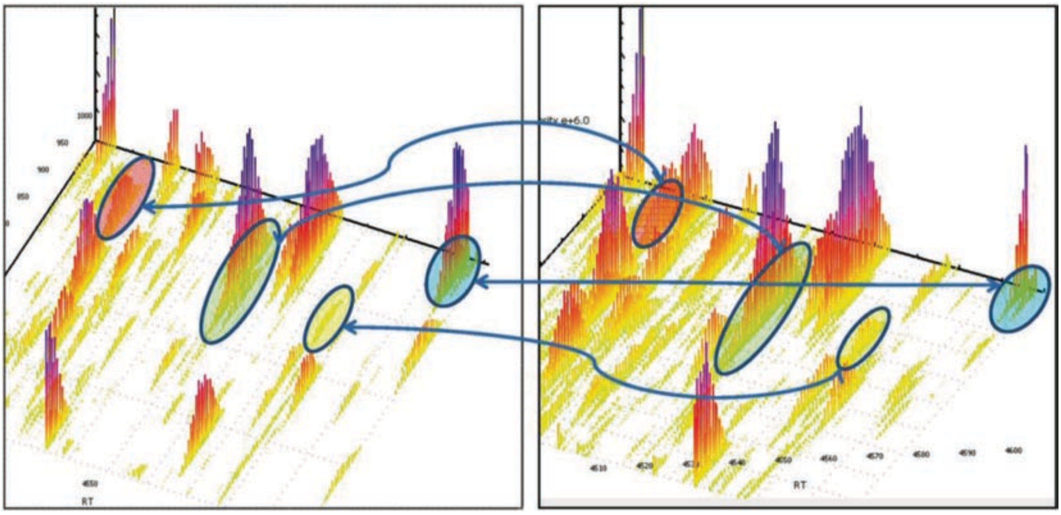


Fig. 8. Label-free quantification requires an accurate identification and mapping of features

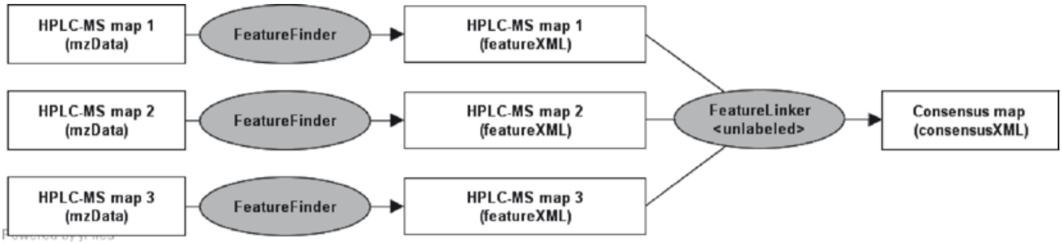


Fig. 9. A simple workflow for label-free quantification based on TOPP tools

After feature detection, the next important step in a typical analysis workflow is the combination of results from multiple experiments, for example, to improve confidence in the obtained measurements or to compare results from different samples. To this end, a suitable mapping or alignment between the datasets needs to be established. The alignment has to correct for random and systematic variations in the observed elution time and mass-to-charge ratio that are inevitable in experimental datasets. OpenMS offers algorithms to align multiple experiments and to match the corresponding ion species across many samples. A novel and generic algorithm has been developed to correct for the variation of retention time and mass-to-charge dimensions between two maps. It uses an adapted pose-clustering approach (8) to efficiently superimpose raw maps as well as feature maps.

This sort of analysis readily translates into a simple pipeline built with TOPP tools (Fig. 9). Two raw data maps are given to the FeatureFinder tool, which identifies, and quantifies the peptidic charge variants and stores the results in two feature maps using the XML-based format featureXML. Then distortions in

```
for i in `seq 1 32`; do
     # Truncate raw data maps to save time
     FileFilter -ini AddSeries.ini -instance $i
     # Collect peptde feature
     FeatureFinder -ini AddSeries.ini -instance $i
done

# Compute optimal matching of the maps
MapAligner –ini AddSeries.ini

# Find corresponding features in all maps
FeatureLinker –ini AddSeries.ini

# Compute final concentration
AdditiveSeries -ini AddSeries.ini
```

Fig. 10. Code for additive series analysis pipeline

retention time are accounted for and corresponding features in the two measurements are mapped onto each other using the FeatureLinker tool in TOPP and the resulting pairs are stored in another XML-based format (consensusXML). Similar pipelines can be constructed if, for example, repeated measurements need to be combined to increase the statistical significance of the result.

Just how easy it is to string together different TOPP modules can also be seen in the script depicted in Fig. 10, which is a workflow used in a study to quantify the absolute content of myoglobin in human serum (for more details see (9,15)). The script first cuts out relevant portions of the maps, and then it identifies features in each map. Finally, a regression analysis (AdditiveSeries) determines the concentration of the myoglobin.

4. Notes

1. Installation. Since many potential users will be relying on the Microsoft Windows platform, we provide automated installers for precompiled binaries for Microsoft Windows. A package installer is available for MacOS X. The installation of these is self-explanatory; nevertheless, there is additional assistance available on the OpenMS webpage. Installation of the full package including the source code is a bit more involved and is sketched below for Linux, where no binary distribution is available.

   OpenMS and TOPP depend on a number of other Open Source libraries, like ANDI/MS with NetCDF (for platform-independent data storage). The GNU scientific library (GSL, see http://www.gnu.org/software/gsl/), the CGAL library for geometric algorithms (http://www.cgal.org/), XERCES-C for parsing XML (http://xerces.apache.org/xerces-c/), the libSVM (http://xerces.apache.org/xerces-c/), the SeqAn library for sequence analysis (http://www.seqan.de/) as well as Qt (http://trolltech.com/products/qt/) for visualization and mySQL database access. These support libraries are bundled together with the OpenMS source code. Building and installing OpenMS thus requires the following steps:

(a) Download the source code archive from the OpenMS website and unpack to an arbitrary directory. Files will reside in a subdirectory named OpenMS.

(b) Change directory to OpenMS/contrib.

(c) Execute the commands "./configure" and "make" to build the contributed libraries.

(d) Change directory to OpenMS/source.

(e) Issue the commands "./configure", "make", and "make install" to build the OpenMS libraries.

(f) Add the path to the dynamic library files (…OpenMS/lib) to the LD_LIBRARY_PATH environment variable.

Once OpenMS is installed and compiled, it is easy to build the TOPP tools:

(a) Change the directory to <path_to_OpenMS>/source/

(b) Write "make TOPP"

(c) To test the tools, write "make TOPPtest"

2. Support and Assistance. If you have problems with OpenMS or TOPP, you should first read the manual and check whether it can solve your problem. If not, there is a mailing list to which every user may subscribe. If you find a bug in OpenMS or TOPP, you can use a bug tracker to submit it to the OpenMS development team, who will address the issue as soon as possible.

3. Integration with other tools. With the establishment of additional standardized (XML-based) file formats, the interoperability of software packages from diverse sources will increase. mzML and analysisXML from the Proteomics Standards Initiative of the HUPO are important milestones in that respect. This will also allow the integration of TOPP with other software packages. The Trans Proteomics Pipeline (TPP) (10), ProteoWizard (11), and SuperHirn (12) are just some examples of recent alternative projects for promising tools in the area of proteomics data analysis. Joint file formats will render the integration of these tools from different sources virtually seamless allowing the user to profit from up-to-date developments from a wide range of groups worldwide.

Acknowledgments

The authors wish to thank Andreas Bertsch, Marc Sturm, Clemens Gröpl, and Ole Schulz-Trieglaff for assistance with the figures and critical comments on the manuscript, and the whole OpenMS team for their support.

References

1. Orchard S, Hermjakob H, Taylor CF, Potthast F, Jones P, Zhu W, Julian RK Jr, Apweiler R. Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17-20th April 2005). Proteomics (2005), 5(14):3552–5.
2. Perkins, D. N. and Pappin, D. J. C. and Creasy, D. M. and Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999, 20, 3551–3567.
3. Eng, J.K., McCormack, A.L. and Yates, III, J.R. J. Am. Soc. Mass Spectrom, 1994, 5, 976–989.
4. Robertson Craig and Ronald C. Beavis, Bioinformatics, 2004, 20, 1466–7.
5. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. J Proteome Res. 2004 Sep–Oct; 3(5):958–64.
6. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V: InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 2005, 77(14):4626–4639.
7. Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. Proteomics , 2003, 3, 1454–1463
8. Lange, E, Gröpl, C, Schulz-Trieglaff, O, Reinert, K. A Geometric Approach for the Alignment of Liquid Chromatography-Mass Spectrometry Data. In: Proceedings of the 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB). pp. i273–i281, 2007.
9. Mayr, B, Kohlbacher, O, Reinert, K, Sturm, M, Gröpl, C, Lange, E, Klein, C, and Huber, CG (2006). Absolute myoglobin quantitation in serum by combining two-dimensional liquid chromatography-electrospray ionization mass spectrometry and novel data analysis algorithms. J. Proteome Res. 5:414–421.
10. Keller, A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol. (2005), 1:2005.0017.
11. Kessner, D, Chambers, M, Burke, R, Agus, D, Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. Bioinformatics (2008).
12. Mueller, Lukas N, Rinner, Oliver, Schmidt, Alexander, Letarte, Simon, Bodenmiller, Bernd, Brusniak, Mi-Youn, Vitek, Olga, Aebersold, Ruedi, Müller, Markus: SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling, Proteomics, 2007 vol. 7 (19) pp. 3470–80.
13. Sturm, M, Bertsch, A, Gröpl, C, Hildebrandt, A, Hussong, R, Lange, E, Pfeifer, N, Schulz-Trieglaff, O, Zerck, A, Reinert, K, and Kohlbacher, O (2008). OpenMS - An open-source software framework for mass spectrometry. BMC Bioinformatics 9:163.
14. Kohlbacher, O, Reinert, K, Gröpl, C, Lange, E, Pfeifer, N, Schulz-Trieglaff, O, and Sturm, M (2007). TOPP - The OpenMS Proteomics Pipeline. Bioinformatics 23(2):e191–e197.
15. Gröpl, C, Lange, E, Reinert, K, Kohlbacher, O, Sturm, M, Huber, C, Mayr, B, and Klein, C (2005). Algorithms for the automated absolute quantification of diagnostic markers in complex proteomics samples. In: Proceedings of the 1st Symposium on Computational Life Sciences (CLS 2005), edited by M. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, I. Fischer. Springer LNBI 3695, pages 151–161.

Chapter 15

Trans-Proteomic Pipeline: A Pipeline for Proteomic Analysis

Patrick G.A. Pedrioli

Abstract

Mass spectrometry has quickly become an essential tool in molecular biology laboratories. Here, we describe the Trans-Proteomic Pipeline, a collection of software tools, to facilitate the analysis, exchange, and comparison of MS data. The pipeline is instrument-independent and supports most commonly used proteomics workflows, including quantitative applications such as ICAT, iTRAQ, and SILAC. Importantly, the pipeline uses open, standard data formats and calculates accurate estimates of sensitivity and error rates, thus allowing for meaningful data exchange. In this chapter, we will introduce the various components of the pipeline in the context of three typical proteomic use-case scenarios.

Key words: TPP, Trans-Proteomic Pipeline, mzXML, PeptideProphet, ProteinProphet, XPRESS, ASAPRatio, Libra, Pep3D, QualScore, Proteomics data analysis, ICAT, iTRAQ

1. Introduction

The Trans-Proteomic Pipeline (TPP) is an open source project of the Seattle Proteome Center (SPC), which includes valuable contributions from external collaborators such as the Fred Hutchinson Cancer Research Center, Insilicos Life Science Software, and LabKey Software (see Note 1). The TPP aims at providing an easy way to analyze, compare, publish, and exchange proteomics data in an MS-independent, standardized and reproducible fashion. To achieve these goals, it leverages open data formats and accurate sensitivity/error rates calculation for peptide and protein assignments. The TPP includes multiple programs which perform distinct tasks. These components can be arranged according to the analytical workflow requirements, to build a pipeline in which the output from one is taken as the input for the next one. The TPP relies on, and integrates in its workflow, external search engines (e.g., SEQUEST (1), MASCOT (2), Phenyx (3), Comet

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_15, © Humana Press, a part of Springer Science+Business Media, LLC 2010

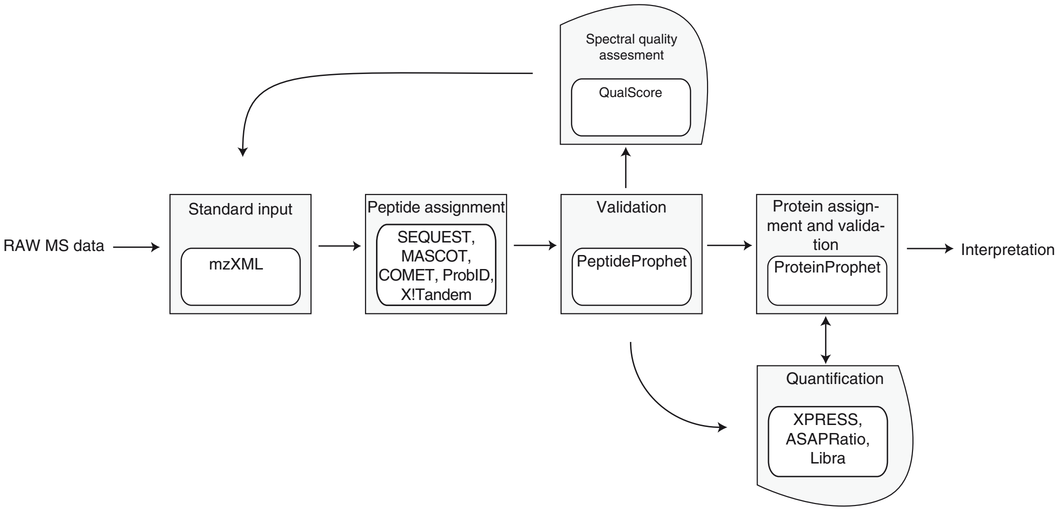


Fig. 1. Simplified TPP software components overview. MS instrument output is first converted to the uniform mzXML format. Fragmentation spectra are then matched to putative peptides by a database search engine. Next, PeptideProphet calculates sensitivity and error rates of these putative assignments. ProteinProphet uses the output from PeptideProphet to derive the list of proteins most likely to have been present in the original sample. In the case of quantitative experiments, XPRESS, ASAPRatio, and Libra calculate the relative abundance of peptides and proteins. Finally, high-quality unassigned fragmentation spectra are identified and extracted by QualScore so that they can be re-searched in a second, more comprehensive, database search

(4), ProbID (5), and X!Tandem (6)) while focusing on tasks other than the spectral matching step itself.

The major TPP components are represented in Fig. 1. First, the raw MS-generated data are converted into a standardized, open, XML-based representation. This helps maintaining compatibility between different MS instruments and facilitating raw data exchange by removing the requirement of corresponding acquisition software to read them. Although the TPP native format for raw MS data is mzXML (7), mzData and mzML formats (8) are also supported. After the database search, Peptide- and Protein-Prophet (9, 10) assign probabilities to the peptide and protein assignments, respectively. Corresponding sensitivity and error rates are also calculated, enabling meaningful comparison of results between multiple experiments and laboratories. The “Prophets” store their analyses in two open formats called pepXML and protXML. XPRESS (11), ASAPRatio (12), and Libra can be inserted in the pipeline flow to analyze different types of quantitative proteomic experiments. Pep3D (13) visualizes microcapillary reverse-phase HPLC electrospray tandem mass spectrometry (μLC-ESI-MS-MS/MS) data as a two-dimensional density plot, providing a simple method to evaluate and troubleshoot MS data acquisition and analysis. QualScore identifies high-quality, unassigned fragmentation spectra and isolates them so that they can be searched during a second, more comprehensive, protein database search. Finally, Pep- and Prot-XML Viewers provide a way to visualize, filter, explore, and export the results of the TPP analysis.

The Methods section will discuss how to set up and start a TPP analysis for three typical proteomic data analysis workflows: (1) Identification of proteins from a low complexity sample; (2) Identification and quantification of proteins from multi-fraction isotopically labeled samples (e.g., ICAT); and (3) Identification and quantification of proteins based on isobaric reporter ions (e.g., iTRAQ). The end of the Methods section will also describe how to validate and read the output of the tools that compose the TPP.

2. Materials

2.1. Stylistic Conventions

Commands to be entered at the DOS prompt (see Notes 2 and 3) are indicated with the following style:

```
> command.exe parameter1 parameter2 file1 file2
```

A command to be entered in a Unix, or Cygwin shell will use the following style:

```
$ command parameter1 parameter2 file1 file2
```

> and $ indicate the beginning of the line in MS-DOS and the Unix shell respectively (i.e., commands printed over multiple lines that do not start with either the > or $ sign, should be entered as one single line). > and $ should be omitted when executing the command.

XML elements are indicated with the following style:

```
<Element attribute="This is an XML element with one attribute">Important parts are in bold </Element>
```

TPP_ROOT refers to the TPP root installation directory (i.e., the one under which the TPP tools are installed) and varies from installation to installation. Its exact location can be found by typing:

```
$ which xinteract
```

which returns the location of the TPP component xinteract (e.g., `/usr/local/tpp/bin/xinteract`). The TPP_ROOT will be equal to the path up to, and not including, the bin directory. Whenever you see TPP_ROOT in this chapter, remember to substitute it for its actual value (e.g., /usr/local/tpp/).

We will assume the existence of the following directory tree:

`/mnt/Data/`

`|___ Databases/`	Fasta protein databases
`|___ MMB/`	Data from user MMB
`  |___ Yeast_IP/`	Analysis directory for the first workflow
`  |___ Yeast_ICAT/`	Analysis directory for the second workflow
`  |___ Yeast_iTRAQ/`	Analysis directory for the third workflow

This directory organization was only chosen for our examples and is not a TPP requirement.

2.2. Command Line Essentials

While the Windows version of the pipeline also has a graphical user interface called Petunia, all versions can be run from the command line. Although this latter method might at first sound intimidating, it is the most general and flexible one and will therefore be discussed in this chapter. In addition to the TPP specific ones, some Unix commands will be used in this chapter. Here is a short reference list for your convenience:

`man <command>`	Displays the manual page for command.
`mkdir <dir>`	Makes a new directory called dir.
`cp <src> <dest>`	Copies source file to destination file.
`cd <dir>`	Changes directory to dir.
`ls`	Lists the content of the current directory.
`sudo <command>`	Executes command with super-user (administrator) privileges.
Pattern matching	
*	The asterisk wildcard is used to match any string (e.g., ls *.mzXML would list all files with extension .mzXML present in the current directory)

2.3. Useful Links TPP-Related Links

Internet sites:

http://tools.proteomecenter.org/wiki/index.php?title=Main_Page

Seattle Proteome Center Software Tools Wiki. This is the official wiki page of the SPC and discusses in great detail the TPP as well as its related tools.

http://www.proteomecenter.org

Main site of the SPC. Here, you will find information ranging from the general description of the SPC organization to the announcement of training courses for the TPP.

http://sourceforge.net/projects/sashimi

This site hosts all the code from the TPP.

https://www.labkey.org

Home of LabKey Software, the makers of CPAS, is a web-based system for managing, analyzing, and sharing high volumes of tandem mass spectrometry data that incorporates parts of the TPP.

Mailing lists:

http://groups.google.com/group/spctools-discuss

A moderate-volume mailing list used for TPP related discussions ranging from installation to data processing. If you have any question regarding the TPP, post it here.

http://groups.google.com/group/spctools-announce

A low-volume mailing list used to announce new developments in SPC proteomics tools.

2.4. Hardware and Operating System Requirements

Any Pentium, or higher, computer with at least 512 MB of RAM can run the TPP software. Clearly, the hardware requirements will scale with the analytical ones. Therefore, in a production environment, a more realistic setup will comprise a computer cluster to perform the database searches and TPP analyses, a dedicated web-server to serve the results and a RAID server to store the data and results of the analyses.

The TPP is compatible with *nix, Windows and Mac (although only the first two are officially supported). This tutorial was tested under Ubuntu Server Edition, however, the same commands will also work under other operating systems.

2.5. Getting the TPP

The TPP stable source code is available at "http://sourceforge.net/projects/sashimi/". This tutorial was tested on version 4-0-1. The most recent development version of the TPP can be obtained via anonymous Subversion (SVN) checkout:

```
$ svn co https://sashimi.svn.sourceforge.net/svnroot/sashimi/trunk/trans_proteomic_pipeline sashimi
```

2.6. Software Dependencies

The TPP requires a web-server with read and write access (see Note 4) to the analysis directories and Server Side Includes (SSI) enabled (see Note 5). By default, the TPP is set to use Apache ("http://httpd.apache.org"). GNUPlot version 4.2 or higher is also required ("http://www.gnuplot.info/"), as well as an XSLT Processor such as xsltproc ("http://xmlsoft.org/XSLT/xsltproc2.html") or xalan ("http://xalan.apache.org"). To build the TPP from the source code, the gnu compilers, or Microsoft Visual Studio, as well as development versions of the following libraries will also be required: boost ("http://www.boost.org"); xerces-c ("http://xerces.apache.org/xerces-c"); libgd ("http://www.libgd.org"); libpng ("http://www.libpng.org/"); and zlib ("http://www.zlib.net/").

2.7. Building and Installing the TPP from Source

Detailed compilation and installation instructions are included in the TPP distribution. Additional information can also be found on the SPC wiki page (http://tools.proteomecenter.org/wiki/index.php?title=Main_Page).

Briefly:

Under linux, create a file called Makefile.config.incl in the src directory of the TPP. Use this file to redefine TPP_ROOT (the default setting is: TPP_ROOT=/usr/local/tpp/); TPP_WEB (the default setting is: TPP_WEB=/tpp/); and XSLT_PROC (the default setting is: XSLT_PROC=/usr/bin/xsltproc) values. In the src directory, enter the following commands:

```
$ make configure
$ make all
$ sudo make install
```

Under Windows, move into the src directory of the TPP and type:

```
$ make windows
$ make install-windows
```

3. Methods

3.1. Processing MS Data Using the TPP

3.1.1. Workflow 1: Protein Identification from a Low Complexity Sample

In this first workflow, it is assumed that the following proteomic experiment was performed:

- Grow, harvest, and lyse a yeast culture expressing an endogenously tagged protein.
- Immuno-precipitate the tagged protein and its interactors.
- Reduce and carboxamidomethylate cysteine residues with iodoacetamide.
- Digest the immuno-precipitated proteins with trypsin.
- Analyze the peptides using µLC-ESI-MS-MS/MS.

This strategy will be used to answer the question: “Which proteins interact with the tagged protein?” (see Note 6). The list of proteins most likely to have been present in the immuno-precipitated sample with their associated probabilities and estimated sensitivity and error rates will be derived.

3.1.1.1. Converting the RAW Output of MS Instruments to an Open Representation

The goal of this step is to convert the native output of the MS instrument into a vendor neutral format that can be read by the TPP software tools (i.e., mzXML (7), mzData, or mzML (8)).

1. Instrument-specific converters to the mzXML (see Note 7) format can be found as part of the standard TPP distribution under src/mzXML/converters/. Since these converters often rely on proprietary software libraries provided by the

MS instrument vendors, their use is generally restricted to computers that have such software installed. The choice of the appropriate converter depends on the format of the MS data:

File format	*Converter synopsis*
Thermofinnigan Xcalibur (see Notes 8 and 9)	`ReAdW.exe [OPTIONS] SOURCE [OUTPUT]` example: > `ReAdW.exe -c D:\Data\Xcalibur-data.RAW`
Waters MassLynx (see Notes 10 and 9)	`massWolf.exe [OPTIONS] SOURCE [OUTPUT]` example: > `massWolf.exe -c D:\Data\MassLynx-data-dir.raw\`
ABI/MDS Sciex analyst (see Notes 11 and 9)	`mzWiff.exe [OPTIONS] SOURCE [OUTPUT]` example: > `mzWiff.exe -c D:\Data\Analyst-data.wiff`
Bruker (see Note 12)	`CompassXport.exe -a SOURCE -o DESTINATION` example: > `CompassXport.exe –a D:\Data\Bruker-data.yep -o D:\Data\Bruker-data.mzXML`
SEQUEST .dta (see Notes 13 and 9)	`dta2mzxml [OPTIONS] FILE` example: `$ dta2mzxml -recount *.dta`

2. Once the conversion is completed, transfer the mzXML file to the computer that will perform the TPP analysis.
   Create a directory, hereafter referred to as the analysis directory:

   ```
   $ mkdir /mnt/Data/MMB/Yeast_IP
   ```

   Now copy the mzXML file (for the purpose of this tutorial we will assume it is called yeast_ip.mzXML) into the analysis directory.

At the end of this step, the MS data will have been converted into the mzXML representation and will have been stored in a new directory that will contain all results of the analysis.

3.1.1.2. Protein Database Search

The goal of this step is to assign peptide sequences to the fragmentation spectra represented in the mzXML file.

1. The TPP supports multiple database search engines (e.g., SEQUEST (1), MASCOT (2), COMET (4), ProbID (5), and X!Tandem (6)). Here, we will focus on X!Tandem (see Note 14) with the k-score scoring plug-in developed by the Fred Hutch Cancer Research Center (14), because it is included in recent releases of the pipeline. X!Tandem behavior is controled using three files (see Notes 15 and 16):

default_input.xml	Contains default settings for X!Tandem search parameters. These values can be overridden in input.xml.
taxonomy.xml	Instructs the search engine on where to find the protein sequence databases.
input.xml	Provides a way to override defaults specified in default_input.xml. This is especially practical for sample specific parameters such as the protein sequence database to search and static/variable amino-acid modifications (see Note 17).

2. In the TPP, these files are located in the TPP_ROOT/bin/ directory. The default_input.xml file is called isb_default_input_kscore.xml and has been fine tuned to give the best performances when using the k-score version of X!Tandem in conjunction with the TPP (see Note 18). For the scope of this tutorial, we will not need to edit this file. However, please note that it is human readable and can be modified using any text editor.
3. In the same directory, there is also a file called taxonomy.xml. Using a text editor, look for the following lines:

```
<taxon label="database_identifier_A">
<file format="peptide" URL="full_path_to_database_A"/>
```

   They specify the location of the protein sequence database (second line) and a unique identifier associated with it (first line). Edit them to point to the protein sequence database to be searched. In our example we will use a yeast database:

```
<taxon label="yeast_nci_20070223">
<file format="peptide" URL="/mnt/Data/Databases/yeast.nci.20070223.fasta"/>
```

   Save the changes and close the file (see Notes 19 and 20).
4. Finally, always in the same directory, there is a file, corresponding to input.xml, called "tandem_params.xml". The first time X!Tandem is run after installing the TPP, all instances of "**_DEFAULT_INPUT_LOCATION_**" should be replaced with "**TPP_ROOT/bin/**". This will allow X!Tandem to locate the default_input.xml and taxonomy.xml files you want to use (see Note 21).
5. Copy tandem_params.xml into the analysis directory:

```
$ cp TPP_ROOT/bin/tandem_params.xml /mnt/Data/MMB/Yeast_IP/
```

6. Now edit the copy to match the desired search criteria. Start by specifying that the database associated with the identifier yeast_nci_20070223, as defined in the taxonomy.xml file, should be searched. Replace:

```
<note type="input" label="protein, taxon">protein_database</note>
```

   with:

```
<note type="input" label="protein, taxon">yeast_nci_20070223</note>
```

7. Next, let the search engine know that cysteine residues are carboxamidomethylated. This is equivalent to adding a static modification of 57.021464 Da, such that the monoisotopic mass of cysteine used by the search engine will be 160.03065 instead of 103.00919 Da. X!Tandem specifies static modifications with this element:

```
<note type="input" label="residue, modification mass"></note>
```

The general syntax for a modification is <delta_mass>@<residue> (see Note 22). Therefore, change this line to:

```
<note type="input" label="residue, modification mass">57.021464@C</note>
```

8. Methionine oxidation is a common artifact introduced during sample preparation. To account for it, specify variable methionine oxidation so that the search engine will test two masses for every methionine residue. Variable modifications are specified with this element:

   ```
   <note type="input" label="residue, potential modification mass"></note>
   ```

   using the same syntax as for a static modification:

   ```
   <note type="input" label="residue, potential modification mass">15.994915@M</note>
   ```

9. Finally, specify the name of the input (i.e., the mzXML file to be searched) and output (i.e., the file where X!Tandem will save the results) files by changing these two elements:

   ```
   <note type="input" label="spectrum, path">full_mzXML_filepath</note>
   <note type="input" label="output, path">full_tandem_output_path</note>
   ```

   to:

   ```
   <note type="input" label="spectrum, path">yeast_ip.mzXML </note>
   <note type="input" label="output, path">yeast_ip.xtan.xml</note>
   ```

10. Save the changes and close the file (see Note 23).
11. Move into the analysis directory:

    ```
    $ cd /mnt/Data/MMB/Yeast_IP/
    ```

    and start the search (see Note 24):

    ```
    $ tandem.exe tandem_params.xml
    ```

At the end of this step, a database search against a yeast protein sequence database, using a static modification of 57.021464 Da at cysteine and a variable modification of 15.994915 Da at methionine residues will have been started. The MS/MS scans to be searched will have been extracted from the file called yeast_ip.mzXML and the results will have been stored in a file called yeast_ip.xtan.xml

3.1.1.3. Calculating Probabilities for Peptide and Protein Assignments

The goal of this step is to assign probability values to the peptide and protein assignments made during the previous step.

1. The X!Tandem search will have created "yeast_ip.xtan.xml" in the analysis directory (see Note 25), which needs to be converted into the TPP pepXML format (see Note 26):

```
$ Tandem2XML yeast_ip.xtan.xml yeast_ip_raw.pep.xml
```

This instructs Tandem2XML to extract the information required by the TPP components and to store it in a file called yeast_ip_raw.pep.xml. Using a standardized representation allows the TPP to deal with different search engines in a consistent manner.

2. At this point, the following programs need to be run:

InteractParser	Combines results from multiple searches (e.g. from multi-fraction samples, or repeat runs), into a single pepXML.
PeptideProphetParser	Runs PeptideProphet (9).
RefreshParser	Extracts a list of all proteins corresponding to the identified peptides.
ProteinProphet	Runs ProteinProphet (10).

The TPP wrapper application xinteract (see Note 27) takes care of sequentially launching these tools. Start it with the following command:

```
$ xinteract –Op –Nyeast_ip.pep.xml yeast_ip_raw.pep.xml “
```

–Op” and “–Nyeast_ip.pep.xml” instruct xinteract to run ProteinProphet (PeptideProphet is run by default) and to save the results in a file called “yeast_ip.pep.xml” respectively.

At the end of this first workflow the probabilities, as well as the estimated sensitivity and error rates, for each identified peptide will have been calculated. These will, in turn, have been used to derive the list of proteins, with their corresponding probabilities, most likely to have been present in the original sample. We will further discuss the interpretation of these results in Subheading 3.2.

3.1.2. Workflow 2: Protein Identification and Quantification Based on Elution Profiles in Multi-Fraction Samples

In this second workflow, it is assumed that the following proteomic experiment was performed:

- Harvest and lyse two yeast cultures grown under different conditions. Hereafter referred to as yeast_A and yeast_B.
- Reduce and label the protein samples with acid cleavable ICAT reagent.
- Combine yeast_A and yeast_B and digest the resulting mixture (yeast_A_B) with trypsin.
- Enrich for ICAT labeled peptides.
- Fractionate the digested sample using SCX cation exchange chromatography into 5 fractions (i.e., yeast_A_B_1, yeast_A_B_2, … , yeast_A_B_5).
- Analyze the 5 fractions using μLC-ESI-MS-MS/MS.

This strategy will be used to answer the question: "Which proteins are up-/down-regulated in condition A versus condition B?". The list of proteins most likely to have been present in both samples with their associated probabilities; estimated sensitivity and error rates; and relative abundances will be derived.

3.1.2.1. Converting the Files and Preparing the Analysis Directory

1. As with the first workflow, start by converting the native output from the MS instrument into mzXML as described in Subheading "Converting the RAW output of MS instruments to an open representation".
2. Create an analysis directory and copy all mzXML files (i.e., yeast_A_B_1.mzXML, … , yeast_A_B_5.mzXML) in it:

   ```
   $ mkdir /mnt/Data/MMB/Yeast_ICAT/
   ```

3.1.2.2. Adjusting X!Tandem Input Parameters Files

1. Copy tandem_params.xml from the TPP_ROOT/bin directory into the analysis directory (make sure all instances of "_ **DEFAULT_INPUT_LOCATION_**" have been substituted as described in Subheading "Protein database search."):

   ```
   $ cp TPP_ROOT/bin/tandem_params.xml /mnt/Data
   /MMB/Yeast_ICAT/tandem_params_A_B_1.xml
   ```

   Notice that the name was changed while copying the file. This is connected to the fact that multiple mzXML files are being searched and X!Tandem requires one tandem_params.xml file for each one of them.
2. Edit tandem_params_A_B_1.xml to match the desired search criteria.
3. Let the search engine know that cysteine residues have been labeled with acid cleavable ICAT reagent (see Note 28). Specify a static modification at cysteine residues corresponding to the light reagent (i.e., 227.13 Da) and a variable modification corresponding to the mass difference between the heavy (i.e., 236.14 Da) and light reagents:

   ```
   <note type="input" label="residue, modifica-
   tion mass">227.13@C</note>
   ```

   ```
   <note type="input" label="residue, potential
   modification mass">9.01@C, 15.994915@M</note>
   ```
4. Finally, indicate which mzXML file to search and where to store the results:

   ```
   <note type="input" label="spectrum, path">
   yeast_A_B_1.mzXML</note>
   ```

   ```
   <note type="input" label="output, path">
   yeast_A_B_1.xtan.xml</note>
   ```
5. Save this file, but do not close it.
6. Repeat step 4 by specifying the next SCX fraction (i.e., yeast_A_B_2 as input and output files).

7. Save this file as “tandem_params_A_B_2.xml”.
8. Repeat steps 4–7 until one tandem_params file has been created for each fraction to be analyzed (see Note 29).
9. All files are now ready to start the searches (*s*ee Notes 30 and 31):

```
$ tandem.exe tandem_params_A_B_1.xml; tandem.
exe tandem_params_A_B_2.xml; tandem.exe tan-
dem_params_A_B_3.xml;tandem.exetandem_params_
A_B_4.xml; tandem.exe tandem_params_A_B_5.xml
```

3.1.2.3. Running XPRESS or ASAPratio from Xinteract

1. Convert X!Tandem output into the TPP input format (see Note 32):

```
$ Tandem2XML yeast_A_B_1.xtan.xml yeast_A_B_1_
raw.pep.xml
$ Tandem2XML
yeast_A_B_2.xtan.xml yeast_A_B_2_raw.pep.xml
$ Tandem2XML yeast_A_B_3.xtan.xml
yeast_A_B_3_raw.pep.xml
$ Tandem2XML yeast_A_B_4.xtan.xml
yeast_A_B_4_raw.pep.xml
$ Tandem2XML yeast_A_B_5.xtan.xml yeast_A_B_5_
raw.pep.xml
```

2. As in the simple case of the first workflow, Peptide- and Protein-Prophet will be run. Additionally, xinteract will be instructed to execute components responsible for ICAT quantification. The TPP ships with two such programs: XPRESS (started with the command line option –X) and ASAPRatio (started with the command line option –A). The programs are independent from each other and it is up to the user to decide which one to use. In fact, xinteract can also be instructed to run them both in parallel (see Note 33):

```
$ xinteract –Op –X –A -Nyeast_A_B.pep.xml
*raw.pep.xml
```

At the end of this second workflow the probabilities, as well as the estimated sensitivity and error rates, for each identified peptide will have been calculated. These will, in turn, have been used to derive the list of proteins, with their corresponding probabilities, most likely to have been present in the original sample. Finally, relative protein abundances in samples yeast_A and yeast_B will have been calculated using two independent approaches.

3.1.3. Workflow 3: Protein Identification and Quantification Based on MS2 Reporter Ions

In this third and last workflow, it is assumed that the following proteomic experiment was performed:

- Grow, harvest and lyse four yeast cultures under four different growth conditions. Hereafter referred to as yeast_A, yeast_B , yeast_C and yeast_D.
- Digest the labeled samples with trypsin.
- Label the peptides with iTRAQ reagents.
- Combine the four samples.
- Analyze the resulting sample (we will call it yeast_iTRAQ) using μLC-ESI-MS-MS/MS.

This strategy will be used to answer the question: "How are protein expression levels affected by the four different growth conditions?". The list of proteins most likely to have been present in the samples with their associated probabilities; estimated sensitivity and error rates; and relative abundances will be derived.

3.1.3.1. Converting the Files and Preparing the Analysis Directory

1. As for the first workflow, start by converting the native output from the MS instrument into mzXML as described in Subheading "Converting the RAW output of MS instruments to an open representation".
2. Create an analysis directory and copy all mzXML files in it:

```
$ mkdir /mnt/Data/MMB/Yeast_iTRAQ/
```

3.1.3.2. Adjusting X!Tandem Input Parameters Files

1. Copy tandem_params.xml from the TPP_ROOT/bin directory into the analysis directory (make sure all instances of "**_DEFAULT_INPUT_LOCATION_**" have been substituted as described in Subheading "Protein database search."):

```
$ cp TPP_ROOT/bin/tandem_params.xml /mnt/Data
/MMB/Yeast_iTRAQ/
```

2. Edit tandem_params.xml to match the desired search criteria.

   Specify the identifier for the protein sequence database to be searched:

```
<note type="input" label="protein, taxon">
yeast_nci_20070223</note>
```

3. Let the search engine know that lysine residues, as well as the N-terminus of each peptide, are modified (see Note 34):

```
<note type="input" label="residue, modifica-
tion mass">144.1@[,144.1@K</note>
```

4. Finally, specify which mzXML file to search and where to store the results.

```
<note type="input" label="spectrum, path">
yeast_iTRAQ.mzXML</note>
<note type="input" label="output, path">
yeast_iTRAQ.xtan.xml</note>
```

5. Save and close this file.
6. Start the search:

```
$ tandem.exe tandem_params.xml
```

3.1.3.3. Running Libra from Xinteract

1. Convert the X!Tandem output into the TPP pepXML format:

```
$ Tandem2XML yeast_iTRAQ.xtan.xml yeast_iTRAQ_raw.pep.xml
```

2. The TPP component responsible for iTRAQ sample quantification is called libra. This program accepts user input in the form of a properly formatted xml file, which can be created using the web-interface at: "http://db.systemsbiology.net/webapps/conditionFileApp/" and saved in the analysis directory as condition.xml.

As in the simple case of the first workflow, Peptide- and Protein-Prophet will be run. Additionally, xinteract will be instructed to execute Libra by adding the –L<condition_file> command-line argument:

```
$ xinteract -Op -Lcondition.xml -Nyeast_iTRAQ.pep.xml yeast_iTRAQ_raw.pep.xml
```

At the end of this third workflow, the probabilities, as well as the estimated sensitivity and error rates, for each identified peptide will have been calculated. These will, in turn, have been used to derive the list of proteins, with their corresponding probabilities, most likely to have been present in the original sample. Finally, the relative abundances of proteins in samples yeast_A, yeast_B, yeast_C, and yeast_D will have been calculated using isobaric report ions.

3.2. Examining and Validating the TPP Output

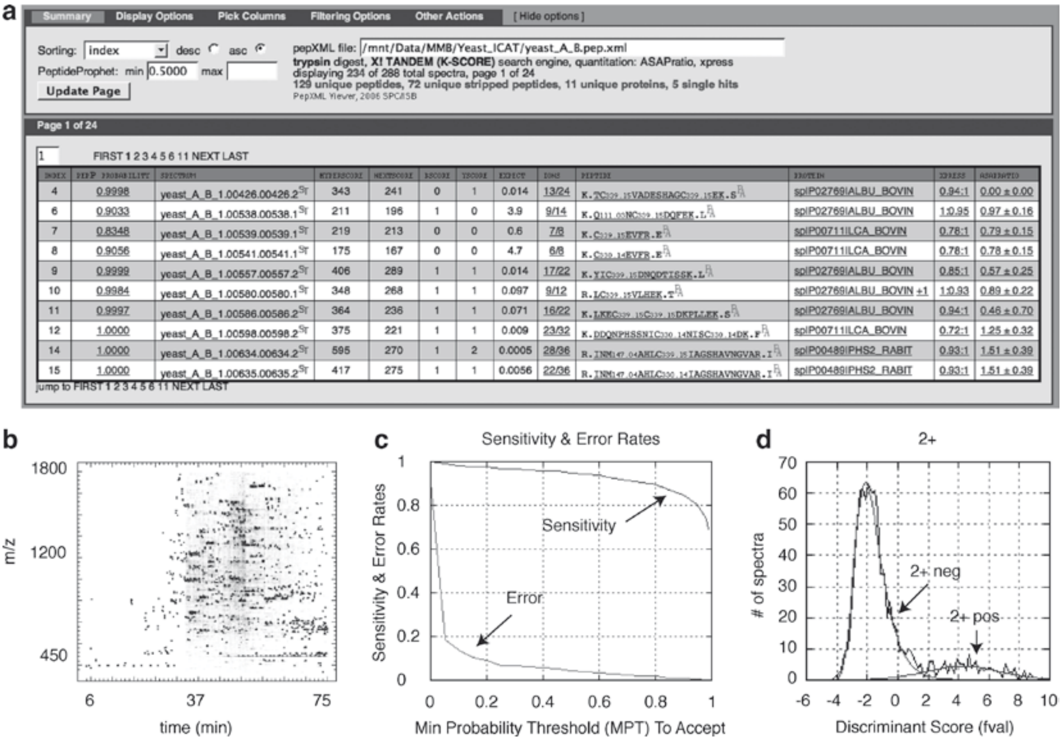
3.2.1. Peptide Level: PepXML Viewer

1. For reasons that go beyond the scope of this chapter, assignments from database search engines need to be validated to filter out incorrect ones. PeptideProphet (9) automates this task by using an Expectation-Maximization algorithm to learn the distribution of a discriminant function F that best separates negative and positive assignments for each precursor ion charge state. The probability of an assignment being correct given an observed score F_{val} is the ratio between the number of correct search results with score F_{val} and the total number of search results with score F_{val}. Our second workflow can be used as a practical example.
2. Point your web-browser to the analysis directory and click on the file with the extension .pep.shtml to open the PepXML Viewer. Each row in the table that appeared represents an MS/MS assignment and has the following columns (see Fig. 2a):

Index	Unique row-number.
PepP probability	PeptideProphet probability for this assignment. Click for a detailed report of PeptideProphet performances.
Spectrum	Has the following format file_name.start_scan.end_scan.charge_state. For instance, "yeast_A_B_2.00058.00058.2" would indicate an assignment from scan 58 in the mzXML file yeast_A_B_2.mzXML based on a precursor ion charge state of +2 (see Note 35).
Hyperscore; NextScore; BScore; YScore; and Expect	X!Tandem specific scores.
Ions	Fraction of theoretical fragment ions present in the MS/MS scan. Click to show the fragmentation spectra, the sequence of the assigned peptide and the difference between the theoretical and measured precursor masses.
Peptide	Sequence of the peptide assignment. For instance, K.C339.15EVFR.E would indicate that: (1) the spectra was assigned to the peptide CEVFR; (2) the cysteine residue was modified by heavy cleavable ICAT reagent; (3) the amino-acid preceding the first one from this peptide is a K; and (4) the amino-acid right after the end of this peptide is an E (see Note 36). Click to BLAST this peptide using the NCBI BLAST interface.
Protein	The protein(s) in the sequence database containing the assigned peptide. Click for a short description of the protein, and the location of the peptide in the protein sequence.
XPRESS	Only present if XPRESS was run. It shows the light to heavy ratio calculated for isotopically labeled peptides. Click to adjust the integration ranges used to calculate the area of the elution peaks.
ASAPRatio	Only present if ASAPRatio was run. It shows the light to heavy ratio calculated for isotopically labeled peptides with associated errors. Click to adjust the integration ranges used to calculate the area of the elution peaks.
Libra	Only present if Libra was run. There will be one such column for each of the iTRAQ reagents used, showing the intensity of the reporter ion for that particular iTRAQ reagent.

3. At the top of the PepXML Viewer page there are 5 tabs called "Summary" "Display Options" "Pick Columns" "Filtering Options" and "Other Actions", by clicking on any one of them different options to interact with the data displayed in the table will become available.

4. Quality of the chromatographic run and data acquisition should be checked first. Start Pep3D from the "Other Actions" tab by clicking on "Generate Pep3D". Accept the default settings and click on the "Generate Pep3D image" button. Shortly, the MS run will appear as a two-dimensional density plot (see Fig. 2b) with the retention time plotted on the x-axis, the m/z ratios on the y-axis, and the



Sorting: index desc ○ asc ●
PeptideProphet: min 0.5000 max

pepXML file: /mnt/Data/MMB/Yeast_ICAT/yeast_A_B.pep.xml
trypsin digest, X! TANDEM (K-SCORE) search engine, quantitation: ASAPratio, xpress
displaying 234 of 288 total spectra, page 1 of 24
129 unique peptides, 72 unique stripped peptides, 11 unique proteins, 5 single hits
PepXML Viewer, 2006 SPC/ISB

1 FIRST 1 2 3 4 5 6 11 NEXT LAST

INDEX	PEP PROBABILITY	SPECTRUM	HYPERSCORE	NEXTSCORE	BSCORE	YSCORE	EXPECT	IONS	PEPTIDE	PROTEIN	XPRESS	ASAPRATIO
4	0.9998	yeast_A_B_1.00426.00426.2	343	241	0	1	0.014	13/24	K.TC$_{339.15}$VADESHAGC$_{339.15}$EK.S	sp|P02769|ALBU_BOVIN	0.94:1	0.00 ± 0.00
6	0.9033	yeast_A_B_1.00538.00538.1	211	196	1	0	3.9	9/14	K.Q$_{111.03}$NC$_{339.15}$DQFEK.L	sp|P02769|ALBU_BOVIN	1:0.95	0.97 ± 0.16
7	0.8348	yeast_A_B_1.00539.00539.1	219	213	0	0	0.6	7/8	K.C$_{339.15}$EVFR.E	sp|P00711|LCA_BOVIN	0.78:1	0.79 ± 0.15
8	0.9056	yeast_A_B_1.00541.00541.1	175	167	0	0	4.7	6/8	K.C$_{330.14}$EVFR.E	sp|P00711|LCA_BOVIN	0.78:1	0.78 ± 0.15
9	0.9999	yeast_A_B_1.00557.00557.2	406	289	1	1	0.014	17/22	K.YIC$_{339.15}$DNQDTISSK.L	sp|P02769|ALBU_BOVIN	0.85:1	0.57 ± 0.25
10	0.9984	yeast_A_B_1.00580.00580.1	348	268	1	1	0.097	9/12	R.LC$_{339.15}$VLHEK.T	sp|P02769|ALBU_BOVIN +1	1:0.93	0.89 ± 0.22
11	0.9997	yeast_A_B_1.00586.00586.2	364	236	1	1	0.071	16/22	K.LKEC$_{339.15}$C$_{339.15}$DKPLLEK.S	sp|P02769|ALBU_BOVIN	0.94:1	0.46 ± 0.70
12	1.0000	yeast_A_B_1.00598.00598.2	375	221	1	1	0.009	23/32	K.DDQNPHSSNIC$_{330.14}$NISC$_{330.14}$DK.F	sp|P00711|LCA_BOVIN	0.72:1	1.25 ± 0.32
14	1.0000	yeast_A_B_1.00634.00634.2	595	270	1	2	0.0005	28/36	R.INM$_{147.04}$AHLC$_{339.15}$IAGSHAVNGVAR.I	sp|P00489|PHS2_RABIT	0.93:1	1.51 ± 0.39
15	1.0000	yeast_A_B_1.00635.00635.2	417	275	1	1	0.0056	22/36	R.INM$_{147.04}$AHLC$_{330.14}$IAGSHAVNGVAR.I	sp|P00489|PHS2_RABIT	0.93:1	1.51 ± 0.39

jump to FIRST 1 2 3 4 5 6 11 NEXT LAST

Fig. 2. (**a**) PepXML Viewer showing the peptide identifications from an ICAT experiment. (**b**) Pep3D representation of an MS analysis as a two-dimensional density plot. (**c**) PeptideProphet sensitivity and error rates plot. (**d**) Actual and learned distributions of the PeptideProphet discriminant score *F* for doubly charged precursor ions

peak intensities represented in shades of grey. Overlaid on this representation, the CID attempts will be plotted as squares and the probabilities assigned by PeptideProphet will be shown using a green to red color gradient (see Note 37). This type of representation can prove very powerful in troubleshooting a suboptimal run. A good quality LC-MS-MS/MS run will show evenly distributed, sharp, signals and CID attempts (13).

5. Next, check how well PeptideProphet learned the score distributions. From the main PepXML Viewer window, click on any link from the "Probability" column. A new page will open with a "Sensitivity & Error Rates" plot (see Fig. 2c), which can be used to select a PeptideProphet Minimum Probability Threshold to accept (MPT) value. Sensitivity represents the fraction of all correct assignments that will pass a given MPT filter. Error represents the fraction of incorrect peptide assignments that will pass a given MPT filter. For instance, in Fig. 2c, an MPT of 0.50 would give an estimated sensitivity of 95% and an error of 4%. The shape of these curves depends on how much discriminating power the model learned by

PeptideProphet has to separate correct and incorrect peptide assignments. The closer the 2 curves get to the y = 1 - x diagonal, the less discriminating power the model has. In the section called "Model results", the actual and learned distributions of the discriminant function *F* are displayed for each precursor ion charge state (see Fig. 2d). A good model will have a gamma distribution for the incorrect assignments and a normal distribution for the correct assignments, with the positive one being smaller and to the right of the negative one (see Note 38).

6. In the main PepXML Viewer view, enter the MPT in the "PeptideProphet min" input field of the "Summary" tab. Select "Sorting by descending probability" from the pull-down menu and click on "Update Page". Unique peptides and proteins numbers shown in the "Summary" tab will change to reflect the new filtering criteria.
7. Interesting peptide assignments (see Note 39) can be further validated by examining the MS/MS spectra that opens when clicking on the "Ions" hyperlink. Good assignments will match most of the high intensity ions and show a continuous series of y and b ions. If the sample was digested with trypsin, the y-ion series will typically be the most intense.
8. Finally, ICAT labeling and enrichment efficiency can be assessed. Open the "Display Options" tab, enter "C" in the "highlight peptide text (regex)" field and click on "Update Page". Cysteine residues will be highlighted, thus providing an overview on how well the enrichment for cysteine containing peptides worked (see Note 40).

After completing these initial steps, one should have a good idea on how well the sample preparation, sample separation, MS data acquisition, and bioinformatic data analysis performed. If everything went well, this is a good time to close PepXML Viewer and move on to the protein level.

3.2.2. Protein Level: ProtXML Viewer

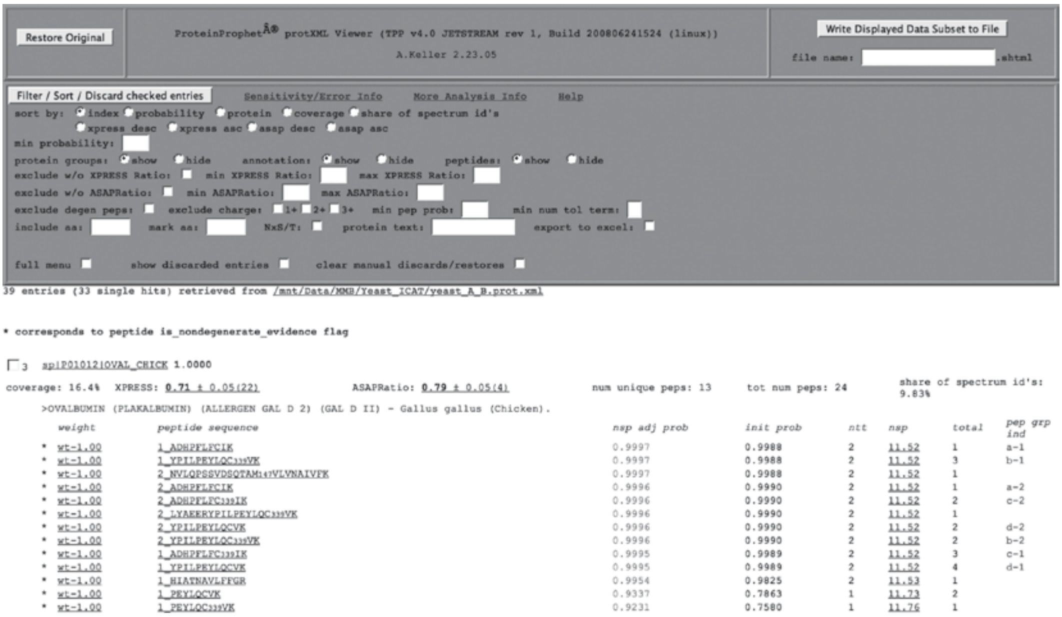
1. In bottom-up proteomics, the connectivity information between peptides is lost, making the transition from the peptide to the protein level non-trivial. ProteinProphet groups peptides while trying to generate the list of proteins most likely to have composed the original sample. The probability of each protein is calculated based on the probabilities of the individual peptides adjusted to account for the fact that groups of peptides generated from the same protein are more likely to be correct. In ProteinProphet, this concept is described with the notion of the Number of Sibling Peptides (NSP). The NSP of a given peptide is calculated as the sum of PeptideProphet probabilities for all other peptides assigned

to the same protein. Probabilities are then increased for peptides with high NSP and lowered for peptides with low NSP. These adjusted peptide probabilities are finally used to compute the probability of their corresponding proteins. When a peptide can be associated with more than one protein, the model tries to generate the shortest possible list of proteins that could explain all the peptides seen in the analysis (Occam's razor). Thus, the contribution each peptide gives to the final protein probability is weighted proportionally to the number of proteins that contain that specific peptide.

2. Point your web-browser to the analysis directory and open the file with extension .prot.shtml to open the ProtXML Viewer. By clicking on the link called "Sensitivity/Error Info" in the top part of the page, a plot of the sensitivity and error rates versus MPT will appear. This is the protein equivalent of the PeptideProphet "Sensitivity & Error Rates" plot. Select an MPT that gives acceptable sensitivity and error rates.

3. Enter the protein MPT in the field called "min probability" of the main ProtXML Viewer page and click on "Filter/Sort/ Discard checked entries". The red summary line indicating the number of proteins being displayed at the bottom of the filtering panel will change to reflect the new filter.

4. ProtXML Viewer displays identified proteins (see Fig. 3) in the following format:


Restore Original
Write Displayed Data Subset to File
Filter / Sort / Discard checked entries
Sensitivity/Error Info
More Analysis Info
Help

ProteinProphetÂ® protXML Viewer (TPP v4.0 JETSTREAM rev 1, Build 200806241524 (linux))

A.Keller 2.23.05

file name: ____ .shtml

```
sort by: ● index ○ probability ○ protein ○ coverage ○ share of spectrum id's
         ○ xpress desc ○ xpress asc ○ asap desc ○ asap asc
min probability: 
protein groups: ● show ○ hide    annotation: ● show ○ hide    peptides: ● show ○ hide
exclude w/o XPRESS Ratio: □  min XPRESS Ratio:     max XPRESS Ratio: 
exclude w/o ASAPRatio: □  min ASAPRatio:     max ASAPRatio: 
exclude degen peps: □  exclude charge: □1+ □2+ □3+   min pep prob:      min num tol term: 
include aa:      mark aa:      NxS/T: □   protein text:           export to excel: □

full menu □      show discarded entries □     clear manual discards/restores □
```

39 entries (33 single hits) retrieved from /mnt/Data/MMB/Yeast_ICAT/yeast_A_B.prot.xml

* corresponds to peptide is_nondegenerate_evidence flag

□ 3 sp|P01012|OVAL_CHICK 1.0000

coverage: 16.4% XPRESS: 0.71 ± 0.05(22) ASAPRatio: 0.79 ± 0.05(4) num unique peps: 13 tot num peps: 24 share of spectrum id's: 9.83%

>OVALBUMIN (PLAKALBUMIN) (ALLERGEN GAL D 2) (GAL D II) - Gallus gallus (Chicken).

	weight	peptide sequence	nsp adj prob	init prob	ntt	nsp	total	pep grp ind
*	wt-1.00	1_ADHPFLFCIK	0.9997	0.9988	2	11.52	1	a-1
*	wt-1.00	1_YPILPEYLQC$_{339}$VK	0.9997	0.9988	2	11.52	3	b-1
*	wt-1.00	2_NVLQPSSVDSQTAM$_{147}$VLVNAIVFK	0.9997	0.9988	2	11.52	1	
*	wt-1.00	2_ADHPFLFCIK	0.9996	0.9990	2	11.52	1	a-2
*	wt-1.00	2_ADHPFLFC$_{339}$IK	0.9996	0.9990	2	11.52	2	c-2
*	wt-1.00	2_LYAEERYPILPEYLQC$_{339}$VK	0.9996	0.9990	2	11.52	1	
*	wt-1.00	2_YPILPEYLQCVK	0.9996	0.9990	2	11.52	2	d-2
*	wt-1.00	2_YPILPEYLQC$_{339}$VK	0.9996	0.9990	2	11.52	2	b-2
*	wt-1.00	1_ADHPFLFC$_{339}$IK	0.9995	0.9989	2	11.52	3	c-1
*	wt-1.00	1_YPILPEYLQCVK	0.9995	0.9989	2	11.52	4	d-1
*	wt-1.00	1_HIATNAVLFFGR	0.9954	0.9825	2	11.53	1	
*	wt-1.00	1_PEYLQCVK	0.9337	0.7863	1	11.73	2	
*	wt-1.00	1_PEYLQC$_{339}$VK	0.9231	0.7580	1	11.76	1	

Fig. 3. Screen-capture of ProtXML Viewer showing the protein identifications from an ICAT experiment

(a) Each entry starts with an index number, a protein name, and its associated protein probability (see Note 41).

(b) The next line indicates the protein coverage (i.e., what % of this protein sequence was seen in the MS), the protein ratios calculated by XPRESS; ASAPRatio; and Libra, the number of unique as well as the total number of peptides assigned to this protein, and the share of spectrum id's (i.e., what % of spectrum identifications from this analysis is accounted for by this protein).

(c) The next line shows a brief description of the protein.

(d) The next table displays information specific to the peptides assigned to this protein:

Weight	How much weight has been given to a peptide when calculating the probability of a protein (i.e., to account for how many different proteins contain the same peptide). Some weights are preceded by a * to indicate that no other sequence database entry shares this particular peptide.
Peptide sequence	Charge state and sequence of the peptide. Click to open a PepXML Viewer like representation.
nsp adj prob	NSP adjusted PeptideProphet probability.
init prob	PeptideProphet probability before NSP adjustment.
ntt	Number of tryptic termini.
nsp	NSP value.
Total	Times this peptide, with this charge state, has been sequenced.
pep grp ind	Connections between peptides with the same sequence (represented by the letter before the "-" sign) and different charge states (represented by the number after the "-" sign).

5. Proteins isotopic ratios (see Note 42), can be fine-tuned by clicking on the XPRESS ratio to adjust the integration ranges used to calculate the area of the elution peaks for the corresponding peptide pairs. Alternatively, the ASAPRatio link can be used to adjust elution peaks integration intervals, and to accept or reject individual peptide quantifications.

6. The protein list can be saved by clicking on "Write displayed data subset to file" after selecting a file name, or exported in a tab separated format suitable for Excel and SBEAMS (http://www.sbeams.org) by selecting "export to excel" and clicking on the "Filter/Sort/Discard checked entries" button.

After completing this section, one should have generated a validated list of protein assignments that can be further analyzed to address the original experimental question.

4. Notes

1. While some third party software packages, such as the CPAS LIMS from LabKey Software, include components of the TPP, in this chapter, we discuss only the original SPC version of the TPP.
2. An MS-DOS window can be opened from the Windows Start menu by clicking on Run and typing cmd.
3. Except when starting a program from the same directory where the executable is found, one must specify the full path to it (e.g., C:\mzXML\Converters\ReAdW.exe). Alternatively, the directory with the converters can be added to the DOS PATH variable.
4. The user running the web-server (e.g. www-data) needs write permissions in the analysis directory. To achieve this, create a group that contains the physical user (e.g., mmb) and the web-server user, give ownership of the analysis directory to this group, and grant it read and write permissions:

```
$ sudo groupadd analysis
$ sudo usermod -a -G analysis mmb
$ sudo usermod -a -G analysis www-data
$ sudo chown :analysis /mnt/Data/MMB/
Yeast_IP/
$ sudo chmod g+w /mnt/Data/MMB/Yeast_IP/
```

5. In Apache2, to enable SSI, add to your site configuration (e.g., /etc/apache2/sites-available/default):

```
Options +IncludesAddType text/html .shtm-
lAddHandler server-parsed .shtml
```

   in the `<Directory /var/www/>` section.
   Save and restart the web-server.
6. More complex approaches can better discriminate specific and unspecific interactors (15).
7. Some useful tools for the mzXML format are:

   readmzXML: prints the peaks in any given scan (included in the TPP).

   MzXML2Search: extracts MS/MS scans from an mzXML file and saves them in formats compatible with database search engines (included in the TPP).

validateXML: validates the integrity of an mzXML file (http://tools.proteomecenter.org/validateXML.php).

mzXMLViewer and Insilicos Viewer: visualize the scans in an mzXML file (http://tools.proteomecenter.org/mzXMLViewer.php and http://www.insilicos.com/free_downloads.html).

8. This program depends on the XRawfile library from ThermoFinnigan and will only work on a computer with Xcalibur installed.
9. For a list of options execute the command without any argument.
10. This program depends on the DAC library from Waters and will only work on a computer with MassLynx installed.
11. This program depends on a library from Analyst and will only work on a computer with Analyst installed.
12. This program is developed and maintained by Bruker. Further information can be found at:
    http://www.ionsource.com/functional_reviews/CompassXport/CompassXport.htm
    and at
    http://www.bioinformaticssolutions.com/products/peaks/support/bruker.php.
13. More options can be seen by typing:

    ```
    $ dta2mzxml -help
    ```

    mzXML files created using this converter contain only MS2 scans and are not be suitable for applications requiring MS1 information.
14. X!Tandem is developed and maintained by the Global Proteome Machine Organization ("www.thegpm.org").
15. The names of these files can change as long as the changes are reflected in the content of input.xml (i.e., the elements `<note type="input" label="list path, taxonomy information">`and`<note type="input" label="list path, default parameters">` need to be appropriately updated) and in the way X!Tandem is called (i.e., the first parameter passed to tandem.exe must be the name of the file corresponding to input.xml).
16. Syntax of these files is documented at: http://www.thegpm.org/TANDEM/api/index.html
17. A static modification is applied to any instance of a particular amino acid (e.g., carboxamidomethylation of all cysteines).

    A variable (also called potential) modification is considered, but not necessarily applied, for any instance of a particular amino acid (e.g., phosphorylation).

18. Existing users of X!Tandem trying to integrate the TPP into their workflow, should use this file as the default_input.xml rather than the one originally distributed with X!Tandem.

19. Any document edited throughout the course of this chapter should be saved as plain text.

20. More protein sequence databases can be added by appending these lines to the taxonomy.xml file:

    ```
    <taxon label="database_identifier_A"><file format="peptide" URL="full_path_to_database_A"/>
    ```

21. Users that have compiled X!Tandem with the native score should edit this file to use isb_default_input_native.xml instead of isb_default_input_kscore.xml:

    ```
    <note type="input" label="list path, default parameters">TPP_ROOT/ /bin/isb_default_input_native.xml</note>
    ```

22. Multiple modifications can be separated with a comma.

23. Enzymatic cleavage parameters can be changed in tandem_params.xml:

    ```
    <note type="input" label="protein, cleavage semi">yes</note>
    <note type="input" label="scoring, maximum missed cleavage sites">2</note>
    ```

    Since the default enzyme defined in "isb_default_input_kscore.xml" is trypsin, these lines instruct the search engine to consider semi-tryptic peptides (i.e., peptides with a non-tryptic cleavage at either one of their termini), and to allow for up to 2 missed cleavages per peptide (i.e., K or R residues in the middle of the peptide sequence other than those followed by a P) respectively.

24. X!Tandem searches on remote machines should be started with nohup to prevent them from being terminated when the user logs out:

    ```
    $ nohup tandem.exe tandem_params.exe
    ```

25. After copying tandem-style.xsl from trans_proteomic_pipeline/extern/tandem-linux-07-07-01-2/bin/ into the analysis directory one can visualize the content of yeast_ip.xtan.xml in a web-browser.

26. The TPP contains equivalent programs to convert the output from other search engines to pepXML (i.e. Out2XML for SEQUEST; Comet2XML for COMET; Mascot2XML for MASCOT).

27. A list of all xinteract command line arguments can be seen by typing:

```
$ xinteract
```

28. To search samples with Lysine and Arginine residues labeled using the SILAC protocol, the following line can be used instead:

```
<note type="input" label="residue, potential
modification mass">8.0@K,10.0@R</note>
```

29. These steps can be automated by creating a file called "prepare_tandem_params.sh" in the TPP_ROOT/bin/ directory with the following content:

```
#!/bin/bash
for ARG in "$@"
do
base_name=${ARG%.mzXML}
out_name=${ARG/.mzXML/.xtan.xml}
echo Preparing parameters for $base_name
sed s/full_mzXML_filepath/$ARG/ \
tandem_params.xml | \
sed s/full_tandem_output_path/$out_name/ > \
tandem_params_$base_name.xml
done
rm tandem_params.xml
```

Making it executable with:

```
$ chmod 755 prepare_tandem_params.sh
```

And executing it in the analysis directory that contains all the mzXML files and a version of tandem_params.xml with all the search parameters set correctly, except for the input and output files, as follows:

```
$ prepare_tandem_params.sh *.mzXML
```

Petunia, the GUI interface for the TPP, will automatically take care of this tedious task for you.

30. To shorten this command, create a file called "start_tandem.sh" in the TPP_ROOT/bin/ directory with the following content:

```
#!/bin/bash
find ./ -name "tandem_params*" | \
xargs -I {} tandem.exe {}
```

Make it executable with:

```
$ chmod 755 start_tandem.sh
```

In the analysis directory type:

```
$ start_tandem.sh
```

31. Commands separated by semicolons will be executed sequentially.
32. To simplify this step, create a file called “convert_tandem2xml.sh” in the TPP_ROOT/bin/ directory with the following content:

```
#!/bin/bash
find ./ -name "*.xtan.xml" | \
sed s/.xtan.xml// | \
xargs -t -I {} Tandem2XML \
{}.xtan.xml {}_raw.pep.xml
```

Make it executable with:

```
$ chmod 755 convert_tandem2xml.sh
```

In the analysis directory, type:

```
$ convert_tandem2xml.sh
```

33. In XPRESS, quantification reagents other than ICAT can be specified with the -n<residue>,<delta_mass> command line argument. For instance, for a SILAC experiment using Lysine and Arginines, xinteract should be run like this:

```
$ xinteract -Op -X-nK,8.0-nR,10.0 silac_
analysis_raw.pep.xml
```

34. X!Tandem uses “[“and”]” to indicate the N- and C- terminus, respectively.
35. If start_scan and end_scan differ, then the current peptide assignment was done by integrating the signal in all scans from start_scan to end_scan.
36. The C- and N-terminus of the protein are indicated with a “-” symbol.
37. Often only a small percentage of MS/MS spectra leads to successful peptide assignments (i.e., typically 10-20%). This is partially due to factors that would be too time prohibitive to account for in the initial search. The TPP program QualScore extracts high quality unassigned spectra from an MS run. To use it, run a normal database search, convert the results to pepXML and run xinteract with parameter -p0 to keep peptides with probability lower than 0.05. Finally, start QualScore as follows:

```
$ java -jar qualscore.jar <pep.XML_file_
to_be_analyzed>
```

Substitute “<pep.XML_file_to_be_analyzed>” for the name of the xinteract generated pep.xml file. This will create a new directory containing all the MS/MS scans, in SEQUEST .dta format, deemed to be of high quality by QualScore. This subset of fragmentation spectra can now be converted to an mzXML file as explained in Subheading “Converting the

RAW output of MS instruments to an open representation" and a more exhaustive database search can be run.

38. If PeptideProphet is unable to learn the score distributions for a particular charge state, it assigns a negative probability. These assignments might be good ones, but they require manual validation.
39. PepXML Viewer only displays the highest scoring identification for each queried spectrum. However, sometimes the correct assignment is a lower scoring peptide. When using SEQUEST, click on the "Spectrum" link in the PepXML Viewer to see lower scoring peptides.
40. The actual number of unique peptides containing a cysteine residue can be seen in the "Summary" tab, after filtering the results for cysteine containing peptides using the "required peptide text" field from the "Filtering Options" tab.
41. Multiple proteins that could account for the same peptide list are grouped in one entry.
42. Rather than looking at all proteins, the results can also be filtered for minimal and maximal XPRESS and/or ASAPRatio ratios using the "min/max XPRESS Ratio" and "min/max ASAPRatio" fields.

Acknowledgments

We thank Jimmy Eng, Deena Leslie Pedrioli, and James Eddes for critical reading of the manuscript. We also thank Matthias Peter and the Roche Research Foundation for providing support for this work.

References

1. Yates, J. R., III, Eng, J. K., McCormack, A. L., and Schieltz, D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **67,** 1426–36.
2. 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,** 3551–67.
3. Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3,** 1454–63.
4. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **1,** 2005 0017.
5. 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,** 1406–12.
6. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–7.
7. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K.,

Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **22,** 1459–66.

8. Orchard, S., Montechi-Palazzi, L., Deutsch, E. W., Binz, P. A., Jones, A. R., Paton, N., Pizarro, A., Creasy, D. M., Wojcik, J., and Hermjakob, H. (2007) Five years of progress in the Standardization of Proteomics Data 4th Annual Spring Workshop of the HUPO-Proteomics Standards Initiative April 23–25, 2007 Ecole Nationale Superieure (ENS), Lyon, France. *Proteomics* **7,** 3436–40.
9. 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,** 5383–92.
10. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75,** 4646–58.
11. 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,** 946–51.
12. 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,** 6648–57.
13. Li, X. J., Pedrioli, P. G., Eng, J., Martin, D., Yi, E. C., Lee, H., and Aebersold, R. (2004) A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry. *Anal Chem* **76,** 3856–60.
14. MacLean, B., Eng, J. K., Beavis, R. C., and McIntosh, M. (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **22,** 2830–2.
15. Ranish, J. A., Yi, E. C., Leslie, D. M., Purvine, S. O., Goodlett, D. R., Eng, J., and Aebersold, R. (2003) The study of macromolecular complexes by quantitative proteomics. *Nat Genet* **33,** 349–55.

Chapter 16

Informatics and Statistics for Analyzing 2-D Gel Electrophoresis Images

Andrew W. Dowsey, Jeffrey S. Morris, Howard B. Gutstein, and Guang-Zhong Yang

Abstract

Despite recent progress in "shotgun" peptide separation by integrated liquid chromatography and mass spectrometry (LC/MS), proteome coverage and reproducibility are still limited with this approach and obtaining enough replicate runs for biomarker discovery is a challenge. For these reasons, recent research demonstrates that there is a continuing need for protein separation by two-dimensional gel electrophoresis (2-DE). However, with traditional 2-DE informatics, the digitized images are reduced to symbolic data through spot detection and quantification before proteins are compared for differential expression by spot matching. Recently, a more robust and automated paradigm has emerged where gels are directly aligned in the image domain before spots are detected across the whole image set as a whole. In this chapter, we describe the methodology for both approaches and discuss the pitfalls present when reasoning statistically about the differential protein expression discovered.

Key words: 2-D gel electrophoresis, Image alignment, Spot detection, Spot matching, Differential expression analysis, Clustering, DIGE

1. Introduction

Since its beginnings in 1975 (1, 2), two-dimensional gel electrophoresis (2-DE) has established itself as the principal approach for separating proteins from cell and tissue samples (3). While recent progress in "shotgun" peptide separation with liquid chromatography and mass spectrometry (LC/MS) (4, 5) has brought some significant analytical benefits, recent bench comparisons have shown that proteome coverage is complementary to 2-DE rather than encompassing (6). Furthermore, currently there are issues with the reproducibility of LC/MS that are difficult to correct retrospectively by alignment, plus there are practical issues limiting the number of

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_16, © Humana Press, a part of Springer Science+Business Media, LLC 2010

replicate runs that can be made and therefore experimental power for biomarker discovery. For these reasons, protein modelling, quantification, and differential expression analysis with 2-DE continues to be an important workhorse method for proteomics research.

The first step in proteomic informatics analysis is image acquisition, either of gels or mass spectra (in LC/MS). The traditional 2-DE informatics pipeline then attempts to identify spot boundaries and quantify individual spots on each gel before proteins are compared for differential expression by matching cognate spots between gels. With existing commercial software, errors in each step contribute to a highly subjective and labour-intensive correction. For example, it has been noted that increasing the number of gels in an experiment dramatically reduces the percentage of correct automated spot matches (7). Recently, a more robust and automated concept has emerged (8), where gels are directly aligned in the image domain (9, 10) so that subsequent spot detection can be based on the integration of the spot appearances in every gel. It has been shown that through preservation of the raw image information contained in each spot and its statistical "fusion" over the gel set, increased power and reliability in quantification is possible which further improves as the sample size is increased (11). Statistical rather than deterministic treatment is key to this new paradigm (12).

In this chapter, we describe the methodology for both approaches and discuss the pitfalls present when reasoning statistically about differential protein expression, with particular emphasis given to the need to perform power analyses and control the false discovery rate. In the remainder of this section, an overview of the established proteome informatics methods will be provided so that the choice of software detailed in Subheading 2 can be better understood. Subheading 3 then details step-by-step instructions for performing the analyses.

1.1. Proteome Informatics

There are a number of challenges in 2-DE proteome informatics (13). Despite the high resolution, diversity of cellular proteins often leads to spot co-migration. Some spots also tend to have severe tails in either dimension, confounding spot modelling. Contrast variations due to stain exposure, sample loading errors, and protein losses during processing inhibit the reliability of volume quantification. Furthermore, geometric distortions due to casting, polymerization, and the running procedure make the deduction of corresponding spots between gels demanding and therefore differential analysis challenging. The DIGE protocol (14) allows up to three samples to be run on the same gel with consequently little geometric discrepancy between them. However, typical experiments require a considerably greater sample size than a single DIGE gel to attain adequate power, and so inter-gel alignment is still a problematic issue. The typical steps in proteome informatics for 2-DE are (13):

1. *Image acquisition*: This prepares each raw acquisition for subsequent comparative analysis. After scanning, the images are pre-processed by cropping (manual delineation), noise suppression, and background subtraction (e.g., with mathematical morphology or smooth polynomial surface fitting).
2. *Conventional analysis (Spot Detection » Spot Matching)*: Each protein spot is delineated and its volume quantified. Typically, the spots are segmented first by the watershed transform (15), where spots are treated as depressions in a landscape which is slowly immersed in water. Spot boundaries (watersheds) are constructed where the pools start to meet. Co-migrating spots with separate peaks are then separated by parametric spot mixture modelling *e.g.*, optimizing the parameters of one or more 2-D Gaussians to minimize the squared residuals. Point pattern matching is then employed to match the spots between gels, which finds the closest spot correspondence between a point pattern (source spot list) and a target point set (reference spot list).
3. *Image-based analysis (Gel alignment » Consensus Spot Modelling)*: With current techniques, a "reference" gel is chosen and the other "source" gels are aligned to it in pair-wise fashion. For the new image-based paradigm, "direct image registration" is applied which defines a transformation that warps (deforms) the source gel and a similarity measure which quantifies the quality of alignment between the warped source gel and the reference gel. The aim is to automatically find the optimal transformation that maximizes the similarity measure. Spot detection is then performed on an image or "spot mask" created from the set, which is then propagated to each individual gel for spot quantification.
4. *Differential analysis*: At this stage, we have a list of spots, and for each spot, a quantified abundance in each gel. The abundances are first normalized to remove systemic biases between gels and between channels in DIGE gels. Variance stabilization can then be employed to remove the dependence between the mean abundance of a protein and its variance *e.g.*, a simple logarithmic transformation to fold-change values. Significance tests are then performed to obtain *p*-values for rejecting the null hypothesis that the mean spot abundance between groups is unregulated.
5. *Advanced techniques*: Since multiple hypothesis testing leads to a large number of false positives, it is essential to control the False Discovery Rate (FDR). The FDR is the estimated percentage of false positives within the detected differential expression rather than within the set of tests as a whole. Power analysis is also essential, which estimates the false negative rate that determines the optimal sample size needed to detect a specific fold change to a particular confidence level. Typical software packages do not contain these important methods.

6. *Diagnostics*: It is useful to look at various diagnostics to assess quality control of the gels in a given experiment, and to search for evidence of any artifacts that may indicate some problems in the gels. Hierarchical clustering can be used to assess which gels are most similar to each other, which can reveal experimental design or other quality control issues in the data. Further, one should visually assess any spots detected as differentially expressed to ensure that outliers unrelated to the biological groupings do not drive the result.

2. Materials

2.1. Image Acquisition

An image capture device is required, for which there are three main categories:

1. *Flatbed scanner*: This mechanically sweeps a standard charge-coupled device (CCD) under the gel and can be used to obtain 12–16 bits of greyscale or colour densitometry from visible light stains. Noise can be an issue due to size and cooling restrictions on the moving sensor and the need for reconstruction through image "stitching." Calibration is often required to provide linearity. Based on high-end document scanners but fully sealed, flatbed scanners are typically the least expensive offerings. Examples: ImageScanner (GE Healthcare, Chalfont St. Giles, UK), ProteomeScan (Syngene, Cambridge, UK) and GS-800 (Biorad, Hercules, CA).
2. *CCD camera*: Since the sensor is fixed, its greater size and cooling provides a dramatic improvement in noise and therefore dynamic range (up to 10^4). Different filters and transillumination options allow a wide range of stains to be imaged, including visible light, fluorescent, reverse, chemiluminescent, and radioactive signals. However, the fixed sensor limits image resolution, while vignetting (reduction of brightness at the periphery) and barrel distortion requires dark frame and flat frame correction respectively, affecting quantification. Examples: LAS (Fuji Photo Film, Tokyo, Japan), ImageQuant (GE Healthcare), Dyversity (Syngene), BioSpectrum2D (UVP, Upland, CA, USA) and VersaDoc (Biorad).
3. *Laser scanner*: Photomultiplier detectors are combined with laser light and optical or mechanical scanning to pass an excitation beam over each target pixel. While slower than CCD cameras, spatial resolution is excellent and logarithmic response leads to a dynamic range of up to 10^5. However, acquisition is limited to dyes whose excitation spectra match that of the installed laser sources, which are costly. With some products, visible light stains can be negatively imaged by

using a fluorescent back board. Examples: FLA (Fuji Photo Film), Typhoon (GE Healthcare) and PharosFX (Biorad).

Please see (16) for further details. Most specialized acquisition devices come with software to crop the resulting scans, but if this is unsuitable, the packages described in the next two sections have this facility, as does ProteomeGRID (http://www.proteomegrid.org/). See Note 1.

2.2. Conventional Analysis Software

A commercial software package is required, such as:

1. ImageMaster 2D or DeCyder (GE Healthcare, Chalfont St. Giles, UK)
2. Dymension (Syngene, Cambridge, UK)
3. Melanie (GeneBio, Geneva, Switzerland)
4. PDQuest (BioRad, Hercules, CA)
5. ProteinMine (BioImagene, Cupertino, CA)

These products are all quite expensive, so comparative personal evaluation is essential. As a guide, comparative assessments appear in the literature (7).

When choosing an acquisition device, it is important to ensure that the output format is compatible with the input format of the analysis software. While typically this involves standardized interchange with the TIFF format, few vendors adhere fully to the standard and therefore incompatibilities do occur. TIFF is also limited to 16 bits of linear dynamic range so some packages implement formats such as Fuji "IMG", which supports logarithmic image capture, and GE Healthcare "GEL", which supports square-root image capture. See Note 2.

2.3. SEA Image-Based Analysis Pipeline

Two commercial packages exist that adopt elements of the image-based analysis paradigm:

1. Delta2D (Decodon, Greifswald, Germany)
2. Progenesis SameSpots (Nonlinear Dynamics, Newcastle, UK)

Both packages perform image alignment before consensus spot detection. However, the alignment performed is only semi-automated with considerable user interaction, and the quantification is based on heuristic delineation of spot boundaries rather than more reliable peak detection (11). To utilize automated image-based alignment and fully harness strength borrowed from the whole gel set in the spot modelling phase, the following techniques can be combined:

1. RAIN (9, 10) (http://www.proteomegrid.org/) for automatic gel alignment.
2. Pinnacle (11) for automated spot detection and quantification that borrows strength between gels in determining what is a real spot.

2.4. Differential Analysis

The commercial packages described in Subheading 2.2 and 2.3 contain the standard tools for determining the statistical significance for regulation of an isolated protein between treatment groups. However, at the time of writing only Delta2D (Decodon) and the add-on Progensis Stats module (Nonlinear Dynamics) have facilities to correct for multiple hypothesis testing with FDR. If your software does not include FDR estimation or you wish to use freely available tools, one of the following microarray analysis suites can be used:

1. The R language and BioConductor repository (http://www.bioconductor.org/)
2. TM4 (http://www.tm4.org/)

The above suites also contain the more advanced normalization and power analyses described herein.

For data quality assessment and to investigate hidden factors in the data, the majority of commercial packages described in Subheading 2.2 and 2.3 contain basic data mining techniques. If a required technique is not available in your software, a range of advanced classification and data summarization methods can be found by using the microarray analysis suites above or with commercial solutions including:

1. Progenesis Stats (Nonlinear Dynamics)
2. Decyder EDA (GE Healthcare)
3. Genedata Expressionist (Genedata, Basel, Switzerland)

3. Methods

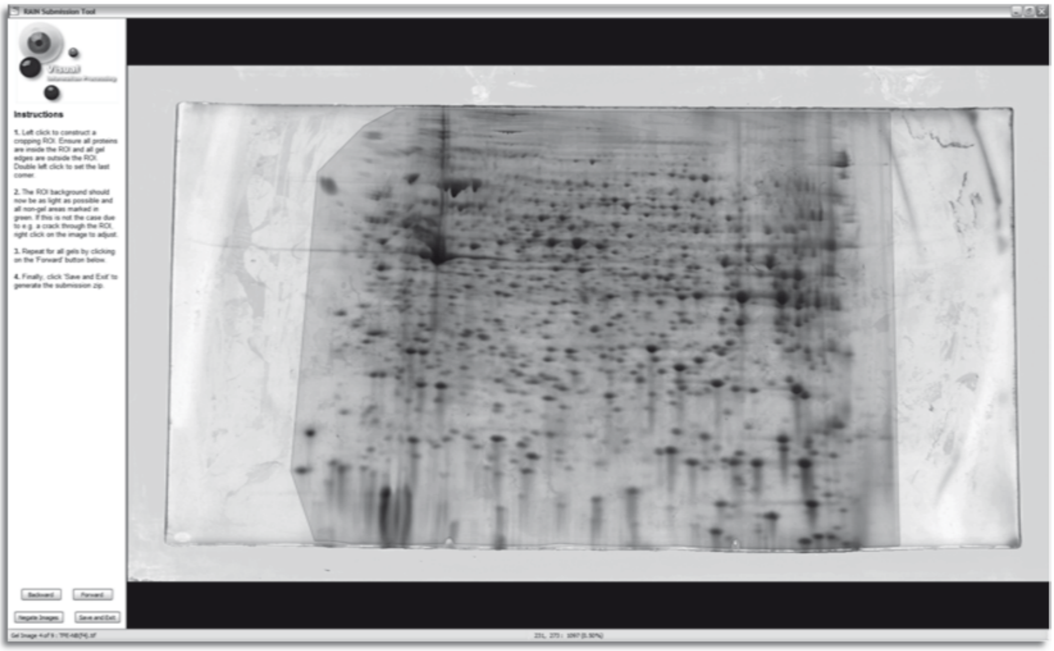
Today, typical experimental design should include enough biological replicate gels in each treatment group to confidently detect differential expression, though the optimal number is highly dependent on the tissue, sample preparation, and running protocols. It is therefore necessary to perform a few test experiments to optimize power as detailed in subheading 3.4 step 4. A good example of such a study is by Hunt et al. (17), where they determined that a sample size of 7–8 biological replicates would permit detection of a 50% change in protein expression in plasma samples. Since proteomics studies are challenging and time consuming, thorough planning of the experimental design is needed to protect against systematic bias. Therefore, standard design principles such as blocking and randomization of sample runs should be applied (18). Also, technical replicates should never be run at the expense of biological replicates. The study by Hunt et al. makes this point in dramatic fashion, showing much greater improvements in statistical power by increasing sample size rather than numbers of technical replicates. Nevertheless, if both must be mixed in the same experiment, both sources of error should be handled, see subheading 3.4 step 5.

A number of suggestions to bear in mind during 2-D gel running if informatics is to be facilitated are:

1. In general, the second dimension running should be consistent between samples. The spot matching and gel alignment algorithms will be confounded if some spots are visible in some gels and not in others. In any case, these spots will lack full statistical support for ascertaining differential expression.
2. Similarly, if one gel looks markedly different than other gels in the same treatment group, it should be discarded rather than incorporated into the analysis, since this will likely add significant outliers and therefore violate the assumption that the biological variation is normally distributed.
3. Saturation of abundant spots must be avoided as this will introduce increased error into the quantification (19). Moreover, the splitting of saturated complex spots is inaccurate regardless of the approach used.
4. Background and noise should be minimized otherwise dynamic range will be compromised, resulting in impaired sensitivity and specificity in spot detection. Danger areas include inadequate sample preparation and destaining, contaminated gels and too high laser scanner photo multiplier tube (PMT) voltages.
5. Between treatment groups, normalization can be greatly facilitated by employing DIGE and running a pooled sample on each gel as a paired control for gel normalization.
6. Within a given laboratory, studies should be performed to identify the key sources of experimental variability, and those factors should be accounted for using randomized block designs. These should be used to ensure that potentially important experimental factors are not confounded with factors of interest. See Note 3**.**
7. Another important design consideration is whether or not to pool samples. Pooling samples increases the protein load on each gel, which may reduce technical variability, but also results in loss of information about each sample. While sometimes necessary in order to obtain enough protein to reliably run the assay, pooling should be avoided since it results in a reduction of statistical power. When pooling samples, the key sample size factor is the number of pools, not the number of subjects. See Note 4.

3.1. Image Acquisition

1. The full dynamic range of the scanner should be utilized in order to maximize the number of weakly expressed spots visible above the noise floor. If your acquisition device does



Instructions

1. Left click to construct a cropping ROI. Ensure all proteins are inside the ROI and all gel edges are outside the ROI. Double left click to set the last corner.

2. The ROI background should now be as light as possible and all non-gel areas marked in green. If this is not the case due to e.g. a crack through the ROI, right click on the image to adjust.

3. Repeat for all gels by clicking on the 'Forward' button below.

4. Finally, click 'Save and Exit' to generate the submission zip.

Fig. 1. Gel image cropping with the RAIN submission tool. The *shaded area* shows an optimum polygonal crop that removes the gel edges and some artifacts while retaining the protein spots

not provide automatic calibration, this can be done with a step tablet (*e.g.*, UVP, Upland, CA) or a step wedge (*e.g.*, Stouffer Industries, Mishawaka, IN). If unavailable, a generic IT8 scanner target can be substituted as the bare minimum.

2. Once scanned, accurate cropping of the gels is essential. All gel edges must be outside the cropped region otherwise the alignment algorithm will attempt to align the edges at the detriment of aligning the spot patterns and erroneous spots will often be detected along the edges. Special care should be taken to ensure that every spot is inside the cropping region in all gels. If your informatics package supports irregular inclusion and exclusion regions, any artifact that appears on only a subset of the gels should be cropped away, such as cracks, fingerprints, and smudges. Suitable cropping is illustrated in Fig. 1.

3.2. Conventional Analysis Pipeline

1. Since spot detection is the first stage in conventional analysis, the gels must be first background subtracted to remove non-protein elements as well as all streaks and smears that do not adhere to the software's protein spot model. However, the removal is subjective and can interfere with surrounding real spots. See Note 5.

2. The spot detection process is then initiated on each gel separately. Typically, spot detection is controlled by setting a handful of algorithm-specific parameters, which should be optimized

for each experiment but fixed inside the same experiment. Unfortunately, optimization is a subjective process which requires a trade-off between false-positives (noise detected as spots, over-segmented spots) and false-negatives (spots failed to be detected, under-segmentation of merged spots).

3. Even with optimization, a significant amount of manual editing will be required post-hoc, which could introduce an element of subjective bias. In particular, calculation of spot boundaries is fraught with errors yet affects protein quantification significantly. See Note 6.

4. Once spots are quantified, a characteristic vector is extracted from each detected spot, which includes position, volume and perhaps shape and boundary information. These are combined to form a spot list for each gel. A reference gel is then manually chosen (or the software may suggest one), and in turn, each spot list is matched to the reference spot list. Since spots are matched between all the gels using the reference gel as an intermediary, any spots undetected on the reference will not be matched. See Note 7. Typically, the neighbourhood of each spot is used to facilitate the matching process, which is why outliers markedly affect the analysis.

5. It is expected that a significant number of weakly expressed spots will be detected only on a subset of the gels and therefore there will be a number of missing values in the resultant spot match list. These missing values reduce statistical power significantly and can introduce inadvertent bias. Thus, first of all, one should manually edit the spot detection and matching to maximize the number of successfully matched spots across gels. Invariably, there will still be some missing values in the spot match list, which must be dealt with in some way. Simply ignoring these spots for analysis introduces bias, since many of the gels with no matching spot likely had negligible or no expression of the corresponding protein. Substituting zeros or some other small value is a better option, but can still introduce bias, since it is expected that for some gels, there is evidence of some non-zero expression of the protein, but it simply fell below some arbitrarily specified detection threshold. Missing data is one of the major unsolved problems when using the conventional analysis pipeline.

6. Because of the nature of this traditional analysis pipeline, errors in automated spot matching increase as the experiment grows larger (7), meaning that the number of accurately matched spots decreases as increasing numbers of gels are run. This propagation of processing errors encourages researchers to run smaller studies that in turn are underpowered to statistically detect group differences when multiple testing is taken into account. The only current solution to this problem is to employ the image-based analysis paradigm instead.

3.3. Image-Based Analysis Pipeline

1. The first stage of the image-based pipeline is to manually select the gel with the most representative protein pattern and positions to be the reference image. The other gel images are then automatically warped so that their spot patterns are brought into alignment with that of the reference.
2. With Delta2D and Progenesis SameSpots, you must first manually identify a few spots that can be matched unambiguously in every gel in the set. The spots should be spread out evenly over the gel's surface otherwise some regions will be aligned too poorly now to be corrected later. The software will then (or after every landmark) automatically generate a smoothly interpolated warp that aligns these landmarks and estimates the intermediary alignment between them. If available, a further automatic phase can be initiated that adjusts the intermediary alignment to better match the remaining spots. These matches can be iteratively accepted or modified by the user and the algorithm rerun. Finally, alignments must be completed by hand and a "spot mask" applied to the reference gel of each set.

   With RAIN, fully automated image registration is performed by considering basic image gradients at several levels of detail and is therefore able to use extra image features such as global protein distribution, background, streaks, and smears in the alignment, as illustrated in Fig. 2. The set of gels is simply submitted to the ProteomeGRID web service (http://www.proteomegrid.org/), and after remote processing, the set of aligned images is available to download, together with visualisations to confirm the accuracy of the alignment.
3. The aligned images will then be automatically composed to create an image for subsequent spot detection. Pinnacle recommends computing an "average gel" that involves taking pixel-wise means of intensities across all gels. The key advantage of using the mean gel is that noise is reduced by $\sqrt{n}$ for a set of n gels, while the signal for true spots is reinforced across gels, thus substantially improving the sensitivity of detection for weak expression whilst suppressing highly variable features expected to be artifacts. Spot detection on the average gel will tend to have increased sensitivity over individual spot detection on each gel for any proteins present in more than $1/\sqrt{n}$ of the gels (11). Furthermore, variability will decrease as the gel sample size increases. Delta2D's fusion image is constructed by placing more emphasis on dark pixels likely to be protein matter (20). While the resulting image exhibits more spots than the average gel, statistically weak spots are artificially amplified so an increased number of false positives is possible.
4. Background subtraction and normalization may be applied to the average gel at this time. Background estimates can

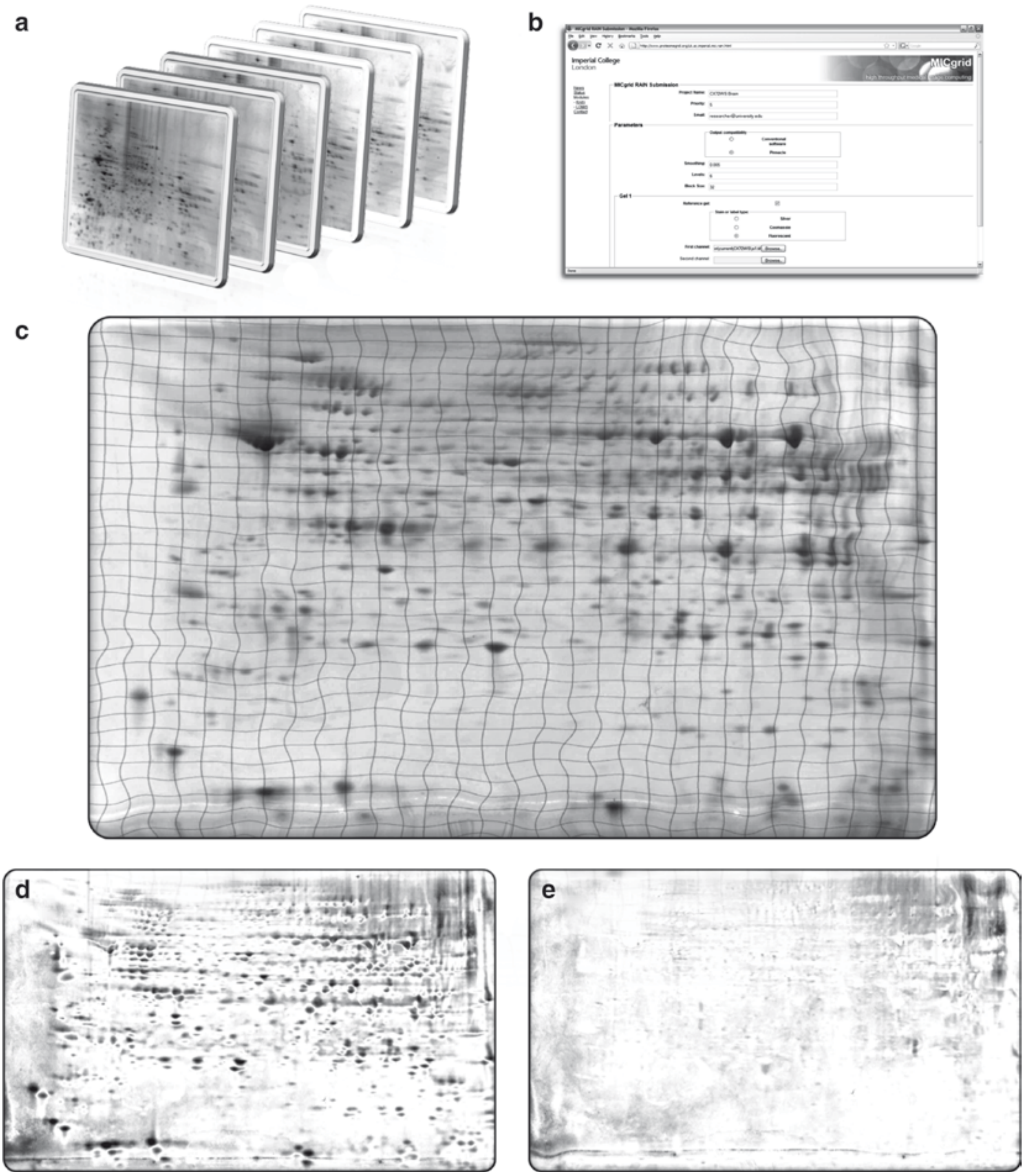


Fig. 2. Automatic image-based gel alignment by http://www.proteomegrid.org/. (**a**) Gels are cropped with the RAIN submission tool (see Fig. 1) and split into treatment groups. (**b**) The images are uploaded together with relevant metadata such as reference gel, stain/label used and DIGE channel. (**c**) Each gel is automatically aligned to the reference. The grid lines show the various levels of image warping needed. (**d**) Pixel-wise difference between reference and sample images before alignment. (**e**) After alignment, the differences should only be due to differential expression and artifacts, which will be differentiated by downstream spot modelling

be global or local (11). Normalization adjusts for gel-specific effects such as protein load. One common global method for normalization is to divide by the total volume, or average intensity, on the background corrected gel. If performed after spot detection and quantification, the sum total of quantified spot protein abundance can be employed instead. In DIGE experiments with a common reference channel, dividing each spot volume by its corresponding reference channel spot volume provides for a more precise normalization.

5. The next step is to detect spots on the average gel while obtaining spot quantifications for each spot on each gel. Progenesis SameSpots and Delta2D compute spot volumes on individual gels after detection of spot boundaries on the fusion image. Spots that are too weak to be detected on some gels by conventional means are able to be quantified by the consensus approach. However, while the resultant spot match list has "no missing values", the correctness of these values is exclusively dependent on their correct alignment. The spot detection results must be manually verified as described in Subheading 3.2 steps 2–3. See (7) for a comparison of Delta2D against conventional approaches. The spot boundaries are then copied onto each individual gel for quantification. See Note 8.

   After wavelet denoising the average gel image, Pinnacle focuses on peaks or "pinnacles" rather than spot boundaries and volumes in its detection and quantification algorithm. The idea is that non-saturated spots have well-defined pinnacles, and the intensity at this pinnacle is highly correlated with the spot volume but less affected by neighbouring spots. If saturation is avoided, peak detection on the average gel is sufficient to separate co-migrated spots, and furthermore, quantification using peak height only is more reliable and has greater validity than that derived through the spot boundary (11). Therefore, given a set of aligned images annotated by their treatment group, Pinnacle automatically outputs a list of peak intensities for each spot and each gel for downstream statistical significance testing as shown in Fig. 3.

3.4. Differential Analysis

After the previously described image processing, we are left with a matrix of spot quantifications for each spot across all gels. This matrix can be analyzed to discern which protein spots are differentially expressed across treatment groups.

1. *Transformation*: Frequently, the raw spot volumes or pinnacle intensities are highly skewed right, with many outliers, and the variance of a spot is related to its mean. These properties violate the assumptions underlying many statistical tests, such as *t*-tests or linear regression. To deal with this problem, it is possible to use some transformation of the spot quantifications before performing statistical analyses. Candidate transformations include the log, square root, and cube root transformations. Frequently, it helps to add a small constant (e.g., ½ or 1) before transforming the volumes in order to

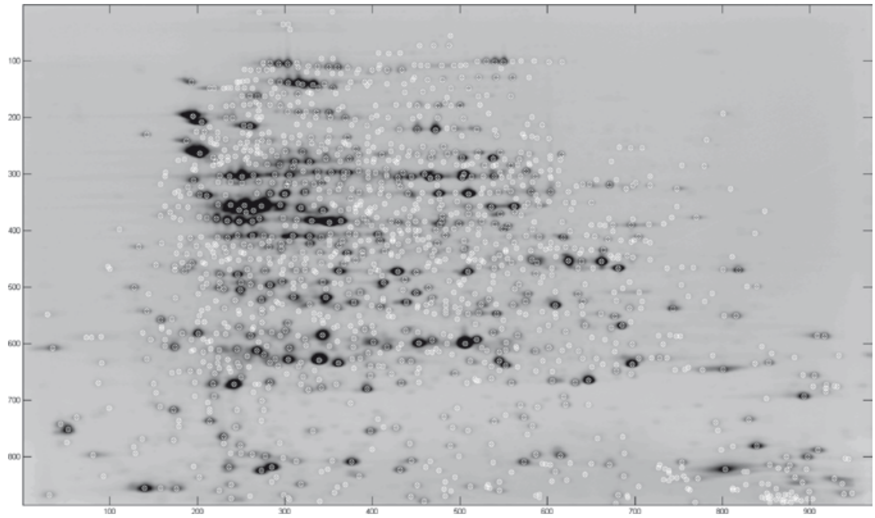


Fig. 3. Average gel computed after alignment using RAIN, with spots detected by Pinnacle marked with an "x"

avoid artifacts near zero intensities. *See* Note 9. Choice of transformation can be assessed by QQ-plots and histograms of the residuals from the statistical test of interest.

2. *Statistical tests for differential expression*: In the past, one way to assess differential expression of spots is to simply flag spots with the largest fold-changes across groups. This approach is statistically flawed, since fold change does not take the variability in the data into account, and thus makes it impossible to gauge the level of false positives. Appropriate statistical tests which take into account experimental variability are Student's *t*-test for two treatment groups, ANOVA for three or more treatment groups, and linear regression for quantitative correlative studies. If adequate normality cannot be obtained, the non-parametric Mann-Whitney and Kruskal-Wallis ANOVA tests can be substituted. After testing, each protein will be associated with a probability (the "*p*-value") that the observed difference could occur by chance. A histogram of the range of *p*-values can be checked for a peak near 0, which is a promising sign for significant differential expression. A peak elsewhere suggests technical problems with the gels.
3. *False-Discovery Rate-Based Thresholds*: Since *p*-values are obtained for each of many spots (100's or 1,000's) in the experiment, a *p*-value threshold of 0.05 would typically lead to a great deal of false positives, since we expect 5% of all spots to have *p*-values less than 0.05 even if there are truly no proteomic differences between groups. In recent years, various methods to estimate and control the false discovery rate (FDR) have arisen, and can be used to find appropriate *p*-value thresholds for declaring statistical significance.

Controlling the FDR at some level, say 0.05, means that of all spots we call differentially expressed, we expect only 5% of them to be false positives, and the other 95% true positives. See Note 10. Various other methods exist that are available for performing this analysis in Bioconductor/R, *e.g.* fdrtool. After using this method, a q-value or overall FDR threshold (typically 0.05, 0.10, or 0.20) is specified, and we obtain a list of differentially expressed spots.

4. *Power Calculations*: Must be performed since it will be necessary to redo the experiment with decreased variance or, usually more attainably, an increased sample size (number of replicates) if the statistical power is found to be too low. Software able to estimate the optimal sample size when the FDR is controlled is available as dictated in subheading 2.4. It is highly recommended that preliminary studies are performed so that the power calculations are based on the ranges of biological and technical variability for a particular experiment before a definitive protocol is laid down.
5. *Mixed Effects*: When multiple gels are obtained, one must take care in performing the statistical analysis since protein levels for replicate gels from the same individual are correlated with each other, violating the independence assumption underlying the test. One approach would be to average the spot quantifications across replicate gels to obtain one measurement per individual, and then analyze using a *t*-test or some other method assuming independence. Another alternative would be to use a method that takes this nested design into account. For example, a generalization of the *t*-test or ANOVA (21) or linear regression for correlated data would be a linear mixed model, including a fixed effect for treatment group, and random effect for the individual. Inference on the fixed effect from this model, then, yields a *p*-value that appropriately takes the correlation between gels from the same individual into account. Mixed models, *e.g.,* PROC MIXED, Cary, NC, can be implemented in standard statistical software, including SAS and R. See Note 11.
6. *Hierarchical Clustering*: This can be applied to the matrix of spot intensity values to see which samples cluster strongly together. In running this clustering, one can see how individuals within the same treatment group are similar. Also, these can be useful diagnostics to see whether there is some experimental factor that may have been strongly influential in the study. For example, if all samples run in the same IEF block cluster strongly together, that could indicate that something happened with that IEF block to make its gels different than the others. These can be valuable indicators to aid in the design of future studies.

4. Notes

1. Never manipulate the images in a generic image-editing package such as Photoshop (Adobe, San Jose, CA) before analysis. Even if the process appears risk-free such as cropping, a number of side-effects can occur silently. For example, calibration curves and metadata are likely to be lost, or the images may be quietly converted to 8 bit.
2. Never attempt quantification on images saved in a lossy compression format such as JPEG. Not only are these limited to 8 bits of dynamic range, but they also remove details that are essential for accurate protein quantification.
3. For example, if we have a case control study and the isoelectric focusing is performed in blocks of 8, ensure that for each run, 4 cases and 4 controls are run in the same block, with the positions determined based on a random number generator. This way, any variability in IEF runs will not mistakenly appear as a case/control effect, which can happen when IEF run and case/control are confounded.
4. It can be shown that maximizing the number of pools minimizes the total variance, yielding maximum power. A disastrous design would be to combine all cases into one pool and all controls into another pool, and then run replicate gels from each pool. If this design is used, it is impossible to assess biological variability, since the variation across gels would only capture technical variability. As a result, it would not be possible to do a valid statistical assessment of differential expression. Thus, if pooling is deemed necessary, one should maximize the number of pools, and make sure to have multiple pools per treatment group.
5. The background subtraction task is never perfectly discriminating, and therefore it is usually performed conservatively to ensure that the accuracy of protein quantification is not adversely affected.
6. In order to maximize the effectiveness of the spot-matching phase and minimize further manual verification, when using conventional analysis methods, it is important to make editing decisions consistently over the gel set. This is obviously a difficult task because of the migration variability between gels, and is a significant limitation of the conventional pipeline.
7. Some software may have an option to match every spot list to every other spot list to avoid this limitation. In this case, the reference gel is used only to define the fixed positional reference frame to which the other gel's spots are migrated to.

8. Typical image-warping techniques do not preserve the amount of protein in each spot, leading to over-expression in areas of dilation and under-expression in areas of contraction. In order to avoid this issue, Progensis SameSpots and Delta2D unwarp each consensus spot boundary so that it can be applied to the original unaligned gels where quantification takes place. RAIN applies a volume-invariant warping procedure, wherein each pixel is weighted by its change in size, thereby allowing accurate quantification on warped gels.
9. One benefit of the log transformation is that a difference in the log scale corresponds to a multiplicative fold-change in the raw scale. For example, if a $\log_2$ transformation is used, a difference of 1 between groups corresponds to a two-fold multiplicative difference.
10. One simple method models the *p*-value histogram as a mixture of two distributions: the null distribution (true negatives and false positives) as the underlying uniform distribution; and the alternative distribution (true positives and false negatives) as a right skewed distribution with mode near zero. From this, it is possible to estimate the probability of a false positive for each *p*-value (this probability is called a q-value), and to estimate a cutpoint on *p*-values that controls the overall FDR at a prescribed level.
11. The mixed models can also be used in the design phase, using preliminary studies on the tissue of interest to estimate levels of technical and biological variability for spots in the study. These estimates can then be used to perform power calculations and make design determinations, e.g., if the technical variability is very large relative to biological variability, then it may be helpful to run several replicate arrays for each biological sample.

References

1. P. H. O'Farrell, High resolution two-dimensional electrophoresis of proteins, *Journal of Biological Chemistry,* vol. 250, pp. 4007–4021, 1975.
2. J. Klose, Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals, *Humangenetik,* vol. 26, pp. 231–43, 1975.
3. T. Rabilloud, Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains, *Proteomics,* vol. 2, pp. 3–10, 2002.
4. D. A. Wolters, M. P. Washburn, and J. R. Yates, An automated multidimensional protein identification technology for shotgun proteomics, *Anal. Chem.,* vol. 73, pp. 5683–90, 2001.
5. M. R. Roe and T. J. Griffin, Gel-free mass spectrometry-based high throughput proteomics: Tools for studying biological response of proteins and proteomes, *Proteomics,* vol. 6, pp. 4678–4687, 2006.
6. K. A. Reidegeld, M. Müller, C. Stephan, M. Blüggel, M. Hamacher, L. Martens, G. Körting, D. C. Chamrad, D. Parkinson, R. Apweiler, H. E. Meyer, and K. Marcus, The power of cooperative investigation, *Proteomics,* vol. 6, pp. 4997–14, 2006.
7. B. N. Clark and H. B. Gutstein, The myth of automated, high-throughput two-dimensional gel analysis, *Proteomics,* vol. 8, pp. 1197–1203, 2008.

8. S. Veeser, M. J. Dunn, and G. Z. Yang, Multiresolution image registration for two-dimensional gel electrophoresis, *Proteomics,* vol. 1, pp. 856–870, 2001.
9. A. W. Dowsey, M. J. Dunn, and G.-Z. Yang, Automated image alignment for 2-D gel electrophoresis in a high-throughput proteomics pipeline, *Bioinformatics,* vol. 24, pp. 950–957, 2008.
10. A. W. Dowsey, J. English, K. Pennington, D. Cotter, K. Stuehler, K. Marcus, H. E. Meyer, M. J. Dunn, and G.-Z. Yang, Examination of 2-DE in the human proteome organisation brain proteome project pilot studies with the new RAIN gel matching technique, *Proteomics,* vol. 6, pp. 5030–5047, 2006.
11. J. S. Morris, B. C. Walla, and H. B. Gutstein, Pinnacle: A fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data, *Bioinformatics,* vol. 24, pp. 529–536, 2008.
12. A. W. Dowsey and G. Z. Yang, The future of large-scale collaborative proteomics, *Proceedings of the IEEE,* vol. 96, pp. 1292–1309, 2008.
13. A. W. Dowsey, M. J. Dunn, and G. Z. Yang, The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics,* vol. 3 pp. 1567–1596, 2003.
14. M. Unlu, M. E. Morgan, and J. S. Minden, Difference gel electrophoresis: a single gel method for detecting changes in protein extracts, *Electrophoresis,* vol. 18, pp. 2071–2077, 1997.
15. L. Vincent and P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 13, pp. 583–598, 1991.

. K. Miura, Imaging technologies for the detection of multiple stains in proteomics, *Proteomics,* vol. 3, pp. 1097–1108, 2003.

17. S. M. N. Hunt, M. R. Thomas, L. T. Sebastian, S. K. Pedersen, R. L. Harcourt, A. J. Sloane, and M. R. Wilkins, Optimal replication and the importance of experimental design for gel-based quantitative proteomics, *Journal of Proteome Research,* vol. 4, pp. 809–819, 2005.
18. G. E. P. Box, W. G. Hunter, and J. S. Hunter, Statistics for experimenters: an introduction to design, data analysis, and model building, Wiley, New York, 1978.
19. J. S. Almeida, R. Stanislaus, E. Krug, and J. M. Arthur, Normalization and analysis of residual variation in two-dimensional gel electrophoresis for quantitative differential proteomics, *Proteomics,* vol. 5, pp. 1242–1249, 04 2005.
20. S. Luhn, M. Berth, M. Hecker, and J. Bernhardt, Using standard positions and image fusion to create proteome maps from collections of two-dimensional gel electrophoresis images, *Proteomics,* vol. 3, pp. 1117–1127, 2003.
21. G. W. Horgan, Sample size and replication in 2D gel electrophoresis studies, *Journal of Proteome Research,* vol. 6, pp. 2884–2887, 2007.

Chapter 17

Automated Generic Analysis Tools for Protein Quantitation Using Stable Isotope Labeling

Wen-Lian Hsu and Ting-Yi Sung

Abstract

Isotope labeling combined with LC-MS/MS provides a robust platform for quantitative proteomics. Protein quantitation based on mass spectral data falls into two categories: one determined by MS/MS scans, e.g., iTRAQ-labeling quantitation, and the other by MS scans, e.g., quantitation using SILAC, ICAT, or ^{18}O labeling. In large-scale LC-MS proteomic experiments, tens of thousands of MS and MS/MS spectra are generated and need to be analyzed. Data noise further complicates the data analysis. In this chapter, we present two automated tools, called Multi-Q and MaXIC-Q, for MS/MS- and MS-based quantitation analysis. They are designed as generic platforms that can accommodate search results from SEQUEST and Mascot, as well as mzXML files converted from raw files produced by various mass spectrometers. Toward accurate quantitation analysis, Multi-Q determines detection limits of the user's instrument to filter out outliers and MaXIC-Q adopts stringent validation on our constructed projected ion mass spectra to ensure correct data for quantitation.

Key words: Computer software, Stable isotope labeling, Mass spectrometry, Quantitative proteomics, Quantitation analysis, Dynamic range, Extracted ion chromatogram, Projected ion mass spectrum

1. Introduction

In the post-genomic era, liquid chromatography (LC) combined with tandem mass spectrometry (MS/MS) (1, 2) has opened up a new dimension in proteomics research, which presents a large-scale, robust, and sensitive technology for protein profiling (3, 4). Recent advances in incorporating stable isotope labeling strategies into MS-based proteomics have further facilitated quantitation studies of differentially expressed levels of proteins in complex biological samples (5). In these studies, experimental samples are separately labeled with isotopically distinct reagents (6). The labeled proteins are then digested, mixed, fractionated,

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_17, © Humana Press, a part of Springer Science+Business Media, LLC 2010

and subjected to LC-MS/MS. In contrast to protein identification by MS/MS data, protein quantitation based on MS signals falls into two categories: one determined by MS/MS data, and the other by MS scans.

MS/MS-based quantitation is determined by the relative intensities of fragment peaks at fixed *m/z* values within an MS/MS spectrum. For example, Tandem Mass Tags (TMT) duplex or sixplex (7) are chemical labels for MS/MS-based quantitation. In addition, a typical example is quantitation using 4-plex/8-plex iTRAQ-labeling (8–10), which is based on a set of four/eight isobaric reagents. Each isobaric reagent comprises three groups: reporter, balance, and reactive groups. After cell lysis and protein digestion, the peptides in four/eight states are separately labeled on N-terminals and lysine residues by the reactive group of iTRAQ. The 8-plex iTRAQ-labeled peptides generate reporter groups with signature ions at *m/z* 113–119 and 121 in an MS/MS spectrum, while the 4-plex iTRAQ-labeled peptides at *m/z* 114–117. In addition to identifying proteins, the intensities of the four/eight MS/MS signature ions represent the quantities of the corresponding peptides, respectively, and are used for quantitation of peptide and protein expression levels.

Stable isotope labeling techniques for quantitation based on MS scans can be divided into three major categories: chemical labeling, e.g., ICAT (11); enzymatic labeling, e.g., ^{18}O-labeling (12); and metabolic labeling, e.g., SILAC (13–16). For quantitation analysis, the extracted ion chromatograms (XICs) of paired heavy- and light-labeled peptide ions are constructed and the areas of the XICs are used to calculate the peptide ratio.

As LC-MS/MS is routinely used following the quantitative labeling strategy, a proteomic experiment usually generates tens of thousands of MS and MS/MS spectra (the size of raw data usually amounts to several gigabytes). Many thousands of peptides can be collectively analyzed by multiple LC-MS/MS runs and hundreds, or even thousands, of proteins can be identified. A protein can be identified and quantified by more peptides, thereby enhancing the confidence of protein quantitation (17). However, in this approach, sample complexity increases substantially, which presents a great challenge for data processing. Furthermore, noise and limitations from the experiments or instruments, e.g., limited dynamic range of the instrument, and coeluting peptides in peptide profiling, further complicate the analysis. Notably, in MS-based quantitation, XICs sometimes have difficulty in achieving systematic quantitation of complex peptide mixtures because problems, such as insufficient chromatographic separation and slow MS acquisition, usually complicate XICs' construction and computation. Bioinformatics tools that can effectively tackle the data analysis challenges are essential for proteomics research.

2. Materials

Several quantitation analysis tools have been made available, e.g., i-TRACKER (18), ProQUANT (Applied Biosystems) and Libra (http://tools.proteomecenter.org/Libra.php) for MS/MS-based quantitation (see Note 1), and XPRESS(19), ASAPRatio (20), RelEx (21), MSQuant (http://msquant.sourceforge.net/) for MS-based quantitation (see Note 2). In this chapter, we present two automated generic tools for quantitation analysis, Multi-Q (22, 23) for MS/MS based quantitation and MaXIC-Q (24) for MS-based quantitation, that aim at convenient and accurate quantitation analysis with much reduced manual validation effort (see Note 3). To achieve these aims, we consider the following issues in the development of our tools:

1. Supporting input data formats from various mass spectrometers and protein identification search engines,
2. Coping with limitations arising from instruments or experiments,
3. Data validation,
4. Visualization of spectral data and quantitation results,
5. Ease of installation and use.

The workflows of Multi-Q and MaXIC-Q are depicted in Fig. 1. Step 1 of both tools is input data preparation and filtering. Only confidently identified peptides and proteins proceed to subsequent analysis steps. The tools are available for download at http://ms.iis.sinica.edu.tw/Multi-Q and http://ms.iis.sinica.edu.tw/MaXIC-Q.

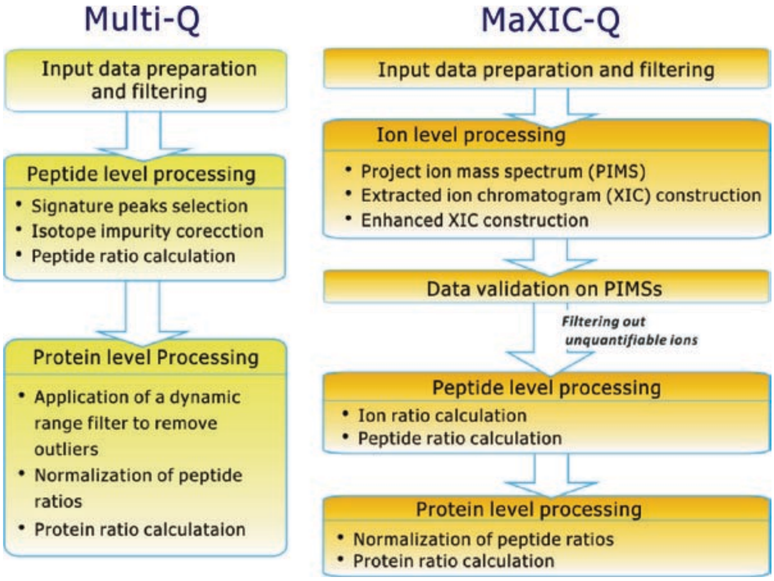

Multi-Q
Input data preparation and filtering
Peptide level processing
• Signature peaks selection
• Isotope impurity corecction
• Peptide ratio calculation
Protein level Processing
• Application of a dynamic range filter to remove outliers
• Normalization of peptide ratios
• Protein ratio calculataion
MaXIC-Q
Input data preparation and filtering
Ion level processing
• Project ion mass spectrum (PIMS)
• Extracted ion chromatogram (XIC) construction
• Enhanced XIC construction
Data validation on PIMSs
Filtering out unquantifiable ions
Peptide level processing
• Ion ratio calculation
• Peptide ratio calculation
Protein level processing
• Normalization of peptide ratios
• Protein ratio calculation

Fig. 1. Workflows of Multi-Q and MaXIC-Q

3. Methods

3.1. Input Data Preparation and Filtering

To accommodate spectral data files generated by different mass spectrometers, Multi-Q and MaXIC-Q adopt the standard mzXML format, developed by Institute for Systems Biology, for spectral data input. As shown in Fig. 2, spectral data files from the major MS manufacturers can be converted into the mzXML format by existing tools, such as mzStar for *.wiff* files from Applied Biosystems, ReAdW (http://tools.proteomecenter.org/ReAdW.php) for *.raw* files from Thermal Finnigan, and MassWolf (http://tools.proteomecenter.org/MassWolf.php) for *.raw* directories from Waters. The converted mzXML format incorporates all the necessary attributes, including MS and MS/MS peak lists and the scan number index for subsequent quantitation.

Multi-Q and MaXIC-Q accept search result files from commonly used database search engines, including Mascot and SEQUEST. SEQUEST users need to input pepXML and protXML files generated by PeptideProphet (25) and ProteinProphet (26) (in the Trans-Proteomic Pipeline, Institute for Systems Biology), respectively. For Mascot users, both tools accept the CSV and XML formats exported directly from the Mascot web interface.

To ensure quantitation from confidently identified peptides and proteins, both tools select only confident search results as input for quantitation. To filter out low-confidence identification hits from SEQUEST, statistical validation by PeptideProphet and

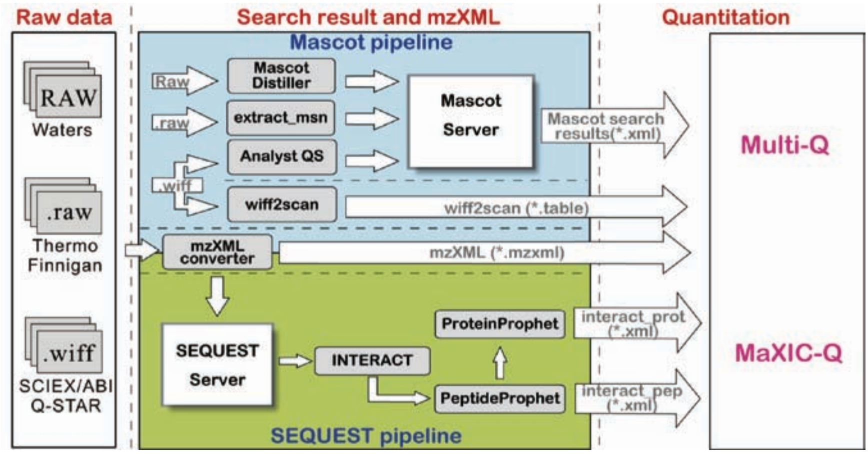


Fig. 2. Input data preparation for Multi-Q and MaXIC-Q. Raw data files generated by various mass spectrometers need to be converted into mzXML files using existing converters. The raw data files or mzXML files are then searched by either the Mascot or SEQUEST pipeline. The files can be searched directly by the Mascot server through varous tools. Users can also use SEQUEST servers to search mzXML files. The searched results need to be processed by PeptideProphet and ProteinProphet to generate pepXML files (referred to as *interact_prot.xml*) and protXML files (referred to as *interact_pep.xml*). The mzXML files and search results then serve as input for both programs

ProteinProphet is usually used to evaluate the confidence of identified peptides and proteins. For Mascot users, both tools use Mascot output identification scores, based on "Standard scoring" or "MudPIT scoring", as a filtering criterion. We discuss Multi-Q in Subheading 3.2 and MaXIC-Q in Subheading 3.3.

3.2. Data Analysis in Multi-Q

After data preparation, Multi-Q selects iTRAQ-labeled peptides with confident MS/MS identification, detects signature ions, and performs automated quantitation of peptide abundance.

3.2.1. Peptide Level Processing

3.2.1.1. Signature Ion Detection and Background Subtraction

To determine the peptide ratios from an MS/MS spectrum, Multi-Q first smoothes the spectrum, selects signature peaks with specific *m/z* values (i.e., 114–117 for 4-plex analyses and 113–119, 121 for 8-plex analyses) from the smoothed spectrum, and performs background subtraction. Spectrum smoothing is implemented by the 3-point moving average (27) method. The mass tolerance of signature peak detection is defined by users based on the mass accuracy of their instruments. For example, the peak apexes within ±0.2 *m/z* of 114, 115, 116, and 117 in 4-plex analyses are selected as signature peaks based on the mass accuracy of the quadrupole time-of-flight (Q-TOF) mass spectrometer in our quantitation analysis. For background subtraction, the spectrum baseline is defined as the mean of all the valleys in the smoothed curve, and the valleys are determined by calculating the first and second order derivatives of the curve. The peak intensity is then calculated by subtracting the baseline from the original data.

3.2.1.2. Isotope Impurity Correction

The isotopically distinct iTRAQ tags cause variations in the true peak intensity such that each batch of iTRAQ reagents contains trace levels of isotopic impurities that must be corrected. We use 4-plex iTRAQ reagents to illustrate isotope impurity correction. As the isotopic distributions of the 114–117 signature peaks interfere with each other, over- or under-representation of signature ions will occur. For instance, the *m/*z 116 peak is a composite peak with contributions from isotope envelopes of *m/z* 114, 115, and 117. Using the impurity information in the "Certificate of Analysis" provided by the iTRAQ reagent manufacturer, the interference of an isotopic reagent with its two predecessors and two successors can be corrected by the following linear equations:

$$\begin{pmatrix} y_{114} \\ y_{115} \\ y_{116} \\ y_{117} \end{pmatrix} = \begin{pmatrix} 1 & k_{115,-1} & k_{116,-2} & 0 \\ k_{114,+1} & 1 & k_{116,-1} & k_{117,-2} \\ k_{114,+2} & k_{115,+1} & 1 & k_{117,-1} \\ 0 & k_{115,+2} & k_{116,+1} & 1 \end{pmatrix} \begin{pmatrix} x_{114} \\ x_{115} \\ x_{116} \\ x_{117} \end{pmatrix} \tag{1}$$

where y_i denotes an observed ion count, x_j denotes a real ion count, $k_{i,j}$ denotes the correction factor of iTRAQ reagent *i's* effect on its *j*-th predecessor/successor signature peak. For example, $k_{114,+2}$ denotes the correction factor of iTRAQ reagent 114's interference with the signature peak at *m/z* 116. All x_i can be obtained easily by performing Gaussian Elimination (28).

3.2.1.3. Peptide Ratio Determination

After peak detection and isotope impurity correction, Multi-Q calculates the peptide ratios according to the peak intensities of the signature peaks. The relative peptide ratio $r_{A/B}(u)$ of a peptide *u* in two different cell states, A and B, is expressed as

$$r_{\mathrm{A/B}}(u) = \frac{|u_{\mathrm{A}}|}{|u_{\mathrm{B}}|} \qquad (2)$$

where $|u_A|$ and $|u_B|$ denote the abundance of peptide *u* in cell states A and B, respectively.

3.2.2. Protein Level Processing

In the design of Multi-Q, we consider corrections of peptide ratio errors caused by limited dynamic range of the instrument used and systematic errors in the experiment procedure or instrument. Prior to protein quantitation, peptide ratios outside the dynamic range must be removed and peptide ratio normalization is performed. Normalized ratios of nondegenerate peptides are used to calculate protein ratios.

3.2.2.1. Determination and Application of a Dynamic Range Filter

We have conducted an experiment with 10-standard-protein mixture to investigate the issue of dynamic range. In this experiment, we observed that peak intensities of high-abundance peptide ions may be underestimated by mass spectrometers due to the saturation effect of detector's responses, which is a common phenomenon, particularly in nanospray or nano-LC experiments. For example, Fig. 3a shows the MS/MS spectrum for the triple-charged iTRAQ labeled peptide (NTDGSTDYGILQINSR from Lysozyme C, P22910) in the *m/z* window 113–120. The peaks' clusters have an abundance ratio of 1:2:1:0.5 in the four samples that conforms with the expected value in a standard protein mixture experiment. In contrast, Fig. 3b shows the double-labeled peptide (labeled on the N-terminal and the lysine residue) from the same protein, CELAAAMK, exhibits nearly identical peak intensities because of the saturated signals in *m/z* 114, 115, 116, and 117. The observed ratio does not reflect the expected ratio 1:2:1:0.5.

To determine the signal saturation threshold of the instrument we used (Q-STAR Applied Biosystems, USA), we illustrate the distribution of paired peak intensities in Fig. 4, which shows the global linear correlation of the paired peak intensities for the total number of peptides in 114 versus 115, 114 versus 116, and 114 versus 117. In the 1:1 experiment (*m/z* 114: *m/z* 116, indicated by solid triangles), the least square regression of the scatter plot reveals a linear fit with a slope of 1.19, which conforms with the expected 1:1 ratio. However, the plot of *m/z* 114 versus

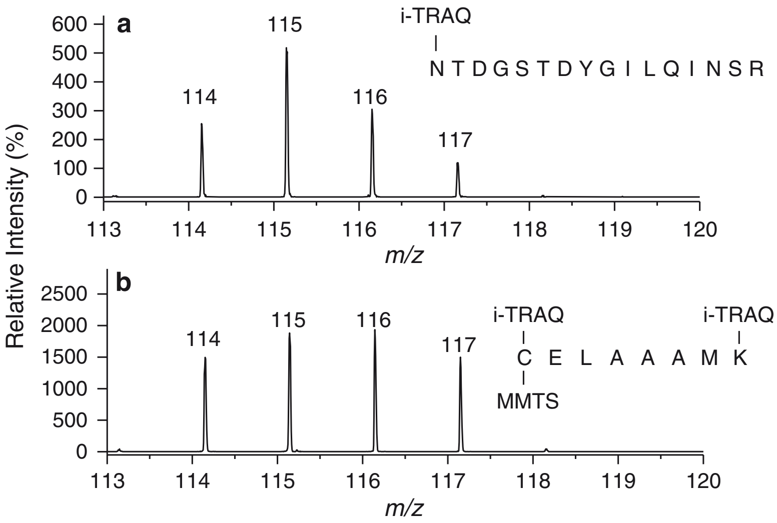


Fig. 3. MS/MS spectra of iTRAQ labeled peptides with different signature peak abundance in the *m/z* 114-118 region. (**a**) The average signature peak intensity reveals clearly 1:2:1:0.5, in agreement with the expected value of NTDGSTDYGILQINSR from Lysozyme C. (**b**) Due to signal saturation effect, the average signature peak intensity of CELAAAML, also from Lysozyme C shows nearly identical peak intensities. (Reprinted with permission from [22]. Copyright © 2006 American Chemical Society)

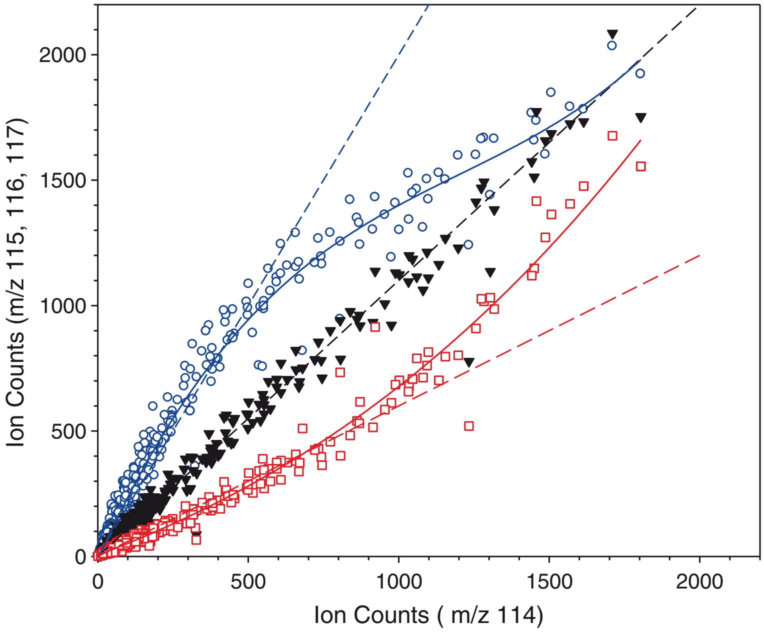


Fig. 4. Peak intensities of peptides with two different labels. Each point represents a peptide with two different labels, as shown by a circle for 115/114 (with expected ratio 2), a triangle for 116/114 (with expected ratio 1), and a rectangle for 117/114 (with expected ratio 0.5). The dashed lines represent the linear regression of unsaturated ion counts, while the solid curves represent fitting curves for ion counts of all peptides. Deflecting trends toward 1:1 ratio (*x*=*y*) are observed when the peak intensity is over approximately 1,000 counts. (Reprinted with permission from [22]. Copyright © 2006 American Chemical Society)

m/z 115, indicated by circles, reflects a non-linear curve with a deflecting trend when the peak intensity is slightly above 1,000 counts. A deviating quadratic trend is also observed in the *m/z* 114 versus *m/z* 117 experiment indicated by squares. The curve substantially deviates from the linear fit with a slope of 0.5 for intensities over 1,000 counts. Thus, we can decide the saturation threshold of the instrument.

In addition to signal saturation, signal fluctuations of low intensity threshold is another factor for determining the dynamic range (see Note 4). To demonstrate the intensity dependence of measured peptide ratios, the *m/z* 116/114 signal ratios (i.e., 1:1) of our experiment are plotted as a function of the average ion counts of the signal *m/z* 116/114 as shown in Fig. 5. There is a significant fluctuation in the 116/114 ratio of low intensity peptides compared to that of high intensity peptides. In particular, low intensity peptides, such as those with ion counts below 100, displayed wide fluctuations.

To avoid the above-mentioned fluctuations and improve quantitation accuracy, Multi-Q determines an intensity threshold to remove peptide ratios derived from low intensity spectra by analyzing the distribution of peptide ratios with a Gaussian fitting (28) function (the inset in Fig. 5) as follows:

$$f(r, n_r, R, \sigma) = n_r \times e^{-(r-R)^2/(2\sigma^2)} \quad (3)$$

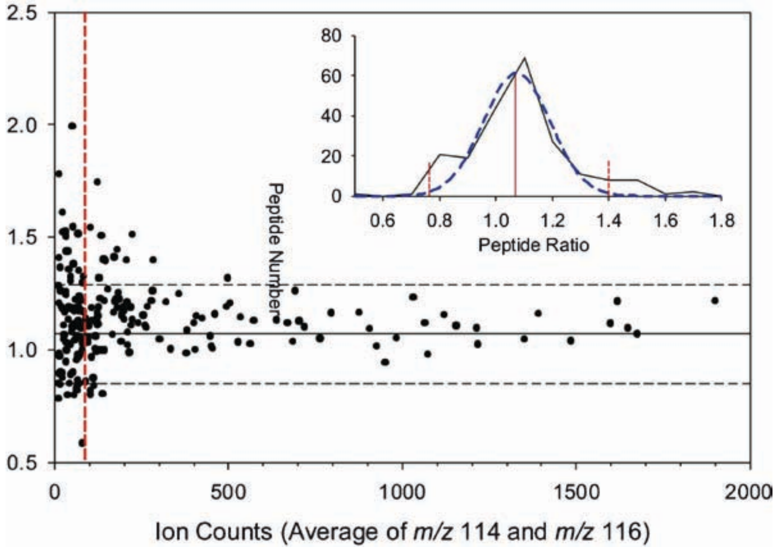


Fig. 5. Effect of ion count on measured peptide ratio. The 116/114 ratios from the 10-standard-protein mixture are plotted as a function of ion count. Fluctuation of peptide ratios is greater in low ion counts than that in high ion counts. *Inset* shows the distribution of peptide ratios, where the solid line is the original data distribution, and the dashed line is the Gaussian fitting curve over the original data. The ratios belonging to the outlier group (99.7% confidence interval) of the fitting curves are considered as deviated ratios. (Reprinted with permission from [22]. Copyright © 2006 American Chemical Society)

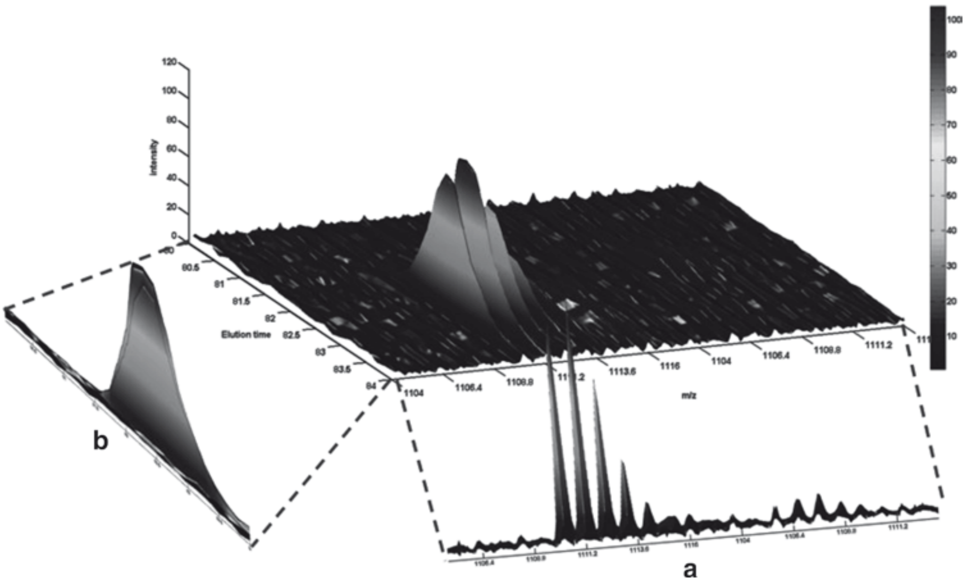


Fig. 6. Construction of the PIMS and XIC. The MS survey scans are represented by a 3D model, in which the different shades of gray show the intensity of the peaks. (**a**) Construction of PIMS: *n* MS survey scans are projected, and the maximal intensities of overlapping peaks are used to construct the PIMS. (**b**) Construction of XIC: each point in the XIC represents the total intensity of signals within the selected *m*/*z* range of the corresponding MS survey scan

where r is a peptide ratio; n_r and σ denote the number and the standard deviation of r, respectively; and R is the mode (also the mean) of the fitted Gaussian distribution that represents the ratio of the most abundant peptides. In the inset of Fig. 5, the standard deviation and mean of the fitted curve are 0.11 and 1.07, respectively. Many low abundance peptides lie in outlier groups of the fitted curve with a 99.7 % confidence interval ($\mu \pm 3\sigma$) (*see* **Note 5**). Since they are considered as deviate ratios, they are discarded before protein ratio determination. Multi-Q calculates the intensity threshold for the low S/N ratio cut-off based on the average intensities of these outlier groups. Multi-Q filters out peptides having intensities outside the dynamic range for subsequent analysis (*see* **Note 6**).

3.2.2.2. Normalization of Peptide Ratios

In principle, most protein expression levels in cells remain unchanged between two different cell states, i.e., most peptide ratios should be equal to 1, and therefore, most of them follow a normal distribution pattern with a mode (i.e., the highest value in the distribution) close to 1. In practice, however, mass spectrometry-based quantitation results have bias due to between-assay variations, e.g., inconsistencies in protein digestion efficiency, iTRAQ labeling yield, and isotope impurities, and between-sample variations, e.g., variable purity or concentrations of independently prepared proteins. As a result, there may be a uniform bias in favor of all peptide ratios such that most ratios deviate from 1 and a normalization procedure is essential.

Multi-Q performs a normalization procedure fitting the peptide ratios into the Gaussian distribution as shown in Eq. 3 (see Note 7). All peptide ratios are multiplied by a normalization factor, the reciprocal of *R*, to correct the systematic bias.

3.2.2.3. Determination of Protein Abundance Ratios Based on Non-degenerate Peptides

Based on nondegenerate peptides, Multi-Q calculates the expression ratio $r_{A/B}(p)$ of a protein *p* in two cell states A and B as the weighted sum of all corresponding non degenerate peptide ratios as follows:

$$r_{\mathrm{A/B}}(p) = \sum_{u \in U} w_u r_{\mathrm{A/B}}(u) \tag{4}$$

where $r_{A/B}(u)$ is defined same as in Eq. 2, *U* is the set containing all non degenerate peptides of *p*, and w_u is the weight of the ratio $r_{A/B}(u)$ given by

$$w_u = \frac{(|u_\mathrm{A}| + |u_\mathrm{B}|)}{\sum_{v \in U} (|v_\mathrm{A}| + |v_\mathrm{B}|)} \tag{5}$$

where $|u_A|$, $|u_B|$, $|v_A|$, and $|u_B|$ denote the abundance of peptides *u* and *v* in cell states A and B, respectively. Also, Multi-Q provides an additional option for calculating protein ratios by unweighted peptide ratios.

3.3. Data Analysis in MaXIC-Q

3.3.1. Ion Level processing

After input data preparation and filtering, MaXIC-Q constructs the projected ion mass spectrum (PIMS) and constructs the corresponding extracted ion chromatogram (XIC) from the elution profile for each isotope-labeled, confidently identified peptide. The area of XIC is used to determine the peptide ion abundance, which is subsequently used for determining peptide ion ratio. Conventionally, an ion mass spectrum is constructed from a single MS scan at the elution time when the peptide ion is identified. However, the presence of other peptides and noises may co-elute during the elution period of the peptide ion that cannot be detected by such ion mass spectrum. Therefore, instead, we construct the PIMS that covers MS scans in a range of the elution time of the identified peptide. PIMS is used for detecting noises and co-eluting peptides in the elution profile and will be validated whether the corresponding XIC is constructed from good spectral data and can be used for quantitation (see Subheading 3.3.2).

3.3.1.1. Constructing the PIMS

MaXIC-Q retrieves every MS scan within 120 s of the elution time of the identified peptide ion to ensure obtaining a complete elution profile of the ion. For each retrieved MS scan, all peaks located in the *m/z* range of the predicted ion isotope distribution, which can be automatically inferred from the identified sequence and its charge state, are extracted and projected to form a raw PIMS. The construction concepts of PIMS and XIC are shown in Fig. 6. The seven-point Savitzky-Golay smoothing algorithm (29) is then applied to the PIMS to remove noise. Next, MaXIC-Q performs peak detection and then background subtraction (as described in Subheading 3.2.1.1) on the smoothed spectrum.

3.3.1.2. Constructing the XIC

To construct the XIC from the PIMS, we first determine the *m/z* range of the precursor ion from the PIMS. If the peak in the PIMS containing the precursor ion has a good signal-to-noise ratio (default: S/N≥2.5), we use the *m/z* width of the peak to construct the XIC; otherwise, the *m/z* range of the precursor within a predefined window (user-defined, default setting: ±0.5 *m/z*) is selected (see Note 8). A raw XIC is constructed by summing all ion intensities after subtracting background in each MS scan within the selected *m/z* range and over the retention time of the peptide ion. Then the smoothed XIC is obtained by the B-spline smoothing algorithm (30), which has been widely used to fit a free-form curve in digital signal processing and microarray analysis.

3.3.1.3. Evaluating the XICs of Light- and Heavy-Labeled Ions to Reconstruct a Refined XIC

The scan range 120 s used to construct the XIC is so broad that interfering intensities (e.g., co-eluting peptides) may be included in the elution profile. To resolve this problem, MaXIC-Q automatically reconstructs the XIC based on a refined scan range determined from the primitive XIC. That is, MaXIC-Q aligns the corresponding elution peaks of the XICs of light-labeled and heavy-labeled ions by their highest points and extracts the scan range of the overlapping areas of the aligned peaks for reconstructing XICs.

3.3.2. PIMS Validation

Though quantitation is based on the area of XICs, the quality of PIMS critically affects the quantitation accuracy. In practice, noise or co-eluting ions usually cause either under- or over-estimated XICs; however, it is difficult to identify such XICs by existing XIC-based methods. MaXIC-Q adopts three criteria, the signal-to-noise ratio (S/N), the charge state (CS), and the isotope pattern (IP), to check the validity of PIMSs and determine whether the corresponding paired ions are quantifiable.

MaXIC-Q first evaluates whether a PIMS is acceptable in terms of the S/N criterion. The procedure checks whether the three peaks nearby the precursor *m/z* and monoisotopic peaks with good S/N, say ≥ 2.5 (see Note 9). If the spectrum fails this criterion, it is noisy or the peptide ion expression is too low and the validation procedure stops. If all three isotopic peaks are valid, the validation procedure proceeds to the next two criteria, which are applied to all the three peaks.

The CS criterion is used to check whether the *m/z* space between two adjacent peaks in the isotopic cluster is consistent with the charge state, e.g., 0.5 for a charge state of +2. Thus, the CS criterion validation step evaluates (1) whether the isotopic pattern matches the mass of the identified peptide; and (2) whether there is co-elution of other peptides that would interfere with quantitation.

The IP criterion is also used to detect co-elution interference. We compare the normalized intensities of the experimental and theoretical isotopic clusters and calculate the correlation score of the two clusters. If the score is greater than a specified threshold, the spectrum fails in terms of this criterion (see Note 10). The effect of the validation procedure is briefly discussed in Note 11.

3.3.3. Calculating Ion and Peptide Ratios

A PIMS that passes the three criteria is considered acceptable. Based on the validation results, MaXIC-Q classifies the quantitation results as unquantifiable (denoted by N/A), or over-expressed (denoted by 0 or 999), or quantifiable (represented as an expression ratio). If both PIMSs of the light- and heavy-labeled paired ions are unacceptable, MaXIC-Q reports that they are unquantifiable. If one PIMS fails the S/N criterion and the other is acceptable, the ion pair is quantified as over-expressed. If one PIMS passes the S/N criterion but fails the CS or IP criteria and the other PIMS is acceptable, the ion pair is deemed unquantifiable. If both PIMSs are acceptable, the ion pair is quantifiable and MaXIC-Q calculates the ion ratio using the overlapping areas of light- and heavy-labeled paired ions. For the calculation of peptide ratios, the ratio of a peptide is the weighted average of all corresponding quantifiable ion ratios, where the weight of an ion ratio is determined by its area. If a peptide contains no quantifiable ions, the peptide is deemed unquantifiable. If a peptide contains at least one ion with over-expressed ratio, the peptide will probably be of interest to the user. Thus, MaXIC-Q reports both the number of over-expressed ions and the number of quantifiable ions.

3.3.4. Calculating Protein Ratios

MaXIC-Q performs the normalization procedure as described in Subheading 3.2.2.2 on ratios of non degenerate peptides. After normalizing the peptide ratios, protein ratios are calculated based on the weighted average of all corresponding non degenerate peptide ratios.

4. Notes

1. The i-TRACKER program is limited to peptide level quantitation and no protein level analysis is provided. ProQUANT has a peptide ratio normalization function to remove systematic errors resulting from isotope impurity and the experiment process. But ProQUANT is limited to instruments

developed by Applied Biosystems. LIBRA does not have GUI interface and is run on the Linux platform.

2. The pioneering XPRESS tool utilizes signals of the precursor ions to reconstruct XICs. ASAPRatio adopts several numerical and statistical methods, such as Savitzky-Golay smoothing filters for smoothing XICs, Dixon's test for detecting ratio outliers, and error analysis for assessing the quantitation results. In addition, ASAPRatio uses signal-to-noise ratios to filter out unquantifiable XICs before ratio determination. RelEx uses the least-squares regression technique to align paired XICs and determine quantifiable areas of the aligned XICs. The ion abundance ratio is determined by the regression slope, and the maximum correlation coefficient serves as a confidence measure of the quantitation results. MSQuant is based on the Mascot search output and utilizes LC profiles to compute quantitation ratios. Readers may refer to (31, 32) for survey of tools.

3. Multi-Q and MaXIC-Q are designed as stand-alone programs that are executable on the Windows platform.

4. The degree of fluctuation depends on the inherent intensity-based stochastic processes in ion signal measurements by mass spectrometers (21, 33).

5. The three-sigma confidence interval for outlier group has been discussed in details in Kunal Aggarwal et al (34).

6. To filter out errors in peptide ratios caused by limited dynamic range, Multi-Q allows users to input the signal fluctuation threshold and the detector's saturation threshold when running the system. Only peptides with intensity values between the two thresholds are extracted for protein quantitation. Before inputting data into the Multi-Q system, we strongly recommend that first-time users apply a standard protein test with predetermined ratios to examine the saturation effect and signal fluctuation phenomena of their mass spectrometers. Multi-Q can automatically calculate the two thresholds and users can use them for subsequent quantitation analyses.

7. Most popular normalization methods multiply all peptide ratios by a global normalization factor, which can be based on the measured median ratios (35), external reference standard (36), or known invariant reference proteins. For high-throughput quantitative proteomics, a normal distribution is usually chosen to minimize the effect of peptides with low S/N ratios on the normalization process.

8. The selection of m/z window and elution time ranges depends on the LC and MS performance of different instruments.

9. For peak validation by S/N, the default threshold of 2.5 only applies to the first monoisotopic peak in the isotope cluster of

the identified ion, and the S/N criterion for the subsequent two peaks is adjusted in proportion to their theoretical intensities.

10. The default threshold is 0.218, which is pre-determined by applying a data mining tool, called C5.0, to our cICAT-labeling datasets.

11. In some labeling experiments, e.g., ICAT labeling experiments, some proteins may be quantified by only a few peptides. It is hard to detect outliers of peptide ratios for such proteins, whereas our validation criteria can filter out these outliers. For example, in our ICAT-labeling experiment, without validation procedure, the protein IPI00218845 was quantified by three peptides CLGSLVFPR, FCVFGLGSR, and CSQLDHLYR with identification confidence 0.99, 0.96, and 0.83, respectively. These peptides have ratios 0.54, 0.82, and 0.98, respectively, calculated from directly computing the areas of XICs. It is not easy to determine outliers. But, the peptide CLGSLVFPR fails the IP criterion and cannot be quantified. Thus, its ratio 0.54 is filtered out.

Acknowledgments

The authors gratefully acknowledge the financial support from the thematic program of Academia Sinica under Grant AS94B003 and AS95ASIA02 and the National Science Council of Taiwan under Grant NSC 95-3114-P-002-005-Y. We would also like to thank our collaborator Dr. Yu-Ju Chen's lab in the Institute of Chemistry, Academia Sinica. Without their help and encouragement, this research would not have been possible.

References

1. Griffin, T. J., Goodlett, D. R., and Aebersold, R. (2001) Advances in proteome analysis by mass spectrometry. *Curr. Opin. Biotechnol.* **12**, 607–612.
2. Domon, B., and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science* **312**, 212–217.
3. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440.
4. Washburn, M. P., Wolters, D., and Yates, J. R., III. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.
5. Tao, W. A., and Aebersold, R. (2003) Advances in quantitative proteomics via stable isotope tagging and mass spectrometry. *Curr. Opin. Biotechnol.* **14,** 110–118.
6. Semmes, O. J., Malik, G., and Ward, M. (2006) Application of mass spectrometry to the discovery of biomarkers for detection of prostate cancer. *J. Cell. Biochem.* **98**, 496–503.
7. Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904.
8. Ong, S. E., and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1,** 252–262.
9. Islinger, M., Li, K. W., Loos, M., Lueers, G., and Voelkl, A. (2006) ITRAQ-quantification

as an analytical tool to describe proteome changes in rat liver peroxisomes after bezafibrate treatment. *Mol. Cell. Proteomics* **5**, S186.

10. Jabs, W., Lubeck, M., Schweiger-Hufnagel, U., Suckau, D., and Hahner, S. (2006) A comparative study of iTRAQ- and ICPL-based protein quantification. *Mol. Cell. Proteomics* **5**, S248.
11. 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.
12. 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.
13. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (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. 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.
15. Ong, S. E., and Mann, M. (2006) A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **1,** 2650–2660.
16. Ong, S. E., and Mann, M. (2007) Stable isotope labeling by amino acids in cell culture for quantitative proteomics. In: *Quantitative Proteomics by Mass Spectrometry*, Sechi, S., ed., *Methods Mol. Biol.* **359**, 37–52.
17. Callister, S. J., Barry R. C., Adkins, J. N., Johnson, E. T., Qian, W., Webb-Robertson B. M., Smith R. D., and Lipton M. S. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* **5***,* 277–286.
18. Shadforth, I. P., Dunkley, T. P., Lilley, K. S., and Bessant, C. (2005) i-Tracker: for quantitative proteomics using iTRAQ. *BMC Genomics* **6**, 145.
19. 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.
20. 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.
21. MacCoss, M. J., Wu, C. C., Liu, H., Sadygov, R., and Yates, J. R., III. (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal. Chem.* **75,** 6912–6921.
22. Lin, W. T., Hung, W. N., Yian, Y. H., Wu, K. P., Han, C. L., Chen, Y. R., Chen, Y. J., Sung, T. Y., and Hsu, W. L. (2006) Multi-Q: a fully automated tool for multiplexed protein quantitation. *J. Proteome Res.* **5,** 2328–2338.
23. Yu, C. Y., Tsui, Y. H., Yian, Y. H., Sung, T. Y., and Hsu, W. L. (2007) The Multi-Q web server for multiplexed protein quantitation. *Nucleic Acids Res.* **35,** W707–W712.
24. Tsou, C. C, Tsui, Y. H., Yian, Y. H., Chen, Y. J., Yang, H. Y., Yu, C. Y., Lynn, K. S., Chen, Y. J., Sung, T. Y., and Hsu, W. L. (2009) MaXIC-Q Web: a fully automated web service using statistical and computational methods for protein quantitation based on stable isotope labeling and LC-MS. *Nucleic Acids Res.* 37, suppl_2 W661-W669.
25. 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.
26. 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.
27. Weisstein, Eric W. "Moving Average." From MathWorld – A Wolfram Web Resource. http://mathworld.wolfram.com/MovingAverage.html
28. Golub, G. H., Van Loan, C.F. (1996) *Matrix Computations.* 3rd edition, The Johns Hopkins University Press: USA.
29. Savitzky, A, and Marcel J.E. Golay (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639.
30. De Boor, C. (1978) *A Practical Guide to Splines*, 1st ed., pp. 114–115, Springer Verlag, NY.
31. Lau, K. W., Jones, A. R., Swainston, N., Siepen, J.A., and Hubbard, S. J. (2007) Capture and analysis of quantitative proteomic data. *Proteomics* **7**, 2787–2799.
32. Muller, L. N., Brusniak, M.Y., Mani, D. R., and Aebersold, R. (2008). An assessment of software solutions for the analysis of mass

spectrometry based quantitative proteomics data. *J. Proteome Res.* 7, 51–61.

33. MacCoss, M. J., Toth, M. J., Matthews, D. E. (2001) Evaluation and optimization of ion-current ratio measurements by selected-ion-monitoring mass spectrometry. *Anal. Chem.* **73**, 2976–2984.
34. Aggarwal, K., Choe, L. H., Lee, K. H. (2005) Quantitative analysis of protein expression using amine-specific isobaric tags in Escherichia coli cells expressing rhsA elements. *Proteomics* **5**, 2297–2308.
35. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15.
36. Ravin, N. V., and Ravin, V. K. (1999) Use of a linear multicopy vector based on the mini-replicon of temperate coliphage N15 for cloning DNA with abnormal secondary structures. *Nucleic Acids Res.* 27, e13.

Chapter 18

An Overview of Label-Free Quantitation Methods in Proteomics by Mass Spectrometry

Jason W.H. Wong and Gerard Cagney

Abstract

Protein quantification represents an important extension to identification proteomics, enabling the comparison of protein expression across different samples or treatments. Comparative protein quantification by mass spectrometry typically employs stable isotope incorporation, but recently, comparative quantification of label-free LC^n-MS proteomics data has emerged as an alternative approach. In this chapter, we provide an overview of the different approaches for extracting quantitative data from label-free LC^n-MS experiments. The computational procedure for recovering the quantitative information is outlined. Examples of statistical tests used to evaluate the relevance of results are also provided.

Key words: Protein quantification, Mass spectrometry based-proteomics, Label-free quantification, Spectral counting, Ion chromatogram extraction

1. Introduction

Identifying the protein components of samples submitted for analysis to proteomics laboratories is now routine. These procedures typically rely on interfacing liquid chromatographic (LC) separation of peptides (proteolytically digested protein mixtures) with the introduction of the ionized material into a mass spectrometer (1). Meanwhile, orthogonal peptide separation techniques, such as multidimensional protein identification technology (MudPIT) (2–4), have further increased the potential throughput of MS/MS experiments, with studies now regularly reporting the identification of hundreds or thousands of individual proteins.

Merely identifying a protein, however, is often only the first step in MS-based proteomics studies. The ability to quantify the levels of proteins present provides an extra dimension of information,

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_18, © Humana Press, a part of Springer Science+Business Media, LLC 2010

and this information can be critical for time-course or comparative condition-dependent experiments. Unfortunately, the data generated in a typical MS experiment is not directly quantitative. The efficiency of the ionization process depends on several factors, including the molecular composition of each molecule, the ionization method, the type of instrument, and the nature of the instrument duty cycle. For instance, the apparent ion intensity for similar peptides at the same concentration often varies with amino acid composition. Other issues such as LC-MS/MS experimental variation over repeated runs and ion suppression effects (5) may also be confounding. Nevertheless, with careful experimental design and data analysis, comparative and even absolute protein quantification by MS is now possible for researchers.

There are currently two main approaches to protein quantification by LC^n-MS/MS, one involving chemical labeling and the other "label-free" (Fig. 1). The first approach uses incorporation of stable isotopes into one or more of the samples being studied (6). This can be carried out in vivo by stable isotope-containing amino acids introduced in the cell culture media (SILAC) (7, 8), for example, carbon-13 substituted arginine (9). Alternatively, the stable isotopes can be incorporated in vitro by chemical (6, 10, 11) or enzymatic means, for instance using oxygen-18 water when performing proteolysis with trypsin (12). Peptides arising from the sample containing the stable isotope will then be "heavier" when

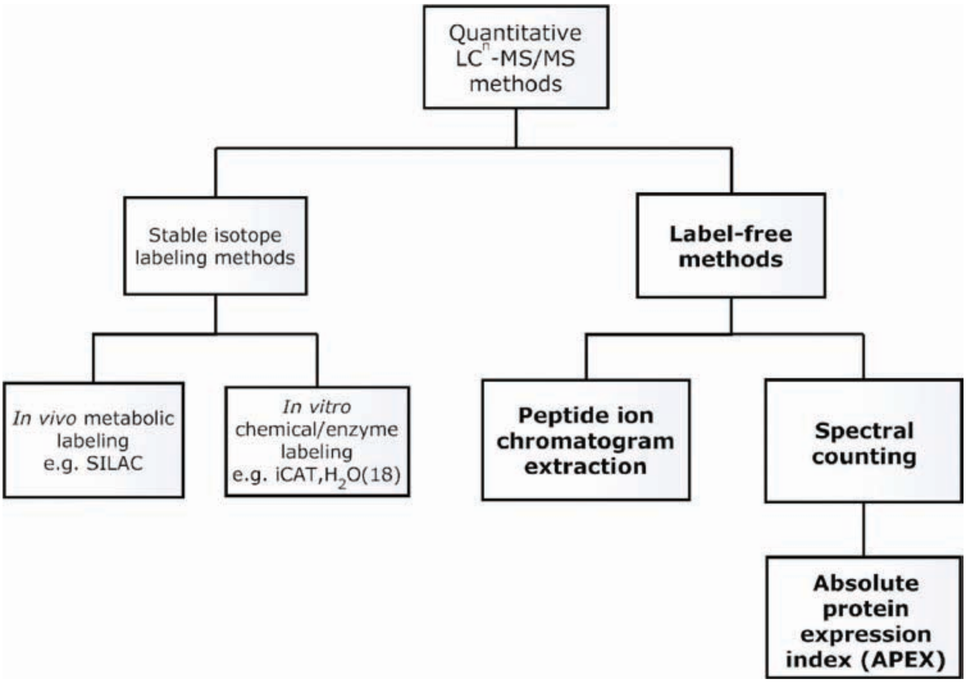

Quantitative LC^n-MS/MS methods
Stable isotope labeling methods
Label-free methods
In vivo metabolic labeling e.g. SILAC
In vitro chemical/enzyme labeling e.g. iCAT,$H_2O(18)$
Peptide ion chromatogram extraction
Spectral counting
Absolute protein expression index (APEX)

Fig. 1. Strategies for comparative proteomics by LC^n-MS. The label-free methods in *bold* are the subject of this review

simultaneously analyzed with the control sample, thereby allowing them to be distinguished in the mass spectrometer. Because the molecular composition of the "heavier" ion is the same as that of the lighter ion, the ionization efficiency will remain the same and therefore the quantities of identical peptides (as represented on an ion chromatograph) can be directly compared. Disadvantages of using stable isotope labeling include the potential for incomplete labeling, and the requirement that cells be culturable (in the case of SILAC). Furthermore, while it is possible to differentially label up to eight biologically different samples using the iTRAQ® Reagent-8Plex kit, the high cost renders routine application prohibitive. In terms of MS data acquisition, isotope labeling may provide an increased challenge because the number of peptides co-eluting will increase, hence possibly reducing the overall peptide coverage. Subsequent computational analysis will also require specific software tools for the recovery of differentially labeled peptides.

The second approach to protein quantification by LC-MS/MS is "label-free". The basis of these methods is to make the assumption that under well controlled conditions with sufficient data redundancy, identical peptides across different LC^n-MS/MS experiments can be compared directly. This has been made possible through technical advances in high-performance (HP) LC systems, mass spectrometers with higher resolution and scanning rates, as well as the use of robots for sample preparation. Semi-quantitative information may be inferred from total peptide ion counts or spectral counts (the number of MS/MS spectra acquired for a protein or peptide). Studies have shown that peptide ion counts across control experiments can be very reproducible (13–16), with results comparable to stable-isotope labeling approaches (17), and label-free comparative quantification studies have gained popularity in recent years (18–25). The major advantages of the label-free approaches are that they typically do not require extra experimental steps and that comparative quantification can be performed across many samples simultaneously. Most recently, it has also been shown that by incorporating information regarding the detectability of a peptide, absolute quantitation is also possible by label-free methods (26, 27).

In this chapter, an overview of procedures for comparative and absolute label-free quantification is provided (see Fig. 2 for schematic overview).

2. Methods

2.1. Experimental Design

As mentioned earlier, a major advantage of label-free quantitation by MS is that no extra steps are typically required compared to a standard MS-based protein identification experiment.

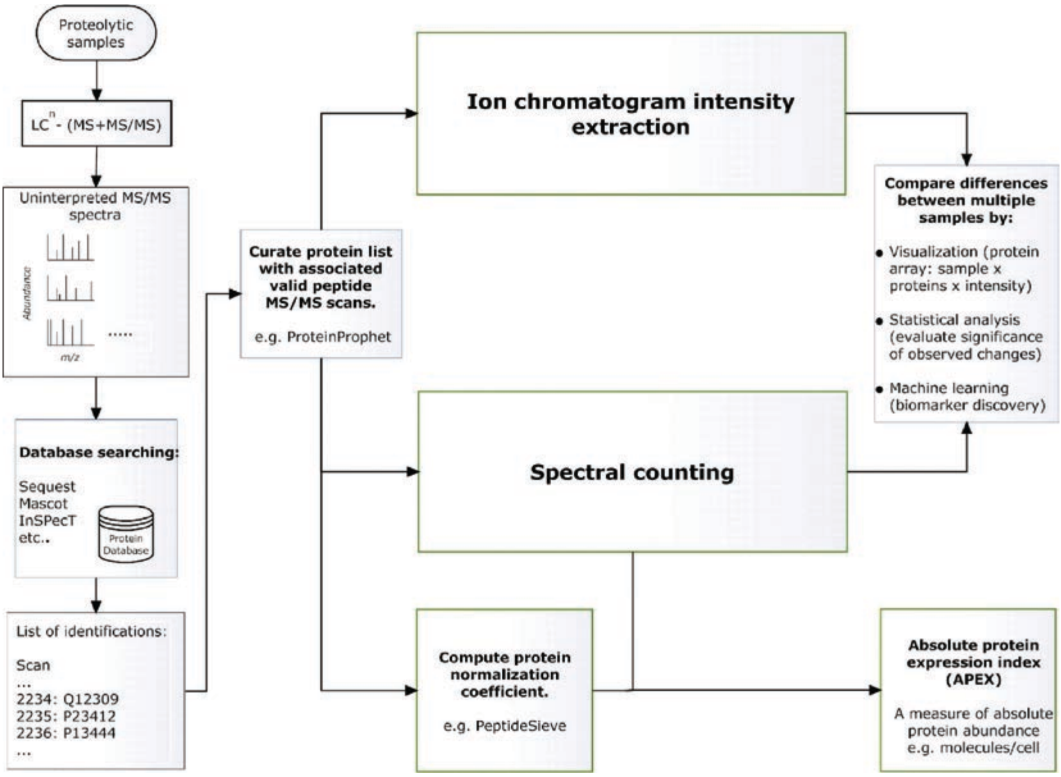

Proteolytic samples
LC^n - (MS+MS/MS)
Uninterpreted MS/MS spectra
Abundance
m/z
.....
Database searching:
Sequest
Mascot
InSPecT
etc..
Protein Database
List of identifications:
Scan
...
2234: Q12309
2235: P23412
2236: P13444
...
Curate protein list with associated valid peptide MS/MS scans.
e.g. ProteinProphet
Ion chromatogram intensity extraction
Spectral counting
Compute protein normalization coefficient.
e.g. PeptideSieve
Compare differences between multiple samples by:
• Visualization (protein array: sample x proteins x intensity)
• Statistical analysis (evaluate significance of observed changes)
• Machine learning (biomarker discovery)
Absolute protein expression index (APEX)
A measure of absolute protein abundance e.g. molecules/cell

Fig. 2. Schematic diagram for label-free quantitative proteomics by LC^n-MS/MS

Nevertheless, in order to ensure the validity of the quantitative information, care must be taken during sample preparation and data acquisition to minimize external influences on the apparent protein quantity.

It is generally advisable to prepare samples to be quantified simultaneously using identical reagents by a single researcher. While LC systems and mass spectrometers are able to produce highly reproducible results, sensitivity can still vary over-time. This could be due to a number of reasons including degradation of the chromatographic column, a change in environmental conditions (e.g., room temperature), as well as regular maintenance, such as tuning of calibration of the mass spectrometer. Therefore, to minimise intra-analysis variations, it is advisable to perform all analysis in a single batch. Furthermore, the analysis of triplicates in tandem repeats (i.e., (A1, B1, C1), (A2, B2, C2), (A3, B3, C3)) will enable any change in LC-MS performance over the course of the experiment to be monitored and to show that any difference in abundance between samples are truly due to intrinsic protein levels as oppose to intra-run variations. Internal controls added to each sample in known concentrations are in principal useful to evaluate experiment-to-experiment variation. In practise however, these controls must be carefully chosen to ensure that they do not interfere with quantitation of the true sample (e.g., via ion suppression). In some cases, signal from experimental reagents present in equal quantities (e.g., trypsin autolysis fragments) can be used to assess reproducibility across samples.

2.2. Data Analysis

2.2.1. Comparative Quantification

2.2.1.1. Ion Intensity

The height or area of a peak at a particular mass-to-charge ratio (m/z) from a mass spectrum reflects the number of ions for that m/z detected by the mass spectrometer at any given time. This is typically known as the ion abundance. Although the ion abundance cannot be used to directly infer absolute protein or peptide concentration (due to different ionization efficiency for each peptide), comparing the ratio of ion abundances between identical peptides obtained in different experiment runs can be used to estimate differential expression.

1. Search all acquired spectra using database searching algorithm (see Note 1).
2. For each peptide identified, extract the ion chromatogram for all charge states (e.g., +1 to +3) (see Note 2).
3. Where necessary, combine the extracted ion chromatogram for each unique peptide from each sample and calculate an abundance ratio across samples (see Note 3).
4. Peptide ratios are then combined to form respective protein abundance ratios (see Note 3).
5. Protein abundance ratios are normalized by the mean of all protein ratios. For complex biological samples, it is expected that the majority of protein ratios to be close to 1:1, assuming that an equal amount of total protein was analyzed. Therefore, to normalize the observed protein ratio, all ratios can be divided by the mean protein ratio.
6. Due to the central limit theorem, it is expected that the distribution of normalized protein ratios be normal and distributed about a mean of 1. The z-score for each protein ratio (i.e., the number of standard deviations a protein ratio is from the mean) can be calculated based on the estimated normal distribution. Comparison of z-scores against a z-table will indicate whether any observed differences in relative protein level is likely to have occurred by chance (see Note 4).

2.2.1.2. Spectral Counting

The spectral count for a protein refers to the number of MS/MS spectra acquired from proteolytic peptide ions for that protein during a LC-MS/MS run. The premise of the method is that the more abundant the peptide, the more likely it will be selected for MS/MS analysis. In controlled experiments, it was found that the correlation of protein abundance with spectral count is superior to that of protein sequence coverage or peptide count (28, 29).

1. Search all acquired spectra using database searching algorithm (see Note 1).
2. For each protein, tally all MS/MS spectra for each peptide belonging to the protein (see Note 5).
3. The total spectral count should be the same across samples under identical data acquisition conditions, so spectral counts for one of the samples should be normalized against the other as follows:

$$s_{1n} = \frac{\sum_i s_{2i}}{\sum_i s_{1i}}$$

where *s* is the spectral count for protein *n*, 1 & 2 are the sample numbers and *i* is the *i*th protein. Where there are more than two samples to normalize, the total spectral count of any one sample can be used as the reference.

4. Determine the statistical significance of protein spectral counts across samples. A number of different statistical tests can be used (see Note 6). For illustrative purposes, the two-sample G-test is shown here, due to its ease in computation and the ability for the test to be generalised for comparisons of more than 2 samples. The spectral counts for each protein can be visualised by a two-way table (Table 1). The G-statistic measure the difference in deviance between the protein spectral counts (generally, the greater the deviance the more likely the difference is significant). The statistic is calculated as follows:

$$G = 2\times\left(\sum_i^2 x_i \ln x_i + \sum_i^2 y_i \ln y_i - \sum_i^2 n_i \ln n_i - x\ln x - y\ln y + n\ln n\right)$$

The William's correction factor (*w*) is applied to adjust the G-statistic for proteins with particularly low spectral counts:

$$w = 1 + \frac{\left(\sum_i^2 \frac{n}{n_i} - 1\right)\left(\frac{n}{x} + \frac{n}{y} - 1\right)}{6n(m-1)}$$

Table 1
Display of spectral count data in a two-way table for statistical testing

	Sample 1	Sample 2	Total counts
Protein *x*	x_1	x_2	*x*
Other proteins	n_1	n_2	*n*
Total	t_1	t_2	*t*

The *p*-value can be calculated from the adjusted G-statistic (G/w) which follows a χ^2 distribution with one degree of freedom ($m-1$ for m samples).

2.2.2. Absolute Quantification

While comparative quantification is adequate when the aim of the experiment is to find differences in protein expression between samples, absolute quantification would be useful for comparing protein levels between data generated at different times by different laboratories using different MS-based proteomics setups. Recently, reaction monitoring techniques that incorporate labeled peptide standards of known concentration have been adapted for proteomics work (30, 31). While these assays are currently expensive and need to be optimized for each protein to be monitored, alternative label-free approaches that aim to provide "absolute" quantitation face a number of hurdles in terms of ensuring that the experiment is accurate, sensitive and reproducible. As mentioned earlier, MS is typically not directly quantitative due to different ionization efficiencies for different peptides. Furthermore, certain peptides may not be retained under particular chromatographic conditions, while the mass of a peptide may simply be out of the range of the mass spectrometer. The end result is that only a limited number of peptides from each protein may be detected within an LC^n-MS/MS experiment. To enable absolute quantification, a measure of the "detectability" of a protein must be defined in order to normalize for the actual observed sampling depth (i.e., spectral counts).

1. Estimate the concentration of protein being analyzed (see Note 7).
2. A normalization coefficient, *O*, is required for each protein. At the simplest, *O* can simply be the number of proteolytic peptides that are within mass detection limits of a mass spectrometer (e.g., 1+ ion > 400 m/z and 3+ ion < 2,000 m/z). However, for a more accurate prediction of the coefficient *O*, Lu and co-workers (26) used machine learning to train a classifier that uses various protein and peptide sequence characteristics to define the probability that a given peptide will be detected. Mallick and co-workers (27) have similarly developed a classifier that also accounts for a number of different shotgun MS proteomics setups. Their classifier (PeptideSieve) is available at tools.proteomecenter.org (see Note 8). To generate the coefficient *O* for all proteins for interest using PeptideSieve, the user inputs a list of protein sequences and selects the relevant experimental design (e.g., MUDPIT_ESI). The output is a list of each tryptic peptide with its respective probability for detection. The coefficient *O* for a particular protein is the sum of all probabilities from tryptic peptides of the protein.

3. Obtain spectral count information as in steps 1 and 2 of subheading “Spectral Counting”.
4. Using the predicted detectability of proteins identified as a normalization factor, estimate the concentration of each protein detected based on spectral counts. Lu and co-workers (26) defined the absolute protein abundance expression (APEX) index, which is essentially molecules of protein per cell. The following formula is use to calculate APEX for protein *i:*

$$\mathrm{APEX}_i = \frac{n_i \times p_i}{O_i \displaystyle\sum_{k=1}^{\text{observed proteins}} \frac{n_k \times p_k}{O_k}} \times C$$

where *n* is spectral count for protein *i*, *p* is the probability of correctly identifying protein *i* (from ProteinProphet (32)), *O* is the normalization coefficient for the “detectability” of the protein and *C* is the estimated total protein concentration.

3. Notes

1. A wide variety of database searching algorithms are available. Sequest (33) and Mascot (34) which are commercial products are generally most popular, however, open source tools such as X!Tandem (35), OMSSA (36) and InsPecT (37) are all broadly comparable in performance.
2. To measure the total ion abundance for any peptide ion within a LC-MS experiment, the ion intensity is integrated over time. This process is computationally referred to as ion extraction resulting in an extracted ion chromatogram (Fig. 3). For technical detail regarding computing an extracted ion chromatogram, see ASAPratio (38).
3. When combining extraction ion intensity ratios to form a unique peptide ratio or combining unique peptide ratios to form protein ratios, each ratio should be weighted accordingly using the actual ion intensity.
4. For a description of a more sophisticated statistical test which takes into account ion intensity errors between peptide ions, refer to ASAPratio (38). The incorporation of errors in the statistical test is particularly useful when the number of overall protein ratio available is low.
5. Tools such as PeptideProphet (32) and ProteinProphet (39) will automatically generate tables wherein spectral counts can be easily extracted. Note that software that directs the MS

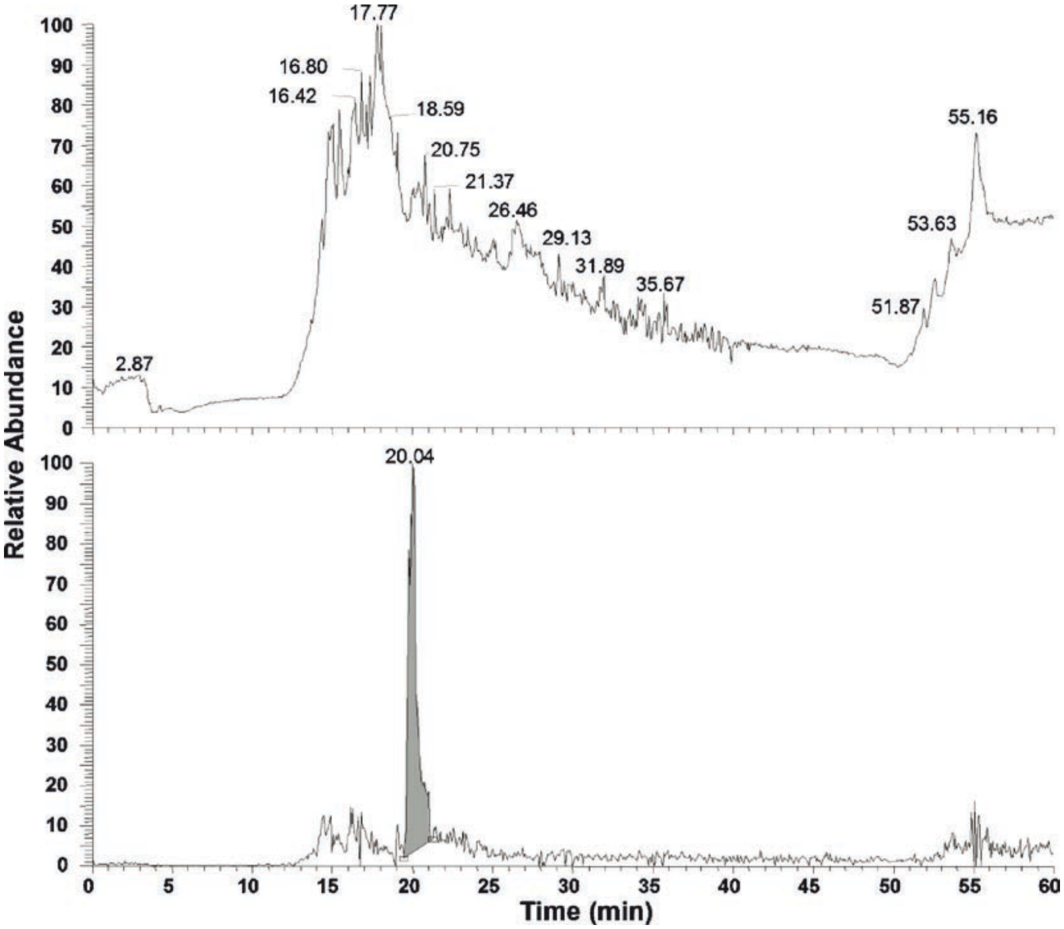


Fig. 3. Example chromatogram of a typical LC^n-MS/MS analysis of a tryptically digested proteome. Peptides were separated on a C18 reverse phase column followed by MS and data dependent MS/MS analysis using a ThermoFinnigan LTQ mass spectrometer. The top shows the total ion chromatogram for the run while the bottom is an extracted ion chromatogram for a particular peptide showing a significant peak. The area of this peak represents the total ion intensity of the peptide

instrument to acquire spectra in a data dependent manner will influence the spectral counts.

6. An evaluation of the performance of five different statistical tests has been performed by Zhang and co-workers (29).
7. Depending on the experimental condition, a number of different methods may be used to estimate protein concentration. For example, the Lowry (40) or Bradford (41) assay can give an estimate for protein concentration in mg/mL.
8. A graphic user interface for PeptideSieve is available at (http://web.bii.a-star.edu.sg/~wongch/peptideSieve/).

References

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics, *Nature* **422**, 198–207.
2. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R. (1999) Direct analysis of protein complexes using mass spectrometry, *Nat. Biotechnol.* **17**, 676–82.
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**, 242–47.
4. Wolters, D. A., Washburn, M. P., and Yates, J. R., III (2001) An automated multidimensional protein identification technology for shotgun proteomics, *Anal. Chem.* **73**, 5683–90.
5. Annesley, T. M. (2003) Ion suppression in mass spectrometry, *Clin. Chem.* **49**, 1041–44.
6. 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**, 994–99.
7. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics, *Mol. Cell. Proteomics* **1**, 376–86.
8. Ong, S. E., Mittler, G., and Mann, M. (2004) Identifying and quantifying in vivo methylation sites by heavy methyl SILAC, *Nat. Methods* **1**, 119–26.
9. Ong, S. E., Kratchmarova, I., and Mann, M. (2003) Properties of ^{13}C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC), *J. Proteome Res.* **2**, 173–81.
10. 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**, 163–70.
11. Ross, P. L., Huang, Y., Marchese, J., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents, *Mol. Cell. Proteomics* **3**, 1154–69.
12. Mirgorodskaya, O., Kozmin, Y., Titov, M., Körner, R., Sönksen, C., 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**, 1226–32.
13. Bondarenko, P., Chelius, D., and Shaler, T. (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry, *Anal. Chem.* **74**, 4741–49.
14. Chelius, D., and Bondarenko, P. (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry, *J. Proteome Res.* **1**, 317–23.
15. Chelius, D., Zhang, T., Wang, G., and Shen, R. (2003) Global protein identification and quantification technology using two-dimensional liquid chromatography nanospray mass spectrometry, *Anal. Chem.* **75**, 6658–65.
16. Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards, *Anal. Chem.* **75**, 4818–26.
17. Zybailov, B., Coleman, M. K., Florens, L., and Washburn, M. P. (2005) Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling, *Anal. Chem.* **77**, 6218–24.
18. Gravett, M. G., Thomas, A., Schneider, K. A., Reddy, A. P., Dasari, S., Jacob, T., Lu, X. F., Rodland, M., Pereira, L., Sadowsky, D. W., Roberts, C. T., Novy, M. J., and Nagalla, S. R. (2007) Proteomic analysis of cervical-vaginal fluid: Identification of novel biomarkers for detection of intra-amniotic infection, *J. Proteome Res.* **6**, 89–96.
19. Wienkoop, S., Larrainzar, E., Niemann, M., Gonzalez, E., Lehmann, U., and Weckwerth, W. (2006) Stable isotope-free quantitative shotgun proteomics combined with sample pattern recognition for rapid diagnositics, *J. Sep. Sci.* **29**, 2793–801.
20. Wang, H. X., Qian, W. J., Chin, M. H., Petyuk, V. A., Barry, R. C., Liu, T., Gritsenko, M. A., Mottaz, H. M., Moore, R. J., Camp, D. G., Khan, A. H., Smith, D. J., and Smith, R. D. (2006) Characterization of the mouse brain proteome using global proteomic analysis complemented with cysteinyl-peptide enrichment, *J. Proteome Res.* **5**, 361–69.
21. Ruth, M. C., Old, W. M., Emrick, M. A., Meyer-Arendt, K., Aveline-Wolf, L. D., Pierce, K. G.,

Mendoza, A. M., Sevinsky, J. R., Hamady, M., Knight, R. D., Resing, K. A., and Ahn, N. G. (2006) Analysis of membrane proteins from human chronic myelogenous leukemia cells: Comparison of extraction methods for multidimensional LC-MS/MS, *J. Proteome Res.* **5**, 709–19.

22. Le Bihan, T., Goh, T., Stewart, II, Salter, A. M., Bukhman, Y. V., Dharsee, M., Ewing, R., and Wisniewski, J. R. (2006) Differential analysis of membrane proteins in mouse fore- and hindbrain using a label-free approach, *J. Proteome Res.* **5**, 2701–10.

23. Kislinger, T., Cox, B., Kannan, A., Chung, C., Hu, P. Z., Ignatchenko, A., Scott, M. S., Gramolini, A. O., Morris, Q., Hallett, M. T., Rossant, J., Hughes, T. R., Frey, B., and Emili, A. (2006) Global survey of organ and organelle protein expression in mouse: Combined proteomic and transcriptomic profiling, *Cell* **125**, 173–86.

24. Fang, R. H., Elias, D. A., Monroe, M. E., Shen, Y. F., McIntosh, M., Wang, P., Goddard, C. D., Callister, S. J., Moore, R. J., Gorby, Y. A., Adkins, J. N., Fredrickson, J. K., Lipton, M. S., and Smith, R. D. (2006) Differential label-free quantitative proteomic analysis of Shewanella oneidensis cultured under aerobic and suboxic conditions by accurate mass and time tag approach, *Mol. Cell. Proteomics* **5**, 714–25.

25. Cao, R., Li, X. W., Liu, Z., Peng, X., Hu, W. J., Wang, X. C., Chen, P., Xie, J. Y., and Liang, S. P. (2006) Integration of a two-phase partition method into proteomics research on rat liver plasma membrane proteins, *J. Proteome Res.* **5**, 634–42.

26. Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation, *Nat. Biotechnol.* **25**, 117–24.

27. Mallick, P., Schirle, M., Chen, S., Flory, M., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B., and Aebersold, R. (2007) Computational prediction of proteotypic peptides for quantitative proteomics, *Nat. Biotechnol.* **25**, 125–31.

28. Liu, H. B., Sadygov, R. G., and Yates, J. R. (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics, *Anal. Chem.* **76**, 4193–201.

29. Zhang, B., VerBerkmoes, N., Langston, M., Uberbacher, E., Hettich, R., and Samatova, N. F. (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics, *J. Proteome Res.* **5**, 2909–18.

30. Anderson, L., and Hunter, C. L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins, *Mol. Cell. Proteomics* **5**, 573–88.

31. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS, *Proc. Natl. Acad. Sci. U. S. A.* **100**, 6940–45.

32. 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**, 5383–92.

33. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectra data of peptides with amino acid sequences in a protein database, *J. Am. Soc. Mass Spectrom.* **5**, 976.

34. 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**, 3551–67.

35. Fenyo, D. (1999) The biopolymer markup language, *Bioinformatics* **15**, 339–40.

36. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X. Y., Shi, W. Y., and Bryant, S. H. (2004) Open mass spectrometry search algorithm, *J. Proteome Res.* **3**, 958–64.

37. Tanner, S., Shu, H. J., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra, *Anal. Chem.* **77**, 4626–39.

38. 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**, 6648–57.

39. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry, *Anal. Chem.* **75**, 4646–58.

40. Lowry, O. H., Rosebrough, N. J., Farr, A. L., and Randall, R. J. (1951) Protein measurement with the folin phenol reagent, *J. Biol. Chem.* **193**, 265–75.

41. Bradford, M. M. (1976) Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principle of protein-dye binding, *Anal. Biochem.* **72**, 248–54.

Chapter 19

The PeptideAtlas Project

Eric W. Deutsch

Abstract

PeptideAtlas is a multi-species compendium of peptides observed with tandem mass spectrometry methods. Raw mass spectrometer output files are collected from the community and reprocessed through a uniform analysis and validation pipeline that continues to advance. The results are loaded into a database and the information derived from the raw data is returned to the community via several web-based data exploration tools. The PeptideAtlas resource is useful for experiment planning, improving genome annotation, and other data mining projects. PeptideAtlas has become especially useful for planning targeted proteomics experiments.

Key words: Proteomics, Data repository, Proteome, Database, SRM

1. Introduction

The advent of tandem mass spectrometry (MS/MS) has enabled the identification of a large number of proteins in a high throughput manner. A wide variety of instruments, sample preparation techniques, and data analysis methods have fostered an innovative research community, and a huge amount of data has been and continues to be generated at significant expense. It has long been recognized that public repositories of data would accelerate the advancement of proteomics (1) as it has done for other fields such as transcriptomics. Making the data easily accessible to the public fosters the validation of results, and more importantly the reuse of the data for purposes beyond the intents of the original researchers.

Making the raw mass spectrometer output files accessible to the community is important because the analysis techniques of proteomics continue to advance markedly over time. Modern analysis of older datasets yields many more identifications and information from the data due to better protein reference information and better

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_19, © Humana Press, a part of Springer Science+Business Media, LLC 2010

informatics software. Indeed, the newest spectral library-searching techniques routinely identify 50% more spectra than sequence-searching techniques, and different search engines are able to identify different peptides in the same datasets. One can expect that future workflows will apply several tools in parallel to achieve an analysis much closer to optimal.

The PeptideAtlas Project is a resource that accepts mass spectrometer output files in a variety of formats along with the metadata associated with the experiment. The raw data are reanalyzed using ever-improving techniques and coalesced into a compendium of identifications for each species. An important part of the resource is the tools that allow the research community to access the data in the PeptideAtlas database for experiment planning, validation of new datasets, and other data mining projects.

In addition to PeptideAtlas, several repositories for proteomics data have emerged over the last few years, including PRIDE (Proteomics Identifications Database) (2), OPD (Open Proteomics Database) (1), Tranche (3), and GPMDB (Global Proteome Machine Database) (4). These repositories have different strengths and fill different niches. The strengths of PeptideAtlas are that only raw data are accepted and are processed through a uniform analysis and validation pipeline to insure high quality results with well-understood false discovery rates (FDR), and an advanced toolset for presenting the results in a manner conducive to experiment planning.

In the following sections, the PeptideAtlas Resource is described in detail. First a brief history of the early motivations and work is presented, followed by a description of the building of PeptideAtlas. Finally, the many ways to use the PeptideAtlas is presented, ending with an outlook on the future of the resource.

2. History

With the increasing number of installed tandem mass spectrometers capable of generating large amount of MS/MS-based proteomics data, it became apparent that there was significant value in collecting and combining many of these datasets. Expected benefits from such work include high coverage of a proteome, sufficient data density for statistical arguments, and the possibility to contribute extensive observational data back to genome annotation projects.

The PeptideAtlas project thus began at the Seattle Proteomics Center as a compendium of peptides observed in a collection of human and *Drosophila* shotgun tandem mass spectrometry datasets acquired at the Institute for Systems Biology. Also available were the annotations describing in which samples the peptides and proteins were observed, which modified forms and how frequently the peptides were observed, and how these peptides mapped onto the genome (5).

Subsequently, additional builds have been added for yeast (6), *Streptococcus pyogenes* (7), and *Halobacterium salinarum* (8). In addition, several specialized builds for subproteomes were released for human plasma (9) and mouse plasma (10) samples. PeptideAtlas builds for several other species (mouse, *E. coli*, rat) and subproteomes (liver, pancreas, et al.) are expected to be released in 2009.

The tools have also evolved considerably in the past four years. In 2004, only basic query and browsing tools were available. As of this writing, there are a large number of tools that support new targeted proteomics strategies as well improvements to traditional approaches.

3. Building of the PeptideAtlas

The build process of the PeptideAtlas has evolved since it was initially described (11). As illustrated in Fig. 1, raw mass spectrometer output files for MS/MS experiments are collected from

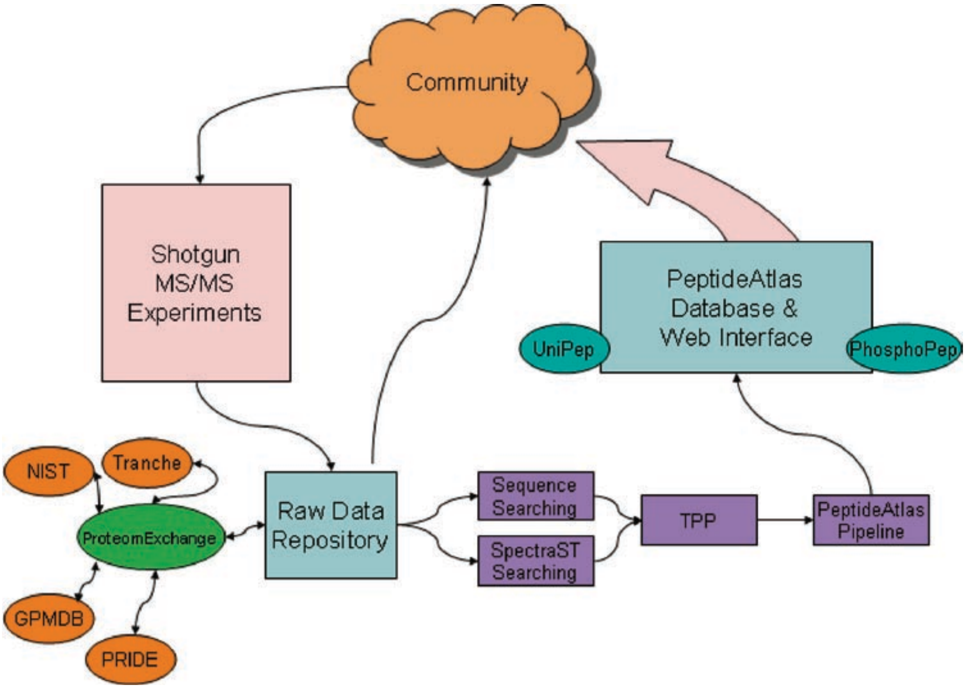


Fig. 1. An overview of the build process of PeptideAtlas. Shotgun tandem mass spectrometry (MS/MS) experimental data are contributed by the community to the PeptideAtlas raw data repository, which is linked to other repositories via the ProteomExchange consortium. The raw data are processed through an evolving but consistent analysis and validation pipeline (Trans Proteomic Pipeline (TPP)) and loaded into the PeptideAtlas database, and made available to the community. Tranche, GPMDB (Global Proteome Machine Database), NIST (National Institute of Standards and Technology), and PRIDE (Protein Identifications Database) are the current major participants in the ProteomExchange consortium

the community, processed through a consistent analysis pipeline, and then loaded into the PeptideAtlas database, thereby returning high-value information back to the community that provided the data. These different phases are described in great detail in the following subsections.

3.1. Acquiring Data and the Raw Data Repository

A key component of PeptideAtlas is a data repository in which raw data and search results are made available to the community. The PeptideAtlas data repository has had an important role in the advancement of research using high throughput technologies, acting as the data provider to several projects, including the spectrum library building at the National Institute of Standards and Technology (NIST), the PepSeeker database (12), as well as large-scale genome annotation efforts (13). In addition to PeptideAtlas, several repositories for proteomics data have emerged over the last few years, including PRIDE, OPD, Tranche, and GPMDB. These repositories have different strengths and fill different niches, but it is obvious that the highest benefit can be gained if all the repositories share data and metadata to allow users to access data from all the same experiments using the repository that best meets their requirements. PeptideAtlas is actively participating in the formation of the ProteomExchange consortium that attempts to facilitate this interoperability between the repositories.

However, most of the aforementioned repositories are largely passive – that is, results are stored and can be queried or downloaded, but the remaining untapped potential within the primary data is not extracted with continually advancing analysis tools. Typically, only a small fraction of acquired MS/MS spectra are confidently identified in the first attempt. Although many of the unidentified spectra are of inadequate quality to be ever identified, a considerable fraction of them can indeed be identified with more effort and newer techniques (14). PeptideAtlas aims to be an active repository, in which only raw data are accepted and these raw data are periodically reprocessed with more advanced techniques for identification and statistical validation as these are becoming available. The results of this advancing analysis of the raw data are then made available back to the community in forms that enable additional research, specifically with tools that support the new targeted proteomics workflows.

3.2. Uniform Processing with Advanced Tools

Once raw mass spectrometer output files are available in the raw data repository, sequence database searching and automated validation of the results using the Trans Proteomic Pipeline (TPP) (15) is performed. This begins with conversion to a common mzXML file format, then sequence searching with either SEQUEST (16) or X!Tandem (17), followed by validation of the top hits with PeptideProphet (18), a program that models the correct and incorrect spectrum-peptide match populations and assigns a probability of being correct to each match.

All PeptideProphet results are then combined using ProteinProphet (19), a program that uses the spectrum-peptide match models from PeptideProphet to derive protein-level probabilities as well as to adjust the peptide-level probabilities based on the information available from the ensemble of experiments. Given a set of confidently identified spectra, the spectral library building tool SpectraST is used to create a consensus spectrum library comprising all observed peptide ions. As part of the library building process, many high scoring but incorrect identifications are rejected. Then, all raw data are subjected to a second round of searching, this time by spectral library searching with SpectraST. This has the effect of identifying many more spectra from the available data, with a higher sensitivity and lower error (20). Output of SpectraST is validated in the same manner as described above with PeptideProphet and ProteinProphet.

3.3. Populating the Database

All peptides are then mapped to a single reference Ensembl (21) build (if available for the species) and mapped to the genome. All this information is loaded into the PeptideAtlas database for browsing or downloading.

The information is loaded in as a discrete build within the PeptideAtlas database. A build represents a particular set of experiments that have been processed as described above at a certain point in time, and mapped to a specific build of the proteome/genome. This build version remains static thereafter. As additional data are acquired, old data reprocessed, or mappings to newer proteome/genome builds are performed, a new build becomes the default, but older builds remain available for comparison or historical reference.

The result of each build process is also made publicly available at the PeptideAtlas web site in several formats. The front-end web-site software is distributed as part of the Systems Biology Experiment Analysis System (SBEAMS) framework (22). A summary of the current state of the various PeptideAtlas builds is provided in Table 1.

4. Using PeptideAtlas

A crucial aspect to the success of the PeptideAtlas Project is the tools available for accessing the information therein. The following subsections highlight some of the most visible and useful features of the PeptideAtlas.

Table 1
Summary of public PeptideAtlas builds

Build	# Exps	# MS runs	Searched spectra	IDs $P > 0.9$	Distinct peptides	Distinct proteins
Human all	219	54 k	49 M	5.6 M	97 k	12,141
Human plasma	76	48 k	16 M	1.8 M	18 k	2,486
Drosophila	43	1,769	7.5 M	498 k	72 k	9,124
Drosophila PhosphoPep	4	448	0.9 M	170 k	10 k	4,583
Yeast	53	2,957	6.5 M	1.1 M	36 k	4,336
Mouse	59	3,097	10 M	1.4 M	51 k	7,686
Halobacterium	88	497	0.5 M	76 k	12 k	1,518
S. pyogenes	5	64	215 k	52 k	7 k	1,068

4.1. Build Overviews

As described above, a build represents a particular set of experiments that have been processed as described above at a certain point in time, and mapped to a specific build of the proteome/genome. For each build, there is a summary page that provides such information as the build date, the number of experiments included, the number of spectra searched to create the build, and the resulting number of identifications.

This is followed by some tables and charts that demonstrate the individual contributions of the experiments to the build. Experiments are usually listed in approximate chronological addition to the PeptideAtlas, and therefore the charts track the growth of the atlas build over time.

4.2. Protein Views

For each protein in the reference proteome for a given build, a dynamic protein view page summarizes the information available for that protein. The page is segmented into several collapsible sections that can be easily minimized when they are not of relevance to the user. Minimized sections persist over multiple page views.

The top section provides basic information about the protein including all the aliases and related names and accessions available in the database, as well as the total number of spectra and distinct peptides that map to the protein.

The following two sections summarize the peptide coverage of the protein. A graphical diagram, similar to the genome browser views, summarizes all the peptides that map either uniquely or redundantly to the proteins plus the information on segments unlikely to be observed with mass spectrometers, as well as the signal peptides and transmembrane information where available.

The actual protein sequence is displayed with amino acids occurring in observed peptides highlighted.

This is followed by a section listing all the peptides observed and mapping to this protein. The table listing includes many attributes of the peptides, including the number of times they were observed with what best probabilities, theoretically calculated hydrophobicities, and the samples in which the peptides were observed. Empirical Observability Score (EOS) and Suitability score metrics are listed as well. The EOS reflects a likelihood that if the protein is detectable in the sample, it is detected via that peptide. The Suitability Score represents a ranking of how suitable the peptide is as a reference or proteotypic peptide. The score includes information about the total number of observations, the EOS, the best probability of identification, and includes penalties if the peptides are not fully tryptic, contain missed cleavages, or undesirable residues that impact a peptides suitability for targeting (such as methionine which is variably oxidized).

Below this is a section about theoretical peptides for the protein. Each protein is digested *in silico* and both the PeptideSieve (23) and DetectabilityPredictor (24) software tools are used to predict which peptides might be most suitable for targeting. This can be compared with the empirical evidence for many proteins. For low abundance or otherwise hard-to-detect proteins, these theoretical predictions are useful.

Finally, the last section provides a summary of the samples in which the protein was observed. This is also potentially quite useful for planning future experiments.

4.3. Peptide Views

For each peptide observed in the data for a given build, a dynamic peptide view page summarizes the information available for that peptide. The page is segmented into several collapsible sections as described above. The first section provides a number of attributes for the peptide including predicted hydrophobicity and pI, as well as the number of spectra supporting the identifications.

The following section diplays the peptide-to-protein and chromosomal mapping information. Since the peptide-to-protein mapping can be multiplex and confusing, this section tries to simplify the mapping information. If the peptide can map to multiple isoforms of the same gene, this is noted, and when a peptide spans an intron in the genome, the chromosomal coordinates reflect this. In order to better visualize a complex mapping relationship, a hyperlink to a secondary page displays all the proteins to which a peptide maps; the proteins are aligned together with an overlap of which peptides are observed for each isoform or different protein.

The next section lists all of the different observed peptide ions, i.e. the different charge states or mass modifications that were observed. For each peptide ion, the predicted monoisotopic precursor *m/z* is listed along with the number of observations, number of experiments, and hyperlinks to visualize the consensus spectra for each peptide ion.

Below this is a listing of every spectrum that supports the identification of this peptide, along with individual attributes of the identifications such as probability of being correctly identified. Each individual spectrum is available for viewing.

Finally, at the bottom is a listing of all the experiments that included the peptide along with some simple charts that depict the relative number of spectral counts in each of the experiments.

4.4. Queries

The previously described peptide and protein views are useful for exploring proteomes, one protein at a time. However, they are impractical for extracting lists of interesting results for many proteins and peptides. For this reason, there are several query pages that can return many peptides or proteins at once. These pages allow users to specify a list of constraints for the desired output and receive a list of either proteins, peptides, or transitions (see Subheading 4.6) based on the specified constraints. The lists may be browsed interactively via the embedded hyperlinks, or downloaded in XML or tab-separated-value formats, or even right into tools like Excel.

4.5. Proteotypic Peptides

For targeted proteomics strategies, it is important to determine which peptides are the optimal ones to target; these are termed proteotypic peptides (25). A proteotypic peptide is one that is easily observable with current mass spectrometry technology and one that maps uniquely to a single protein or isoform. Such peptides make optimum targets and PeptideAtlas provides tools that make the extraction of such proteotypic peptides easy via the query form described above.

The relationships between proteins and constituent peptides can be quite complex in higher eukaryotes and difficult to grasp using ordinary tabular views. We therefore provide a mechanism to visualize peptide and proteins within a PeptideAtlas build using the Cytoscape network visualization software (26). On the protein view web page, below the list of constituent peptides, there is a button to launch Cytoscape. The information on the current page, including the protein and peptides is combined into a Cytoscape-compatible format. Proteins and peptides are nodes in the network; peptides that map to the protein are connected with an edge. Additionally, the network is then grown to include all proteins and peptides that have any relationship with the peptides or proteins already in the network. This final dataset is then packaged up in a jar file and sent to the client with the application via Java Web Start. The user sees a new Java window appear as shown and further described in Fig. 2.

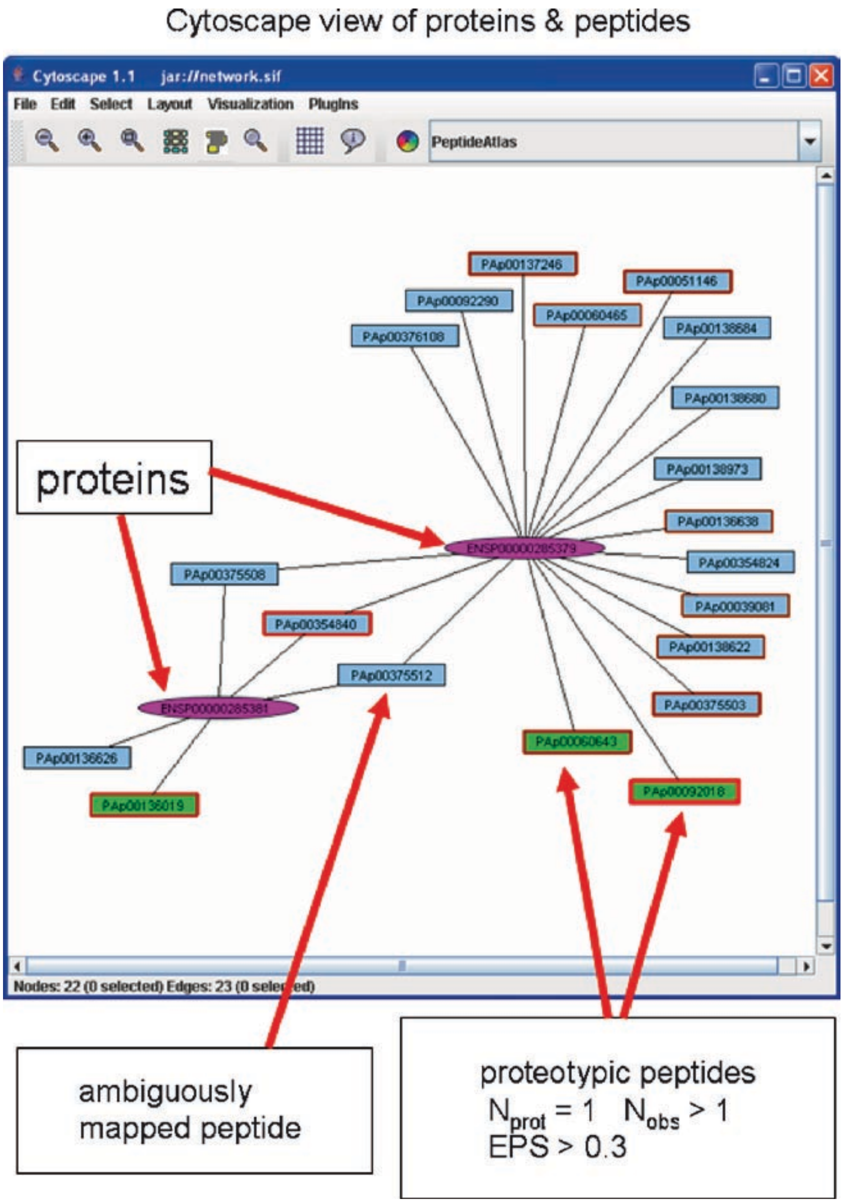


Fig. 2. Cytoscape visualization of a simple set of proteins and peptides. The two proteins are drawn as purple oval nodes. Peptides are drawn as rectangular nodes. Edges indicate the mapping of peptides to proteins. Peptides that have only one edge are uniquely mapping; peptides with two or more edges are ambiguously mapped. Peptide rectangle borders become thicker and redder with greater numbers of observations. Proteotypic peptides (uniquely mapping, multiply observed, and having EOS > 0.3) are shaded in green.

4.6. Selecting Transitions

The emerging targeted proteomics workflows such as selected reaction monitoring (SRM; also called MRM) are gaining popularity. In this workflow, the mass spectrometer is configured to monitor unique ion signatures, called transitions, of predetermined peptides in order to achieve a detection or confident upper limit for desired peptides to the exclusion of all other peptides. A transition is merely the set of precursor *m/z* and one or more product ion *m/z* values. However, the selection of appropriate transitions can be a difficult task. PeptideAtlas has several tools to aid in the selection of transitions, which are the signatures of peptides needed for SRM (27). As described above, the individual and consensus spectra are all available within the PeptideAtlas interface and can be used to select transitions either by hand or in batch queries. The ViewMRMList query allows users to specify a list of input proteins and the desired attributes of the transitions, and the result is a tab-separated-value list of candidate transitions for follow up.

One problem with predicting candidate transitions from PeptideAtlas is that most of the spectra in the atlas are from ion trap instruments, simply because that is what is predominantly submitted. However, the relative intensities of the fragment ions in triple quadrupole mass spectrometers, the one typically used for SRM, can be quite different from that of ion traps, and thus the predicted transitions do need to be validated. However, a special build of the PeptideAtlas, called MRMAtlas, is built using MS/MS spectra only from triple-quad instruments. Only a relatively small number of such spectra are available in the MRMAtlas, and only for certain species. However, the data that are contained therein provide the best available transitions for the proteins represented in these special builds.

4.7. PeptideAtlas Results in Other Resources

Besides the interfaces at http://www.peptideatlas.org described so far, the builds from the PeptideAtlas can be accessed via several other sites on the World Wide Web. At the Ensembl genome browser site, one can overlay PeptideAtlas peptides onto the genome exploration interface by selecting PeptideAtlas in the DAS (distributed annotation server) sources section. Indeed, PeptideAtlas builds are available as DAS sources on our DAS server, and therefore any application that can access genome annotation information via DAS can access PeptideAtlas builds.

Also, we have adapted our interfaces so that they may be easily indexed by the very popular Google search engine. If one performs a Google search for any of the peptides contained in the public PeptideAtlas builds, the top hit will usually be to a PeptideAtlas page that summarizes attributes of the desired peptides, including in which builds the peptide occurs.

As a final example, the iSPIDER resource (28) allows its users to search for proteomics identifications across multiple proteomics databases including PeptideAtlas. When a protein name or accession is entered into iSPIDER, it dynamically queries several repositories, including PeptideAtlas, and summarizes the results for the user.

5. Conclusion

This chapter has provided an overview of the PeptideAtlas proteomics data resource and repository, including a description of its history, the build process, and the many tools that can be used to access the information in PeptideAtlas. Although it has many uses from improving genome annotation to complex data mining projects, PeptideAtlas is also a very valuable resource for the design of experiments for emerging targeted proteomics workflows. Work is underway to make PeptideAtlas an even more valuable resource for SRM experiments. Users will soon be able to shop for the best available transitions for their favorite list of proteins based on the various data types in PeptideAtlas, including community-submitted validated transitions, transitions based on MRMAtlas observations, transitions based on the main PeptideAtlas builds, and finally if insufficient information is available from the previous sources, transitions will be predicted based on the best available theoretical prediction software.

PeptideAtlas is designed as an engine to turn the community's data into information that everyone can use to enable future work. It relies critically on the availability of raw data, which is now starting to become common. As better and more extensive datasets are processed through PeptideAtlas with ever-improving analysis tools, the resource will serve everyone designing future proteomics experiments.

Acknowledgments

The PeptideAtlas Project has involved a great many contributors. The author would like to thank the following for their contributions to the design and implementation of PeptideAtlas: Dave Campbell, Nichole King, Luis Mendoza, David Shteynberg, Natalie Tasman, Abhishek Pratap, Pat Moss, Jimmy Eng, Ning Zhang, Frank Desiere, Terry Farrah, Zhi Sun, Michael Johnson, and Ruedi Aebersold.

The author has been funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179, and from PM50 GMO76547/Center for Systems Biology.

References

1. Prince, J.T., et al., *The need for a public proteomics repository.* Nat Biotechnol, 2004. **22**: p. 471–72.
2. Martens, L., et al., *PRIDE: the proteomics identifications database.* Proteomics, 2005. **5**(13): p. 3537–45.
3. Falkner, J.A. and P.C. Andrews, *Tranche: secure decentralized data storage for the proteomics community.* J Biomol Tech, 2007. **18**(1): p. 3.
4. Craig, R., J.P. Cortens, and R.C. Beavis, *Open source system for analyzing, validating, and storing protein identification data.* J Proteome Res, 2004. **3**(6): p. 1234–42.
5. Desiere, F., et al., *Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry.* Genome Biol, 2004. **6**(1): p. R9.
6. King, N.L., et al., *Analysis of the Saccharomyces cerevisiae proteome with PeptideAtlas.* Genome Biol, 2006. 7(11): p. R106.
7. Lange, V., et al., *Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring.* Mol Cell Proteomics, 2008. 7(8): p. 1489–500.
8. Van, P.T., et al., *Halobacterium salinarum NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage.* J Proteome Res, 2008. 7(9): p. 3755–64.
9. Deutsch, E.W., et al., *Human Plasma PeptideAtlas.* Proteomics, 2005. **5**(13): p. 3497–500.
10. Zhang, Q., et al., *A mouse plasma peptide atlas as a resource for disease proteomics.* Genome Biol, 2008. **9**(6): p. R93.
11. Desiere, F., et al., *The PeptideAtlas project.* Nucleic Acids Res, 2006. **34**(Database issue): p. D655–8.
12. McLaughlin, T., et al., *PepSeeker: a database of proteome peptide identifications for investigating fragmentation patterns.* Nucleic Acids Res, 2006. **34**(Database issue): p. D649–54.
13. Tanner, S., et al., *Improving gene annotation using peptide mass spectrometry.* Genome Res, 2007. **17**(2): p. 231–9.
14. Nesvizhskii, A.I., et al., *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, 2006. **5**(4): p. 652–70.
15. Keller, A., et al., *A uniform proteomics MS/MS analysis platform utilizing open XML file formats.* Mol Syst Biol, 2005. **1**: p. 2005 0017.
16. Eng, J., A.L. McCormack, and J.R. Yates, *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.* J Am Soc Mass Spectrom, 1994. **5**: p. 976–89.
17. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466–7.
18. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Anal Chem, 2002. **74**: p. 5383–92.
19. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry.* Anal Chem, 2003. **75**: p. 4646–58.
20. Lam, H., et al., *Development and validation of a spectral library searching method for peptide identification from MS/MS.* Proteomics, 2007. 7(5): p. 655–67.
21. Hubbard, T.J., et al., *Ensembl 2007.* Nucleic Acids Res, 2007. **35**(Database issue): p. D610–7.
22. Marzolf, B., et al., *SBEAMS-Microarray: database software supporting genomic expression analyses for systems biology.* BMC Bioinformatics, 2006. **7**: p. 286.
23. Mallick, P., et al., *Computational prediction of proteotypic peptides for quantitative proteomics.* Nat Biotechnol, 2007. **25**(1): p. 125–31.
24. Tang, H., et al., *A computational approach toward label-free protein quantification using predicted peptide detectability.* Bioinformatics, 2006. **22**(14): p. e481–8.
25. Kuster, B., et al., *Scoring proteomes with proteotypic peptide probes.* Nat Rev Mol Cell Biol, 2005. **6**(7): p. 577–83.
26. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498–504.
27. Deutsch, E.W., H. Lam, and R. Aebersold, *PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows.* EMBO Rep, 2008. **9**(5): p. 429–34.
28. Siepen, J.A., et al., *ISPIDER Central: an integrated database web-server for proteomics.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W485–90.

Chapter 20

Using the PRIDE Proteomics Identifications Database for Knowledge Discovery and Data Analysis

Philip Jones and Lennart Martens

Abstract

The PRIDE Proteomics Identifications Database provides users with the ability to explore and compare mass spectrometry-based proteomics experiments that reveal details of the protein expression found in a broad range of taxonomic groups, tissues and disease states. A PRIDE experiment typically includes identifications of proteins, peptides and protein modifications. Many of the submitted experiments also include processed peak lists representing the mass spectra that provide the evidence for these identifications.

Since the inception of the PRIDE project, a number of tools supporting submission of data to PRIDE have been developed. Of particular note is the "PRIDE Converter" that has become the tool most frequently used for the production of PRIDE submissions at the time of writing.

The PRIDE XML format has been expanded to provide submitters with the capacity to annotate fragment ion information on to peptide identifications and the fragmentation spectra that provide the experimental evidence for these peptides. A novel algorithm for annotating fragment ion information on to peptides and their evidential mass spectra has also been developed that will ultimately provide a route for evaluating the quality of peptide identifications arising from tandem mass spectrometry. This algorithm allows the visualisation of potential fragment ions on to the identified mass spectra, even where no such information has been submitted.

In this chapter, we describe how PRIDE can be applied as a research tool and how the experiments in PRIDE can be compared and analysed. We also explore how complex queries can be constructed using the PRIDE BioMart. Finally, we will describe how the user can integrate PRIDE data with annotation from other resources, using federated BioMart queries.

Key words: Proteomics identifications database, Fragment ion annotation, Mass spectrometry, Data repository, Data analysis, Data set comparison, BioMart, Distributed annotation system

1. Introduction

The PRIDE Proteomics Identifications Database (1, 2) is a repository for the data and results derived from mass spectrometry-based proteomics experiments, which makes use of public data

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_20, © Humana Press, a part of Springer Science+Business Media, LLC 2010

standards, allowing data from a vast range of instruments and analysis software platforms to be submitted. PRIDE presents this data through a variety of different interfaces, including a bespoke web interface, a BioMart (3) interface for advanced query and a BioMart web service, to provide programmatic access to the data. Detailed data from individual experiments can be obtained from the EBI FTP service (ftp://ftp.ebi.ac.uk/pub/databases/pride/) as compressed XML files.

During 2008, the PRIDE database has developed from its original role as a repository of proteomics identifications arising from mass spectrometry to a database providing tools for complex query and data retrieval, dataset comparison and access to additional automated annotation of submitted datasets. Here are described the methods used to access and exploit some of these new tools together with the web-service capability currently provided by PRIDE.

PRIDE comprises a repository of identifications of proteins, peptides and protein modifications. To support these identifications the mass spectra arising from a wide range of experimental techniques are included. Almost all of the data in PRIDE is submitted by the experimentalist rather than being manually curated into PRIDE. No assumptions are made about the sample, instrumentation or protocol used to generate mass spectra, or the data analysis applied to this data. This presents two challenges. First of all, to allow useful query of the data, a consistent mechanism for annotation of the experiments is required. To allow annotation of the sample under investigation, the instrumentation used for separation and mass spectrometry and the software infrastructure used for identifying proteins and peptides, PRIDE mandates the use of specific controlled vocabularies or ontologies. For example, sample species information is annotated using NCBI taxonomy identifiers (4, 5) and where appropriate, the tissue is annotated using the BRENDA tissue ontology. To annotate the sample, various other ontologies are recommended to indicate sub-cellular location, disease state and cell type. The PRIDE software then makes use of the Ontology Lookup Service (6) (http://www.ebi.ac.uk/ols) to allow powerful, hierarchical query of the data.

A previous publication describing PRIDE in the Humana Press focused upon the mechanisms provided in PRIDE for basic query and data submission (7).

Here, we will describe the use of more recent additions to the PRIDE database and user interface including fragment ion annotation, together with a description of how to make the best use of the BioMart “MartView” user interface.

2. Submitting Data to the PRIDE Database

For the first 2 years of PRIDE as a production service, it was relatively difficult to produce valid PRIDE XML files and submit them to the database. The submission process usually involved writing software, often based upon the PRIDE Java API. For laboratories without sufficient bioinformatics or software engineering support, this often proved impossible without seeking direct programming support from the PRIDE team at the EBI. More recently, the hurdles to submission have been alleviated by the development of several powerful tools.

2.1. The PRIDE Proteome Harvest Data Submission Spreadsheet

The PRIDE Proteome Harvest Data Submission Spreadsheet was the first such tool, which uses a special Microsoft Excel spreadsheet to allow proteomics data producers to build PRIDE XML files without resorting to programming. Even this tool has limitations however, as it does not directly support the inclusion of mass spectra or peak lists in the generated XML, other than by allowing an mzData XML file to be embedded in the resulting PRIDE XML file.

2.2. Pride Wizard

The second tool to be developed in support of PRIDE XML submission was the "Pride Wizard" (http://www.mcisb.org/resources/PrideWizard/index.html), developed at the University of Manchester (8). The development of this tool was part of a larger effort to support the reporting of iTRAQ™data, allowing PRIDE XML to encode quantitative results from multiple samples. The tool itself allows the conversion of several spectral data formats including .mgf, mzXML, .pkl and the HUPO PSI mzData format, together with the output from the Mascot search engine (dat files) into PRIDE XML. This may include the intensities of iTRAQ™ reporter ions. In addition, the tool can be used to automatically capture the output of non-quantitative, large-scale experiments in PRIDE XML format. This tool is freely available under the GNU GPL License.

2.3. The PRIDE Converter

A fruitful collaboration with the University of Bergen, Norway, led to the development of the "PRIDE Converter" (http://code.google.com/p/pride-converter), the most complete solution so far, developed by Harald Barsnes. This tool allows the researcher to convert the output from several makes of mass spectrometer and also the output from multiple proteomics search engines into valid PRIDE XML. The PRIDE Converter tool also incorporates a client to the Ontology Lookup Service (OLS, http://www.ebi.ac.uk/ols) allowing the user to look up suitable terms to annotate all aspects of a PRIDE Experiment that cannot be imported directly from these proprietary data formats, such as details of the sample and the mass spectrometer components. This tool has already proven very popular. At the time of writing, approximately 40% of submissions to PRIDE are generated using this tool. The PRIDE Converter has been developed under the permissive Apache 2 open source license. It should be noted that laboratories that wish to develop reusable pipelines for generating PRIDE XML will find the source code of the PRIDE Converter a valuable resource for solving common conversion problems, together with the core PRIDE Java API, which is available from http://code.google.com/p/ebi-pride.

The timings of these developments supporting data submission are particularly fortuitous as the proteomics journals apply increasing pressure on proteomics researchers to publish their experimental data in public repositories (9–12). Indeed the journal "Proteomics" now mandates this requirement in their instructions to authors: "In particular, novel protein sequences should be deposited in UniProt (www.uniprot.org); molecular interactions in an IMEx partner database (imex.sf.net); and protein identification data in PRIDE (www.ebi.ac.uk/pride), World-2DPAGE (www.expasy.org/world-2dpage/), or a comparable database."

3. Fragment Ion Annotation in PRIDE

Peptide identification usually results from the analysis of fragmentation spectra, which are obtained by fragmenting a selected precursor ion. These spectra, also called MS/MS or MS2 spectra, are valuable as they contain peptide sequence information. The individual ions in the fragmentation spectrum that correspond to parts of the peptide sequence are called fragment ions.

In collaboration with Waters Corporation (http://www.waters.com), the PRIDE team has developed a simple mechanism to allow fragment ion information to be annotated on to peptide identifications in PRIDE XML. These fragment ion annotations are associated with the mass spectrum that has provided evidence for the peptide identification (note that in the PRIDE XML model, one peptide identification is associated with at most one mass spectrum). PRIDE fragment ion annotation can be derived from two sources: the annotation can be submitted by the experimentalist or it can be derived from the automatic annotation of fragment ions as generated by a specialised algorithm developed as part of the PRIDE project.

This algorithm performs two consecutive processing steps. In the first step, it matches the theoretical peptide mass calculated from the peptide sequence in PRIDE, and compares it to the experimental precursor mass annotated in the mass spectrum. If there is a mismatch between the experimental and theoretical precursor mass, this is typically caused by unreported modifications on the peptide that were introduced during the sample processing protocol. In these cases, the algorithm works out what the missing modifications are, based on the mass difference between theoretical and experimental mass and knowledge about the experimental protocol. Once the theoretical and experimental precursor masses have been reconciled, the algorithm can simply predict the theoretical fragment ions and locate them in the fragmentation spectrum.

These fragment ion annotations can be viewed on the PRIDE web interface, using the PRIDE spectrum viewer. The following sequence exemplifies this, using data from the PRIDE experiment with accession number 3, “COFRADIC N-terminal proteome of unstimulated human blood platelets”:

1. Navigate to the PRIDE Home Page at http://www.ebi.ac.uk/pride.
2. Enter a search term into the “Search PRIDE:” text box at the top left hand side of the window and click on the red “Search” button (see Note 1 for an explanation of suitable search terms). As an example, you may choose to enter “3” to search for the PRIDE experiment with accession number 3, “COFRADIC N-terminal proteome of unstimulated human blood platelets”, which includes automatic fragment ion annotation.
3. You will be presented with the “Search Summary View”. If no results match your search, you will be informed, otherwise you will be presented with a table of all of the PRIDE experiments that match your search, with one row per experiment. These rows contain some summary statistics for each experiment.
4. To progress on to view individual mass spectra for one of the experiments in the list, click on the red “View” button for the experiment of interest. You will then be taken to a more detailed top-level summary of the experiment, on the “Experiment View” page. This page includes details including references in the published literature, contact information, details of the sample that was examined, the protocol followed to prepare the sample and towards the bottom, links to view protein identifications (and the corresponding peptide identifications) and mass spectra.
5. The most direct way to access the mass spectrum view for a single mass spectrum is to click on the red “View Spectrum Details” on the “Experiment View” page. You will then be taken to a list of the mass spectra available for the experiment. The complete list is separated into pages with 30 mass spectra per page. You can move through the entire list using the links

found at the top and bottom of the list on each page. Click on one of the numbers in the “Spectrum ID” column (the first column) to view the details of a single mass spectrum.

6. You will now be taken to the “Spectrum Detail View”. At the top of this view is information associated with the mass spectrum, including a list of peptides identified from the spectrum, if the spectrum led to identifications. Scrolling further down, you will find a view of the mass spectrum itself, not including fragment ion annotations.

7. Back at the top of the page, you may see a red “View Automatically Annotated Fragment Ions” button, if the automatic annotation algorithm has been successfully run against the experiment. Alternatively, you may see a red “View Submitted Fragment Ions” button, if the experimentalist has submitted fragment ion information. Clicking either of these buttons will present these annotations on the mass spectrum, as illustrated in Fig. 1. Obviously, if the spectrum was not identified, or no fragment

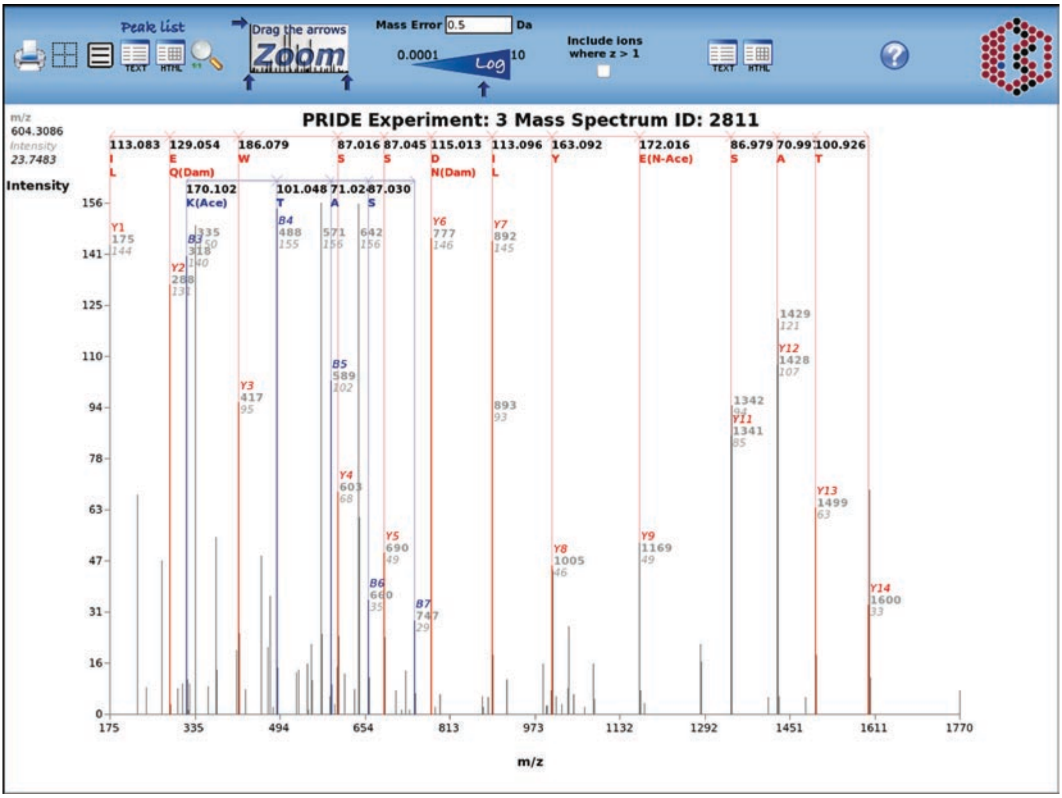


Fig. 1. This figure illustrates the mass spectrum that has been used to identify the peptide AGMKTASGDYIDSSWELR. Automatically annotated fragment ions are included on this mass spectrum view. It can be seen that very good coverage of the Y series fragment ions and some B series fragment ions are annotated, corresponding to the identified peptide sequence

ions are as yet annotated for an identification, you will not be able to see these buttons.

8. You can hide the annotations again by clicking on the red "Hide Spectrum Annotation" button at the top of the "Spectrum Detail View" page.

4. The PRIDE BioMart

The original web interface to PRIDE allows queries based upon a small number of specific parameters, including experiment accession number, protein identifier, peptide sequence, literature reference and sample parameters. All of these types of search result in a standard display of experimental information that can be used to browse the details of the experiment. Very often however, the form of these results will not be well tailored to the requirements of the user. For example, the user may require nothing more than a list of protein accession numbers for proteins identified in a particular tissue. This type of result set was difficult to acquire from the PRIDE web interface.

Fortunately, a robust solution to this problem has been provided in the form of the PRIDE BioMart.

BioMart (http://www.biomart.org) is a query-oriented data management system developed jointly by the Ontario Institute for Cancer Research (http://www.oicr.on.ca/) and the European Bioinformatics Institute (http://www.ebi.ac.uk). The BioMart data management system provides many benefits: fast query across large data sets (due to the standardised query-orientated database structure that it specifies); the ability to define complex queries, both in terms of the filters applied to the data and to the selection of data items (attributes) to be reported; the ability to select from a number of different data output formats, including HTML, a Microsoft Excel spreadsheet and plain text formats; programmatic access to data via both a REST and a SOAP web service and based upon this service and the ability to integrate data from separate BioMarts in seamless queries.

The PRIDE BioMart (http://www.ebi.ac.uk/pride/prideMart.do) provides access to the details of identifications of proteins, peptides and post-translational modifications in all of the public datasets accessible from PRIDE. Mass spectra providing evidence for protein and peptide identifications are referenced and linked from the BioMart, allowing the details of the mass spectra to be viewed and accessed in the "standard" PRIDE web interface. The BioMart database is updated on a regular basis to keep it synchronised with the core PRIDE database.

4.1. Performing a Search of the PRIDE BioMart, Using the MartView Interface

The MartView BioMart interface allows the user to build complex queries interactively, by defining "filters" (to restrict the rows of data returned) and "attributes" (the data items that should be returned, equivalent to the columns in a spreadsheet). The user is then able to preview the results of her search, allowing her to modify the search as necessary. Finally, the user can select the desired format of the results and request them, either for immediate viewing or via an email link to receive notification when the results are ready. This process is described in detail below.

1. Access http://www.ebi.ac.uk/pride/prideMart.do (see Note 2 for alternative URL).
2. Choose the database to access, by selecting one of the items in the "-CHOOSE DATABASE-" pull down list. At the time of writing, the PRIDE BioMart provides federated access to the Reactome (www.reactome.org) BioMart, allowing PRIDE data to be viewed alongside Reactome data. You can select either PRIDE BioMart or Reactome (CSHL) from the list of databases first. The rest of these instructions assume however that you have selected the PRIDE database at this step. See Note 3 for more details on federated data access.
3. A second pull down list will appear below the database pull down list. In the case of the PRIDE BioMart, this will default to the single dataset defined in the PRIDE BioMart, "PRIDE". If, however, the list contains the value "-CHOOSE DATASET-", you will need to select the required dataset from the list. Some more complex BioMarts provide more than one dataset. See Note 4 for an explanation.
4. To build a filter, click on the "Filters" heading in the left panel of the BioMart interface. You can then define your filter on the right hand panel.
5. You will see several expandable sections from which you can select filter criteria (you can select any number and combination of filters from all the sections). To open a section, click on the + symbol.
6. Note that for some of the filters it is possible to specify more than one item. The filter then returns all records that match any of the items specified (using OR logic). For fields that accept a typed, pasted or uploaded list, you can separate values using either white space or commas. You may also upload a text file from your computer containing information formatted in the same way by clicking on the browse button. For fields that contain a list of possible values, you can select multiple values by holding the CTRL key (Microsoft Windows and Linux) or Apple and Shift key (Apple OS X) while clicking on the values that you wish to select.

7. To select attributes, click on "Attributes" in the left panel of the BioMart window.
8. Expand the section that interests you by clicking on the + symbol and then check the check-boxes adjacent to each attribute that you wish to include in the results. *Note that the order in which you check the check boxes will determine the order of the columns in the results table.*
9. If you click on the "Count" button at the top of the BioMart interface, you will be presented with a count of the number of PRIDE experiments that match your query see Note 5.
10. Clicking on the "Results" button will return the first ten result rows in HTML format. At this point, you may wish to modify your filters or the selected attributes. If you are happy with the example results, you can proceed to the next step.
11. To select the format for the results, you should click on the pull down list that defaults to "TSV" (tab separated values file). Here, you can select from TSV, CSV (comma separated values file – another plain text format), HTML (well formatted for the browser) or XLS (Microsoft Excel spreadsheet).
12. If you expect a large number of rows to be returned from the BioMart, you are advised to select the "Compressed web file (Notify by email)" option, adjacent to "Export all results to". Then enter you email address in the text box labelled "Email notification to". You will receive an email containing a hyperlink from which you can download a compressed results file.
13. Click on the GO button. If you have not requested the results by email, they will be displayed in the requested format after a short delay (see Note 6).

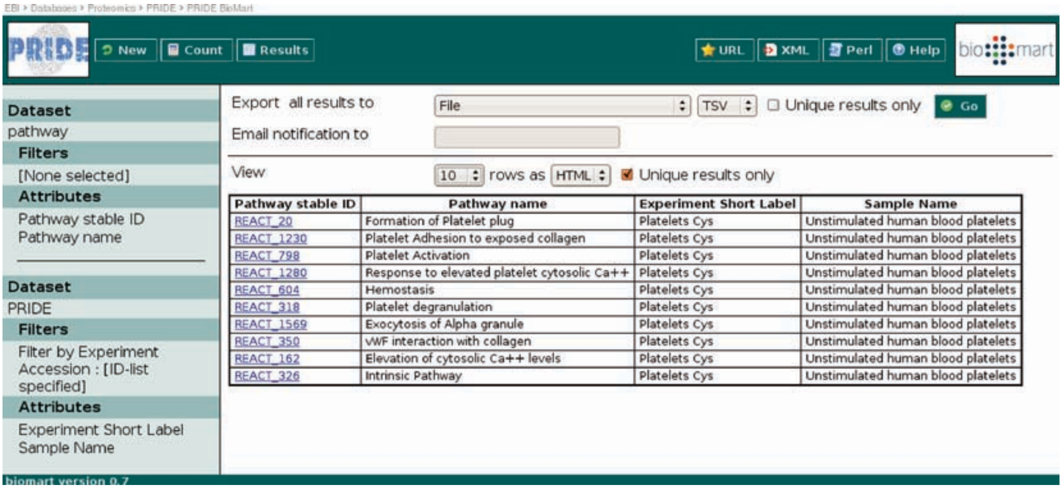
5. Notes

1. The simple search box on the PRIDE home page should be used for the following types of search. Note that for more complex searches, you are advised to use the PRIDE BioMart.
   - PRIDE Experiment accession number. These values are plain integers.
   - Protein accession number. You can use most common protein database identifier types which are mapped to the submitted accession number using the Protein Identifier Cross Reference (PICR) service at the EBI, which was developed by the PRIDE team.
2. In the event that the http://www.ebi.ac.uk/pride/prideMart.do URL is not available, for example in the case of scheduled

downtime of the main PRIDE web interface for maintenance, you can access the PRIDE BioMart directly at http://www.ebi.ac.uk/pride/biomart/martview/.

3. The PRIDE BioMart allows federated access to the Reactome BioMart, at the time of writing, although plans are in place to expand the number of federated databases connected to PRIDE. The link between PRIDE and Reactome is defined through the presence of common UniProtKB protein accession numbers. On the PRIDE side, these accession numbers are attached to protein identifications and on the Reactome side, these protein accession numbers are attached to proteins annotated in biological pathways. As a consequence, it is possible to (for example) view data on specific protein identifications in PRIDE alongside information about their metabolic activity as defined in Reactome. Inversely it is possible to view all identifications stored in PRIDE for a given Reactome pathway. As an illustration of the power of federated queries, *see* Fig. 2, which illustrates a BioMart query including data from PRIDE for a sample of human platelet tissue, linked to the pathway information for the identified proteins from Reactome (limited to the first ten results). This example illustrates a strong correlation between the sample tissue and the biological pathways on which the identified proteins have been annotated in Reactome.

4. BioMart allows separate "datasets" to be defined for one BioMart database. PRIDE keeps this very simple, comprising only a single dataset, "PRIDE" that contains all of the public data available from PRIDE. Other BioMarts split their data



Dataset
pathway
Filters
[None selected]
Attributes
Pathway stable ID
Pathway name

Dataset
PRIDE
Filters
Filter by Experiment Accession : [ID-list specified]
Attributes
Experiment Short Label
Sample Name

Export all results to File TSV □ Unique results only Go

Email notification to

View 10 rows as HTML ☑ Unique results only

Pathway stable ID	Pathway name	Experiment Short Label	Sample Name
REACT_20	Formation of Platelet plug	Platelets Cys	Unstimulated human blood platelets
REACT_1230	Platelet Adhesion to exposed collagen	Platelets Cys	Unstimulated human blood platelets
REACT_798	Platelet Activation	Platelets Cys	Unstimulated human blood platelets
REACT_1280	Response to elevated platelet cytosolic Ca++	Platelets Cys	Unstimulated human blood platelets
REACT_604	Hemostasis	Platelets Cys	Unstimulated human blood platelets
REACT_318	Platelet degranulation	Platelets Cys	Unstimulated human blood platelets
REACT_1569	Exocytosis of Alpha granule	Platelets Cys	Unstimulated human blood platelets
REACT_350	VWF interaction with collagen	Platelets Cys	Unstimulated human blood platelets
REACT_162	Elevation of cytosolic Ca++ levels	Platelets Cys	Unstimulated human blood platelets
REACT_326	Intrinsic Pathway	Platelets Cys	Unstimulated human blood platelets

biomart version 0.7

Fig. 2. An illustration of BioMart federated query, including linked results from both PRIDE and the Reactome database of biological pathways. This example illustrates a strong correlation between the sample tissue and the biological pathways on which the identified proteins have been annotated in Reactome

in various manners. The Reactome BioMart is split into three datasets: “complex”, “pathway” and “reaction”. The Ensembl BioMart is split into taxonomic groups.

5. The “Count” button on the BioMart interface returns the number of PRIDE Experiments that match your query. This is not necessarily the same as the number of results (rows of data) that will be returned. Depending upon the attributes that you have selected, the number of results may be several orders of magnitude greater than the stated count.

6. If there is a long delay after you have clicked GO, this would suggest that your query will result in many rows of data being returned. This may crash your browser on arrival, so it is recommended that you click on the back button and modify your results request to be sent by email.

References

1. Martens, L., Hermjakob, H., Jones, P. et al. (2005) Pride: the proteomics identifications database. *Proteomics* **5**, 3537–45.
2. Jones, P., Côté, R.G., Cho, S.Y. et al. (2008) Pride: new developments and new datasets. *Nucleic Acids Res* **36**, D878–83.
3. Kasprzyk, A., Keefe, D., Smedley, D. et al. (2004) Ensmart: a generic system for fast and flexible access to biological data. *Genome Res* **14**, 160–9.
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J. et al. (2000) Genbank. *Nucleic Acids Res* **28**, 15–18.
5. Wheeler, DL., Chappey, C., Lash, AE. et al. (2000) Database resources of the national center for biotechnology information. *Nucleic Acids Res* **28**, 10–14.
6. Côté, R.G., Jones, P., Martens, L. et al. (2008) The ontology lookup service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res* **36**, W372–6.
7. Jones, P. and Côté, R. (2008) The pride proteomics identifications database: data submission, query, and dataset comparison. *Methods Mol Biol* **484**, 287–303.
8. Siepen, J.A., Swainston, N., Jones, A.R. et al. (2007) An informatic pipeline for the data capture and submission of quantitative proteomic data using itraq. *Proteome Sci* **5**, 4.
9. (2007) Democratizing proteomics data. *Nat Biotechnol* **25**, 262.
10. (2007) Time for leadership. *Nat Biotechnol* **25**, 821.
11. (2007) Mind the technology gap. *Nat Methods* **4**, 765.
12. (2008) Thou shalt share your data. *Nat Methods* **5**, 209.

Chapter 21

Molecular Interactions and Data Standardisation

Sandra Orchard and Samuel Kerrien

Abstract

Molecular interactions are crucial components of the cellular process. In order to understand this complex machinery, one needs to gather published data from various sources. Many projects have initiated the collection of interaction data for this purpose since 2002. However, the lack of standardisation previously made the task of aggregating datasets difficult. This issue has been resolved by the creation of Molecular Interaction standard in 2004 by members of the Proteomics Standards Initiative (PSI), a work group of the Human Proteome Organization (HUPO). Furthermore, major database providers have come together with the goal to exchange data in order to optimise laborious curation tasks. Finally, tools and frameworks have been created based on PSI-MI standards to facilitate the visualisation and analysis of molecular interaction data.

Key words: Molecular Interactions, Data standardisation, PSI-MI XML, PSIMITAB, MIMIx, IMEx

1. Introduction

Protein–protein interaction databases have existed for many years, with early contributors to the field, such as BIND (1) and DIP (2), being joined by other databases containing interactions from the entire spectra of the literature, such as IntAct (3), organism-specific resources such as BioGrid (4) and HPRD (5), and repositories which fulfil specialist biological interests, for example the MINT (6) database that concentrates on the identification of interacting domains of proteins and MatrixDB (http://matrixdb.ibcp.fr), which looks at the molecular interactions made by extracellular proteins. The scope of these databases has increased over time, with the degree of annotation captured becoming richer and some repositories, such as IntAct, beginning to capture all possible interactions within a cell or organism, including protein-small molecule and protein-nucleic acid.

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_21, © Humana Press, a part of Springer Science+Business Media, LLC 2010

By 2002, the user was served by an ever-increasing number of data resources, all of which collected interaction information from the literature, and in some cases, from direct submissions from research workers, but an individual wishing to download the information from a number of sources was faced with parsing multiple, separately constructed databases, each with its own individual structure and data format. Merging the data into a single repository then required further effort, and it could be problematic merely identifying those papers, which had been redundantly curated by more than one database. It was at this point that several of these resources were brought together by the Human Proteome Organisation to tackle these problems and provide an improved service for the user.

2. HUPO and the Proteomics Standards Initiative

The Human Proteome Organisation (HUPO) was formed in 2001 to consolidate national and regional proteome organisations into a single worldwide body (7). The Proteome Standards Initiative (PSI) was established by HUPO with the remit of standardising data representation 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 utilised by laboratory workers (8). The Proteomics Standards Initiative (HUPO-PSI) has concentrated on bringing data standardisation and common data reporting standards to an increasing number of fields within the global umbrella of Proteomics; to date, protein/peptide separations, mass spectrometry and molecular interactions. Following the initial meeting of the HUPO-PSI in 2002, the work has been progressed by a series of workshops, interspersed by regular phone conferences and active mailing lists and web usage by groups of volunteers from all over the world.

Each workgroup within the HUPO-PSI has produced a series of documents and resources to aid in the process of data standardisation and exchange. Minimum Information About a Proteomics Experiment (MIAPE) documents have been developed (9), analogous to the MIAME (Minimum Information About a Microarray Experiment) guidelines for DNA microarray experiments (10), to define those data items that should minimally be reported about a proteomics experiment to allow critical assessment of the experiment. This is a simple textual representation, independent of any formal data format. MIAPE guidelines consist of a general "parent document" (9) and a series of workgroup-specific modules (for example, Refs (11–13)). These guidelines summarise what could be considered "common sense" and what the community has agreed should be present in each and every paper, but is all too often not appropriately reported in publications – the precise identification of a protein entity and the species from which it originated being a simple example of data often missing from articles. To facilitate data management and exchange, each domain area has also developed data exchange formats for which it can at least minimally represent the data items specified in the MIAPE guidelines, but usually additionally allow a much more detailed representation. Normally, the data exchange format is specified as a fully annotated XML schema. HUPO-PSI schemas are developed to facilitate data exchange between databases as well as databases and end users. They explicitly do not propose any internal data representation for databases or tools. XML is well supported by standard mechanisms for querying, native XML databases, and automated mappings to both relational databases and object models, all of which has been taken advantage of in the development of user-friendly tools and services.

The semantics of data elements exchanged are described by a series of controlled vocabularies, either by referencing external resources such as the NCBI taxonomy or developed internally by the HUPO-PSI, for example to describe the details of mass spectrometry or molecular interactions. The combination of reasonably stable XML schemas and regularly maintained controlled vocabularies allows a quick adaptation to new terms and technologies, while providing the stability required for database and software development.

3. The Molecular Interaction Data Exchange Format

In 2004, the Molecular Interaction workgroup published Level 1.0 of the PSI-MI XML interchange schema, with accompanying controlled vocabularies, jointly developed by major producers of protein interaction data and by data providers including BIND, DIP, IntAct, MINT, MIPS and Hybrigenics (14). The PSI-MI format was explicitly intended to develop in an incremental fashion. Version 1.0 focused exclusively on protein interactions, and was widely implemented and supported by both software tool development and data providers. As a direct result of requests from users, database groups and data providers, the original PSI-MI format was considerably extended, resulting in version PSI-MI XML2.5 (15). The range of interactor types that could be described within the format was extended to encompass all biomolecules and the description that can be made of both experimental conditions and experimental features on participating molecules such as the description of purification tags or deletion or point mutations was considerably enhanced and made more flexible.

The abilities to describe kinetic as well as modelled interaction parameters were also added.

Controlled vocabularies (CVs) are used throughout the PSI-MI schema to standardise the meaning of data objects. Their use ensures that the same term used throughout a description by a data producer, instead of a synonym or alternative spelling, and also that the interpretation of the meaning of that term remains consistent between multiple data producers and users. In order to achieve this, all terms have definitions and, where appropriate, are supported by one or more literature references. The controlled vocabularies have a hierarchical structure, in the form of a direct acyclic graph (DAG), higher level terms being more general than lower level descriptors, allowing annotation to be performed to an appropriate level of granularity whilst also enabling search tools to return all mapped objects to both parent and child terms, if required. As with all HUPO-PSI maintained CVs, the molecular interaction CV is made available on the OBO website (www.obofoundry.org) and is actively maintained by an editorial panel responding to user requests.

The PSI-MI XML2.5 format allows a detailed representation of fully annotated interaction records both for inter-database and database end-user data communication. However, to support many use cases, such as fast Perl parsing or loading into Microsoft Excel, that only require a simple, tabular format of interaction records, the MITAB2.5 format was defined as part of PSI-MI 2.5. The MITAB2.5 format only describes binary interactions, one pair of interactors per row in a simple tab-delimitated format.

Almost all major interaction data producers now make data available in PSI-XML2.5 format and many also in MITAB2.5.

4. The Minimum Information About a Molecular Interaction Experiment (MIMIx)

The MIMIx guidelines provide a checklist for anyone preparing interaction data, be it as little as a single interaction within a paper describing the characterisation of a protein, for either publication in a peer-reviewed article, deposition in an interaction database or displaying a large dataset on a website (11). MIMIx represents a compromise between the depth of information necessary to describe all relevant aspects of an interaction experiment and the reporting burden placed on scientists who generate the data. The MIMIx guidelines were assembled by a large number of experts and subjected to public review both on the HUPO-PSI website and through a community review process organised by *Nature Biotechnology*. At all stages, input from the molecular interaction community, which included data producers, data providers and tool and software developers, has been discussed and fed back into the document. Additionally, these guidelines will not be static. They will evolve based on community requirements in the context of a rapidly developing science but it is hoped that they will contribute to an increase in the standard of the reporting of molecular interactions, much of which has previously been incomplete or even misleading.

It is of increasing importance that databases are maintained to, at least, MIMIx compatibility, as increasingly, the tools and services described below are being written on the assumption that this minimum level of information be supplied. For example, the R statistics package (16) is compromised if databases have not included information on interaction directionality (e.g. bait–prey relationships), which is a MIMIx requirement. In order to comply to this, the fundamental information that the databases need to access i.e. the peer-reviewed written articles published by journals, need to contain this information for the databases to extract or, preferably, the authors should submit the information directly to the databases immediately prior to submission to ensure that the information held in these databases is as near totally correct as possible. The MIMIx guidelines are written to assist authors in achieving compliance and encouraging early database deposition of the information.

5. Tool Development Based on the PSI-MI Data Interchange Standard

The development and maintenance of tools to enable the use of these formats by a wide number of users is a community effort to which many people have contributed. Many have been developed jointly with other resources, allowing the interaction community easy access to new resources to further analyse datasets. All tools are freely available and can be accessed and downloaded from the Molecular Interaction workgroup HUPO-PSI web pages (www.psidev.info). These include applications to view and validate the use of the schema, to enable graphical representation of interaction network, to convert between data formats and to facilitate the use of controlled vocabularies. All tools are specific to the PSI-MI XML format apart from the validator.

5.1. The Semantic Validator

As standards become more commonly used, it of increasing importance to ensure that they are adhered to, and are not compromised by either mis- or re-interpretation by an ever-increasing user community. To aid in this, a semantic validator has been developed which allows:

- The verification of correct CVs usage,
- Complex integrity checking of the data.

The framework is free, open-source and can be adapted to any data format.

The Semantic Validator aims at checking a user-submitted data model against predefined mapping rules (http://psidev.info/validator). The framework is composed of multiple components allowing a developer to easily build a custom validator instance for a specific data model. Provided with a set of corresponding rules, this validator instance can process input data (of the specific data type) and return a set of validation messages, each of which reports on inconsistencies found. The message gives an error level denoting the severity of the problem, a description of the problem, a context that should allow the user to locate the erroneous data and the rule that generated the message. The mapping rules ensure the appropriate usage of CV and ontology terms. These rules are solely based on an XML mapping provided by the user that defines which terms are allowed in specific locations of the data model. The locations in the model are defined using the XPath language and ontology terms are specified by their identifier or name.

Current implementations of the validator allow users to check on Molecular Interactions (PSI-MI) as well as Mass Spectra and Protein Identification (PSI-MS and PSI-PI). The framework is free, open-source and can be adapted to any data format. To that end, a tutorial has been made available so one can implement its own validator. More information can be found on the HUPO-PSI web site: http://psidev.info/validator.

5.2. PSI-MI XML View

XML is a powerful means by which to model complex data whilst preserving human readability. However, due to the complexity of the PSI-MI schema an easier way to visualise the data has also been provided. XSLT scripts have been made available in order to convert XML data files to HTML, thus providing user-friendly representation of the data.

5.3. Conversion Between Expanded and Compact form of the Schema

The PSI-MI XML2.5 schema allows two different representations of interactions – compact and expanded. In the compact form, the repetitive elements of a large set of interactions, namely copyright statements, experiment descriptions, and interactors (proteins, small molecules, etc), are only described once, in the respective list elements. The use of the compact form is appropriate for larger datasets, where, for example, one protein would be referred to by multiple interaction elements. The expanded form is more suitable for smaller datasets, as it groups all related data closely together. The format presents further advantages in that interactions are self-contained units, thus rendering streaming of data much simpler and parsing more efficient. A set of XSLT scripts has been made available to the community, which allows users to convert PSI-MI XML2.5 between compact and expanded forms.

5.4. Java XML Parser

Many molecular interaction databases have chosen the PSI-MI format for providing data to their users. In order to ease the development of tools exploiting this data, a Java library providing a user-friendly data model and the core functionalities for reading and writing PSI-MI XML2.5 data has been provided. Furthermore, the parser supports XML streaming, thus providing a versatile and cursor based approach for retrieving interactions, interactors or experiments. The parser has been made available on the PSI-MI website.

5.5. XMLMaker/Flattener

The XMLMaker/Flattener is a Java application that converts any XML schema into tab-delimited ASCII format (flat files) and vice versa, given a user-defined mapping that can be saved and reused on subsequent files. A PSI-MI mapping can be readily created to inter-convert PSI-MI 1.0 or 2.5 XML files to simple flat files.

Whilst some of the tools listed above have limitations with respect to memory requirements when dealing with large data files and expanded/compact forms of the schema, this problem can be addressed by choosing the appropriate XSLT script.

6. Relationship with Other Community Resources

6.1. BioConductor

Bioconductor is an open source and open development software project for the analysis and comprehension of large-scale biological datasets. The Bioconductor package Rintact provides a provgrammatic interface to the IntAct molecular interaction database (16). It translates the primary data encoded in PSI-MI XML2.5 files into R graph objects, which can then be analysed by a variety of computational methods. The package is currently being expanded to take data from all PSI-MI XML2.5 databases.

6.2. BioPAX

The Biological Pathway Exchange (BioPAX) format is a collaboratively developed data exchange format for biological pathway data that currently uses the PSI-MI ontology internally for modelling associated molecular interactions (17). All PSI-MI entries annotated to “physical association” map to the BioPAX *physical-Interaction* class.

6.3. Cytoscape

Cytoscape (www.cytoscape.org) is a powerful open-source resource for analysing and visualising biological networks. The Cytoscape user community has developed numerous plugins allowing the extension of its functionalities in the area of data format compatibility and network analysis. Cytoscape now allows users to load molecular interaction data in PSI-MI XML1.0 and 2.5 formats without installing additional extensions, enabling data retrieval from one or more databases and its subsequent integration with data from other sources such as high-throughput expression experiments. Release 2.6 includes a Web Service client plugin for downloading networks from IntAct from directly within Cytoscape.

6.4. The International Molecular Exchange Consortium

The cooperation between the many molecular interaction databases that produced much of the work described above, has lead to the formation of the IMEx collaboration, an agreement between major public domain databases to develop common curation standards, to share the curation load and to interchange data between the databases (http://imex.sourceforge.net/). The IMEx collaboration exists to provide a network of stable, synchronised, freely accessible molecular interaction databases; and to jointly capture all published molecular interaction data in a standardised format.

Interaction data deposited with, or curated by, one of the participating databases is regularly exchanged to ensure that all databases hold a single, non-redundant dataset for the user to download or to access over the web. The user will have access to a consistent set of records, maintained at the IntAct, DIP, MINT and MatrixDB interaction database sites, with other partners working towards IMEx membership. Work is ongoing to address the archive of legacy data held separately by each database, to ensure this is distributed between members in the foreseeable future. Data deposition is increasing, with submissions varying from e-mail based, PSI-MI XML, web-form, and Excel spreadsheet, with all these options being supported via the IMEx website (18).

7. Summary

The availability of standards and guidelines to assist in the preparation of molecular interaction data for publication, deposition and data exchange has started to have a significant impact on the field. The PSI-MI XML interchange format has resulted in easy data retrieval and merger – one measure of the success of this is the ever-increasing number of compilation databases, which take curated data from a number of primary sources and repackage this in a single resource. The availability of a single data format has also stimulated tool development and will increase the range and diversity of analytical methods available to both the laboratory scientist and the bioinformatician to assess data quality and build up interaction networks of both increasing complexity and validity.

All such efforts require support from the user community and the PSI-MI is actively seeking input and advice from all quarters. Anyone wishing to become involved is invited to visit http://www.psidev.info, to participate in the discussion groups listed, and to contribute to the further development of community standards for proteomics data in general, and molecular interactions in particular.

References

1. Bader, G.D., Betel, D. and Hogue, C.W.V. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31, 248–250
2. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S. and Eisenberg, D. (2002) DIP: The database of interacting proteins. A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305
3. Kerrien, S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.* 35:D561–565
4. Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bahler, J., Wood, V., Dolinski, K., Tyers M. (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.* 36:637–640
5. Mishra, G.R., Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K.S., Sharma, S., Chandrika, K.N., Deshpande, N., Palvankar, K., Raghavnath R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H.G.M., Nagini, M., Kumar, G.S.S., Jose, R., Deepthi, P., Mohan, S.S. Gandhi, T.K.B., Harsha, H.C., Deshpande, K.S., Sarker, M., Prasad, T.S.K. and Pandey. (2006) A. Human protein reference database–2006 update. *Nucleic Acids Res.* 34: 411–414
6. Chatr-aryamontri, A., Ceol, A., Palazzi, LM., Nardelli, G., Schneider, M.V., Castagnoli, L., Cesareni, G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.* 35: 572–574
7. Kaiser, J. *Science*, 2002, 296, 827.
8. Orchard, S., Hermjakob, H. (2008) The HUPO proteomics standards initiative–easing communication and minimizing data loss in a changing world. *Briefings in Bioinformatics* 9:166–173
9. Taylor, C.F., Paton, N.W., Lilley, K.S., Binz, P-A., Julian R.K.Jr , Jones, A.R., Zhu, W.,, Apweiler, R., Aebersold, R.,.Deutsch, E.W., Dunn, M.J., Heck, A.J., Leitner, A., Macht, M., Mann, M. Martens, L., Neubert, T.A., Patterson, S.D., Ping, P., Seymour, S.L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T.M., Whitelegge, J.P., Wilkins, M.R., Xenarios, I., Yates, J.R., III Hermjakob, H. (2007) The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnol.* 25: 887–893
10. Brazma, A, Hingamp P, Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C, Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C, Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365–371
11. Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stümpflen, V., Ceol, A., Chatr-aryamontri, A. Armstrong, J., Woollard, P., Salama, J.J., Moore, S., Wojcik, J., Bader, G.D., Vidal, M., Cusick, M.E., Gerstein, M., Gavin, A-C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., De Las Rivas, J. , Priesto, C., Perreau, V.M., Hogue, C., Mewes, H-W., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, C., Hermjakob, H. (2007) The Minimum Information required for reporting a Molecular Interaction Experiment (MIMIx). *Nat. Biotechnol.* 25:894–898
12. Taylor, C.F., Binz, P-A., Aebersold, R., Affolter, M., Barkovich, R., Deutsch, E.W.,

Horn, D.M., Huhmer, A., Kussmann, M., Lilley, K., Macht, M., Mann, M., Muller, D., Neubert, T.A., Nickson, J., Patterson, S.D., Raso, R., Resing K., Seymour, S.L., Tsugita, A., Xenarios, I., Zeng, R. Julian, R.K. Jnr. (2008) Guidelines for reporting the use of mass spectrometry in proteomics. *Nature Biotechnol.* 26: 860–861

13. Gibson, F., Anderson, L., Babnigg, G., Baker, M., Berth, M., Binz P-A., Borthwick, A., Cash, P., Day, B.W., Friedman, D.B., Garland, D., Gutstein, H.B., Hoogland, C., Jones, N.A., Khan, A., Klose, J., Lamond, A.I., Lemkin , P.F., Lilley, K.S., Minden, J., Morris, N.J., Paton, N.W., Pisano, M.R., Prime, J.E., Rabilloud, T., Stead, D.A., Taylor, C.F., Voshol, H., Wipat, A., Jones, A.R. (2008) Guidelines for reporting the use of gel electrophoresis in proteomics. *Nature Biotechnol.* 26: 863–864

14. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G.N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., Apweiler, R. (2004) The HUPO PSI's Molecular Interaction format - a community standard for the representation of protein interaction data. *Nature Biotechnol.* **22:**177–183

15. Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A.F., Vinod, N., Bader, G.D., Xenarios, I., Wojcik, J., Sherman, D., Tyers, M., Salama, J.J., Moore, S., Ceol, A., Chatr-aryamontri, A., Oesterheld, M., Stümpflen, V., Salwinski, L., Nerothin, J., Cerami, E., Cusick, M.E., Vidal, M., Gilson, M. Armstrong, J., Woollard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R., Hermjakob, H. (2007) Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC biology* 5:44

16. Chiang, T., Li, N., Orchard, S., Kerrien, S., Hermjakob, H., Gentleman, R., Huber, W. (2008) Rintact: enabling computational analysis of molecular interaction data from the IntAct repository. *Bioinformatics* 24: 1100–1101

17. Luciano, J.S. (2005) PAX of mind for pathway researchers**.** Drug Discovery Today. 10:937–942

18. Orchard, S., Kerrien, S., Jones, P., Ceol, A., Chatr-Aryamontri, A., Salwinski, L., Nerothin, J., Hermjakob, H (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics* 7 Suppl 1:28–34

Chapter 22

Mass Spectrometer Output File Format mzML

Eric W. Deutsch

Abstract

Mass spectrometry is an important technique for analyzing proteins and other biomolecular compounds in biological samples. Each of the vendors of these mass spectrometers uses a different proprietary binary output file format, which has hindered data sharing and the development of open source software for downstream analysis. The solution has been to develop, with the full participation of academic researchers as well as software and hardware vendors, an open XML-based format for encoding mass spectrometer output files, and then to write software to use this format for archiving, sharing, and processing. This chapter presents the various components and information available for this format, mzML. In addition to the XML schema that defines the file structure, a controlled vocabulary provides clear terms and definitions for the spectral metadata, and a semantic validation rules mapping file allows the mzML semantic validator to insure that an mzML document complies with one of several levels of requirements. Complete documentation and example files insure that the format may be uniformly implemented. At the time of release, there already existed several implementations of the format and vendors have committed to supporting the format in their products.

Key words: File format, mzML, Standards, XML, Controlled vocabulary

1. Introduction

Mass spectrometry is an important method to analyze biomolecules by measuring the intact mass-to-charge ratios of their in-situ generated ionized forms or the mass-to-charge ratios of in-situ-generated fragments of these ions. The resulting mass spectra are used for a variety of purposes, among which is the identification, characterization, and absolute or relative quantification of the analyzed molecules. The processing steps to achieve these goals typically involve semi-automatic computational analysis of the recorded mass spectra and sometimes also of the associated metadata (e.g., elution characteristics if the instrument is coupled to a chromatography system).

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_22, © Humana Press, a part of Springer Science+Business Media, LLC 2010

The result of the processing can be assigned a score, rank, or confidence measure.

Differences inherent in the use of a variety of instruments, different experimental conditions under which analyses are performed, and potential automatic data preprocessing steps by the instrument software can influence the actual measurements and therefore the results after processing. Additionally, most instruments output their acquired data in a very specific and often proprietary format. These proprietary formats are then typically transformed into so-called peak lists to be analyzed by identification and characterization software. Data reduction, such as peak centroiding and deisotoping, is often performed during this transformation from proprietary formats to peak lists. The peak lists are then used as inputs for subsequent analysis. However, these peak list file formats lack information about the precursor MS signals and about the associated metadata (i.e., instrument settings and description, acquisition mode, etc.) compared to the files they were derived from. The many different and often proprietary formats make integration or comparison of mass spectrometer output data difficult or impossible, and the use of the heavily processed and data-poor peak lists is often suboptimal.

The solution has been to create open file formats that can encode the information in these output files in XML (Extensible Markup Language) and then write software to read and write these formats. Several formats were developed and a unified format called mzML emerged. This chapter will first describe the history leading to mzML and then provide an overview of the mzML specification and components, followed by information about the implementations of the format.

2. History

During 2003–2005, two data formats to store mass spectrometer output in an open, vendor-neutral, XML format were developed. The mzData format (1) was developed by the Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI), primarily as a data exchange and archive format. The mzXML format (2) was developed at the Institute for Systems Biology (ISB), primarily in order to streamline data processing software. Both formats are used extensively but having two formats for essentially the same information causes unnecessary confusion in the community and adds complexity to software development as often both formats must be supported. Therefore the designers of mzData and mzXML, including representatives of instrument vendors, analysis software developers and end users, have joined under the auspices of the PSI and jointly developed a single format intended to replace the previous two. This new format is named mzML and is the subject of this chapter.

The main difference between the two original formats, aside from the primary intent described above, is the design philosophy of flexibility. The mzData format was designed to be quite flexible via the extensive use of a controlled vocabulary. It was hoped that the actual XML schema could remain stable for many years while the accompanying controlled vocabulary could be frequently updated to support new technologies, instruments, and methods of acquiring data. However, a significant complication with this philosophy was that flexibility in the format led to a variety of styles of the format, with each different tool representing data in a different dialect of the format, causing significant trouble for reader software.

On the other hand, mzXML was designed with a very strict schema with most auxiliary information described in enumerated attributes. This simplified software implementations as there was only one way to present various attributes and the validity of the documents could be easily checked with industry-standard XML validators. However, virtually any desired change to the format, even adding one attribute, would require a new version number. This led to several closely related version numbers, and different software programs supporting different versions.

The main challenge in uniting these two formats was therefore resolving the opposing philosophies rather than fundamental technical issues. The result is a format that contains the best aspects of the two original formats so that it may be widely adopted and resolve the previous problem of two formats.

The will to start the unification process was gathered at the spring 2006 PSI workshop in San Francisco, CA. At this meeting, the technical differences between the two formats were examined and an agreement to move toward unification was reached. The various vendors voiced their displeasure with the state of having two formats, being almost uniformly unwilling to support both. They all voiced that they would implement a unified format developed by the PSI. For two years, a relatively small band of PSI participant volunteers met at workshops and tried to push forward progress between workshops (Fig. 1). Finally, two years after the initial agreement, mzML 1.0.0 was released on June 1, 2008, coinciding with the American Society for Mass Spectrometry (ASMS) conference. The PSI MS working group continues to support the format by maintaining the controlled vocabulary, semantic validators, and documentation as described below.

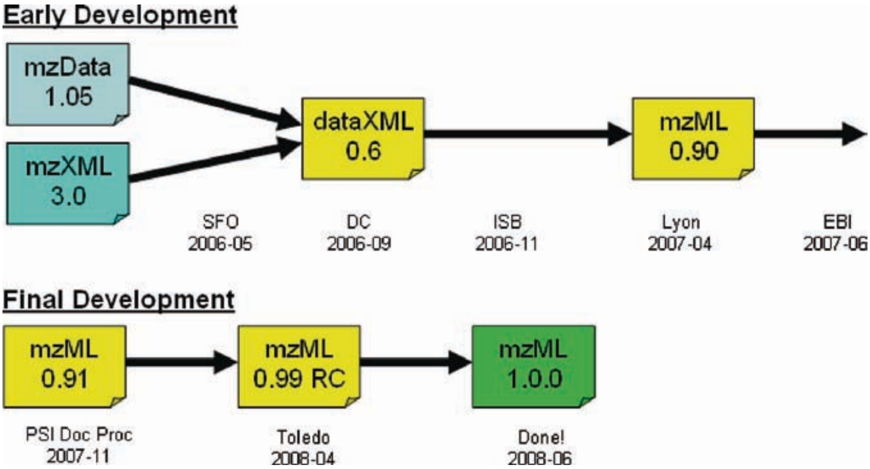


Fig. 1. History of the development of mzML, beginning with the initial unification agreement in May 2006 in San Francisco (SFO), continuing work at PSI workshops in Washington, DC, USA (DC), the Institute for Systems Biology (ISB) in Seattle, USA, Lyon, France, the European Bioinformatics Institute (EBI) in Hinxton, UK, Toledo, Spain, and the 1.0.0 release in June 2008 at the American Society for Mass Spectrometry (ASMS) conference in Denver, USA. Maintenance of mzML will continue with the PSI Mass Spectrometry Standards Working Group. It is expected that the schema will remain stable, but minor updates to the controlled vocabulary and semantic validation rules may be necessary

3. Design of the Format

The design of mzML benefited tremendously from the precursor formats mzXML and mzData. It was not necessary to start from scratch. Rather, the designers could take the best aspects from each of the precursor formats, consider known deficiencies of the previous formats, and apply the lessons learned from years of real-world implementations and use of these formats. In this section, the primary design aspects are presented after a brief discussion of the design philosophy of mzML.

3.1. Design Principles

Since the development of mzML brought together different philosophies, the primary mzML designers agreed on the following design principles that would guide its development:

1. Keep the format simple. Many elaborate extensions were proposed but most were rejected in favor of a simple implementation.
2. Eliminate alternate ways of encoding the same information. Such flexibility, while sometimes touted as a benefit for some products, is bad for data formats.
3. Build in some flexibility for encoding new important information but keep the format stable. There is a strong desire

from companies that develop software for their customers to keep the data format stable over long periods of time with updates to an auxiliary file.

4. Support the features of mzData and mzXML but not much more in version 1.0. The only major new feature deemed crucial was support for selected reaction monitoring (SRM) data (which is also supported in the latest mzXML 3.1 and thus not really new).
5. Finish version 1.0 of the format soon with the resources available. It was felt that the greatest community benefit would be to resolve the mzData/mzXML duality rather than expend limited resources on new features.
6. Validate the new format by implementing software to read and write the format before its release.

There was great temptation to add support for many new kinds of data and representation possibilities. There are many enhancements that have been suggested, but the small group of volunteers that have actively developed this format have opted to focus on the primary goal set before them: develop a single format that the vendors and current software can easily support and thereby obsolete mzData and mzXML. The enhancements not considered compatible with this goal will be entertained for mzML 2.0.

One of the aspects of mzXML that enabled its very swift adoption as a de facto standard was an immediately available set of open source tools that implemented the format. With these tools, many users were able to begin using the format immediately without coding their own software. Therefore, to insure that mzML is a format that will be adopted quickly and implemented uniformly, the format was to be released along with several tools that write, read, and validate the format. It was deemed crucial that at submission to the PSI document process, the following minimum software would implement mzML:

1. Two or more converters that convert from vendor formats to mzML.
2. The popular RAMP parser library that currently supports mzData and mzXML.
3. An mzML semantic validator that checks for correct implementation of files.

These implementations are discussed in Subheading 4.

3.2. XML Schema

An mzML document is designed to contain all the information for a single MS run, including metadata about the spectra plus all the spectra themselves, either in centroided (peak list) or profile mode. At the top of the file, the `<cvList>` element contains information about the controlled vocabularies referenced in the rest of the file. The <fileDescription> element contains some basic information about the type of spectra to be found in the file. The optional <referenceableParamGroupList> element contains a list of groups of controlled vocabulary terms that are used frequently in the file and may simply be defined once and referenced thereafter. Following this basic housekeeping information, the <sampleList> element may optionally contain information about samples that are referenced throughout the file. The <instrumentConfiguration> element contains information about the instrument used for the run (possibly in more than one configuration for hybrid instruments). The <softwareList> and <dataProcessingList> elements provide a history of data processing that may have occurred since raw acquisition. An optional <acquisitionSettingsList> element can hold special input parameters to mass spectrometers such as inclusion lists. This is followed by the actual spectra and optionally some chromatograms. The high-level outline of the schema structure is shown in Fig. 2.

As with its predecessors, mzML is encoded in an XML format. The structure of the format is defined by an XSD (XML schema definition), which is used to insure that documents are properly formed. XML is both easily parsed by computer programs, using

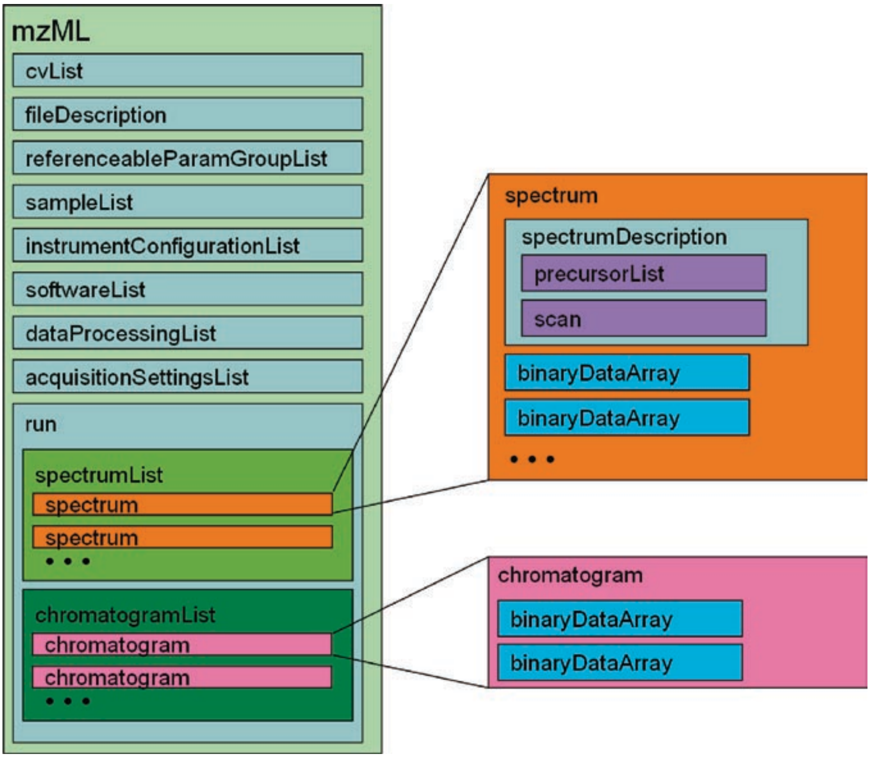


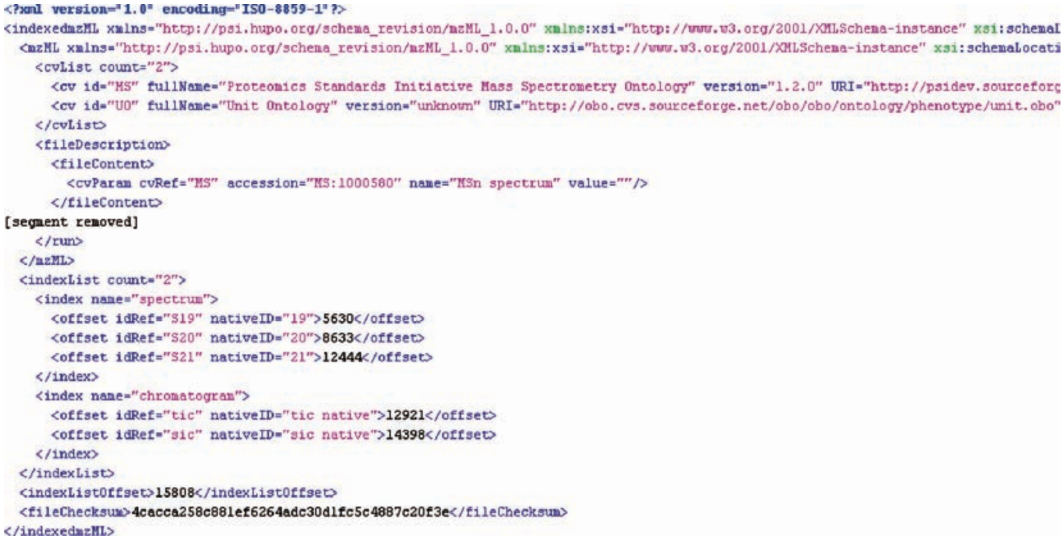
Fig. 2. Overview of the mzML schema. The root mzML element contains elements that provide metadata about a file and run followed by the spectral data itself with optional chromatograms

well-established libraries, and is also relatively easily readable by humans since it is a text-based format. This is a benefit during the design process, and aids in troubleshooting file problems, although it does come at the cost of larger file sizes than with binary formats. However, XML documents do compress well.

One of the requirements for mzML was that it provide a standardized mechanism for a random access index in the same way as mzXML. The use case is that when writing software to process mzML or view individual spectra, it is often necessary to quickly pull out an arbitrary spectrum. If a program needs to display spectrum number 18,345 to display to a user, it must be able to seek to that spectrum in the file rather than read it sequentially if a fast user experience can be expected.

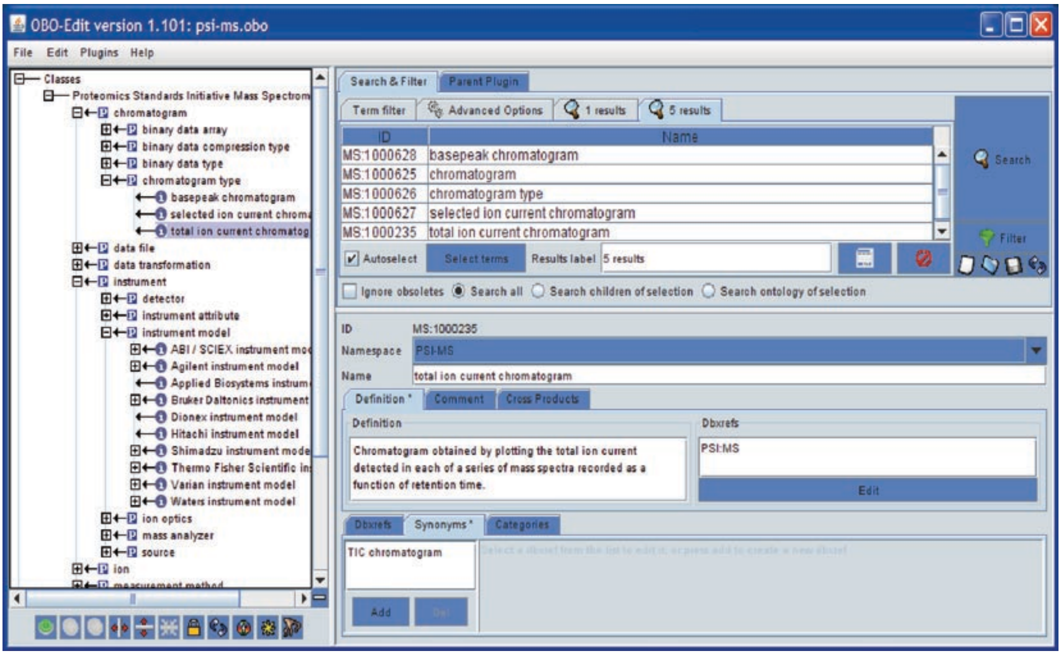
Some have argued that providing a random access index into XML is an anathema to the intent of XML and list the many possible ways in which an index could become broken. However, several years of use of such an index in the mzXML format has shown that the indexing problems are few and the benefits are enormous. Reader software can (and has been) easily written to make use of the index, but automatically rebuild the index if it is noticed to be incorrect.

Since many are not interested in an index, mzML has been designed such that the main part of an mzML document does not contain an index, but that the document may be enclosed in a wrapper schema that includes an index. Therefore, an .mzML file may contain a plain mzML document or an indexed mzML document. Reader software is designed to handle either. A sample snippet of XML showing the wrapper indexing schema is shown in Fig. 3.



```
<?xml version="1.0" encoding="ISO-8859-1"?>
<indexedmzML xmlns="http://psi.hupo.org/schema_revision/mzML_1.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaL
  <mzML xmlns="http://psi.hupo.org/schema_revision/mzML_1.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocati
    <cvList count="2">
      <cv id="MS" fullName="Proteomics Standards Initiative Mass Spectrometry Ontology" version="1.2.0" URI="http://psidev.sourceforg
      <cv id="UO" fullName="Unit Ontology" version="unknown" URI="http://obo.cvs.sourceforge.net/obo/obo/ontology/phenotype/unit.obo"
    </cvList>
    <fileDescription>
      <fileContent>
        <cvParam cvRef="MS" accession="MS:1000580" name="MSn spectrum" value=""/>
      </fileContent>
[segment removed]
    </run>
  </mzML>
  <indexList count="2">
    <index name="spectrum">
      <offset idRef="S19" nativeID="19">5630</offset>
      <offset idRef="S20" nativeID="20">8633</offset>
      <offset idRef="S21" nativeID="21">12444</offset>
    </index>
    <index name="chromatogram">
      <offset idRef="tic" nativeID="tic native">12921</offset>
      <offset idRef="sic" nativeID="sic native">14398</offset>
    </index>
  </indexList>
  <indexListOffset>15808</indexListOffset>
  <fileChecksum>4cacca258c881ef6264adc30d1fc5c4887c20f3e</fileChecksum>
</indexedmzML>
```

Fig. 3. Example top and bottom of an mzML document with the middle segment removed for display purposes. The main part of the mzML document is contained within the `<mzML></mzML>`tags. It is wrapped within an `<indexed-mzML></indexedmzML>` construct, which contains the random access index at the bottom



- Classes
  - Proteomics Standards Initiative Mass Spectrom
    - chromatogram
      - binary data array
      - binary data compression type
      - binary data type
      - chromatogram type
        - basepeak chromatogram
        - selected ion current chroma
        - total ion current chromatog
    - data file
    - data transformation
    - instrument
      - detector
      - instrument attribute
      - instrument model
        - ABI / SCIEX instrument mo
        - Agilent instrument model
        - Applied Biosystems instrum
        - Bruker Daltonics instrument
        - Dionex instrument model
        - Hitachi instrument model
        - Shimadzu instrument mode
        - Thermo Fisher Scientific in
        - Varian instrument model
        - Waters instrument model
      - ion optics
      - mass analyzer
      - source
    - ion
    - measurement method

ID	Name
MS:1000628	basepeak chromatogram
MS:1000625	chromatogram
MS:1000626	chromatogram type
MS:1000627	selected ion current chromatogram
MS:1000235	total ion current chromatogram

☑ Autoselect Results label 5 results

□ Ignore obsoletes ● Search all ○ Search children of selection ○ Search ontology of selection

ID MS:1000235

Namespace PSI-MS

Name total ion current chromatogram

Definition: Chromatogram obtained by plotting the total ion current detected in each of a series of mass spectra recorded as a function of retention time.

Dbxrefs: PSI:MS

TIC chromatogram

Select a dbxref from the list to edit it, or press add to create a new dbxref

Fig. 4. A view of the PSI MS controlled vocabulary as seen in the OBO-Edit program, currently highlighting the term "total ion current chromatogram"

3.3. Controlled Vocabulary

Much of the metadata encoded in the mzML is in the form of a cvParam, an XML element that provides a reference to a specific concept within the PSI MS controlled vocabulary. Each term has an explicit and detailed definition, and may have information about its data type and what kind of units it requires, if any. The controlled vocabulary is edited in OBO format with the OBO-Edit software (3) (Fig. 4) and is used by most readers and writers of mzML. The controlled vocabulary can be easily adjusted and extended without modifying the mzML schema.

3.4. Semantic Validator

The mzData format was a far more flexible format than mzXML. The support of new technologies could be added to mzData files by adding new controlled vocabulary terms, while mzXML often required a full schema revision. However, mzData did suffer from a problem of inconsistently used vocabulary terms and there appeared several different dialects of mzData, encoding the same information in subtly different ways. This was not usually a problem for human inspection of the file, but caused difficulty writing and maintaining reader software.

This problem has been solved (it is hoped) for mzML by releasing a semantic validator with the data format. This semantic validator enforces many rules as to how controlled vocabulary terms are used, not only making sure that the terms are in the CV, but also that the correct terms are used in the correct location in the document and the required terms are present the correct number of times. This allows greater flexibility in the schema, but enforces order in how the CV terms are used. This will require the discipline to use the semantic validator, not just an XML validator, to validate new or updated mzML writer code. The result is that new technologies or information can be accommodated with adjustments to the controlled vocabulary and validator, not to the schema. Opinions differ on whether this is a benefit or not.

Another benefit of using the semantic validator is that different levels of compliance can be defined. For example, if configured with a rules file for basic mzML, the validator will point out any problems that prevent the mzML from being correct at a basic level that should be expected by all parsers. However, for submission to a journal that requires the MIAPE-MS guidelines (4), the validator can be configured to use the MIAPE-MS rules file to check if an mzML file is fully compliant at the MIAPE-MS level. This level of compliance insures significant additional metadata that should be produced for new data and publications but cannot be expected from older data.

Another benefit is that metadata requirements can be adjusted for different types of data. For example, photodiode array (PDA) spectra generated from a mass spectrometer instrument can be encoded in mzML, but the requirements will be different than for mass spectra. These different types of spectra can be encoded using the same tags, just with different metadata, and this can be enforced through the semantic validator.

Semantic rules are encoded in one or more (for different compliance levels) rules mapping files and these can be updated along with the controlled vocabulary without changing the schema. The validator is available as a web page to which any file can be uploaded or as a standalone tool that can validate local files.

3.5. Documentation

The full mzML specification is available as a PDF document at the mzML web site. It describes many of the details of the design of the format and then describes each of the elements of the format in detail, along with figures depicting the structure graphically. The element documentation is autogenerated by some custom software that reads the XSD file, a sample document, the controlled vocabulary, and the rules mapping file and writes out an HTML representation of the information contained within these files. This representation is also imported into the full specification document.

In order to exercise the schema and demonstrate that the various use cases have been adequately modeled, we have developed several example instance documents. Some of the documents are hand-crafted with an ordinary editor, while others are written out as a software test as part of the ProteoWizard reference implementation. In addition, several instance documents are conversions of real data files using ProteoWizard or other implementations of converters.

4. Implementations of the Format

The best way to test a new format is by implementing it in software. Inevitably as a format is implemented, one finds minor inconsistencies or missing features. The initial release of mzML is strengthened by the breadth of implementations that coexisted with the release and have exercised the various use cases.

4.1. Reference Implementation

The ProteoWizard software project (5, 6), initiated by the Spielberg Family Center for Applied Proteomics at the Cedars-Sinai Medical Center, provides a modular and extensible set of open-source, cross-platform tools and libraries. The tools perform proteomics data analyses; the libraries enable rapid tool creation by providing a robust, pluggable development framework that simplifies and unifies data file access, and performs standard chemistry and LCMS dataset computations. During the final stages of mzML development, refinement, and testing, the ProteoWizard library has provided the necessary framework for testing and reference implementation of mzML.

ProteoWizard is modular C++ library with an internal data model that has a one-to-one translation of mzML data elements to C++ data structures. It builds with native compilers on all major platforms (MSVC on Windows, gcc on Linux, XCode on OSX) and is available under the Apache Version 2 license. ProteoWizard has a plug-in Reader interface for reading both open and vendor proprietary data formats: mzML, mzXML, Thermo RAW, MGF; there are additional Readers in development. CLI binding allows use of ProteoWizard libraries from .NET languages (C++/CLI, C#, VB.NET), and SWIG bindings for scripting (from Java, Python, Perl, R) are in development.

ProteoWizard also comes with several tools that make use of the library to perform various processing or display tasks. The msconvert tool provides general file format conversion, including native centroiding and zlib compression. The SeeMS and mspicture tools allow visualization of mass spectral data.

It is worth emphasizing that ProteoWizard has been released under a very permissive license, the Apache Version 2 license, which allows the library to be used in commercial software without influencing the licensing terms of that software. This is in contrast to some other open-source licenses, which require that software that uses such a library also be open source.

4.2. Other Implementations

At the time of this writing, there are many software implementations of the mzML format already in place or emerging. A data format is only as usable as the software that implements it. One of the strengths of mzML is this wide variety of software available that uses mzML. Table 1 lists the available software at the time of this writing. An up-to-date table is available at the mzML web site.

Table 1
A list of software supporting mzML at the time of this writing. An updated list is available at the mzML web site

Author	Product	mzML support
CSHS	ProteoWizard	Full mzML support available today http://proteowizard.sourceforge.net/
ISB	TPP	Full mzML support available today http://tools.proteomecenter.org/
ISB	RAMP, JRAP	Full mzML support available today http://tools.proteomecenter.org/
Insilicos	Insilicos Viewer	Full mzML support available today http://www.insilicos.com/viewer_download.html
GeneBio	Phenyx	Full mzML support available today http://www.phenyx-ms.com/
Vanderbilt	MyriMatch	Full mzML support available today http://fenchurch.mc.vanderbilt.edu/lab/software.php
Thermo Scientific	RAW->mzML conv	Beta
Applied Biosystems	WIFF->mzML conv	Beta
NCBI	NCBI C++ toolkit	Beta http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/
Univ. of Lund	Proteios	Full mzML support available today http://www.proteios.org/
SIB	InSilicoSpectro	Full mzML support available today http://insilicospectro.vital-it.ch/

5. Conclusion

The mzML format is an open, XML-based format for mass spectrometer output files, developed with the full participation of vendors and researchers in order to create a single open format that would be supported by all software. The format includes the best features from preexisting open formats and has additional support for chromatograms and some other features deemed highly desirable. It is expected that the schema will remain stable for at least a year, hopefully more. However, the controlled vocabulary and semantic validation rules will continue to be updated and refined as all authors and vendors finish implementing their software for mzML.

Additional feature requests that cannot be accommodated using the existing schema will be collected and considered for an update release in the next few years.

During the early design phase, RDF (resource description framework) was considered as an alternative to XML. In many ways, the type of flexibility that has been worked into mzML, notably the adapting controlled vocabulary and the semantic validation, are concepts that RDF has the potential to solve nicely. However, it was determined that the developers and the implementer community were not yet ready to try to implement a standard in RDF, a significant departure from custom XML schema. Moreover, the primary goal of the designers was to fix the two-format problem, rather than set out a bold new course. The new mzML 1.0 release fulfills the goals set before the designers. It may well be that the next major release, mzML 2.0, not expected for several years, will be designed using RDF, at a time when the designers and implementers are ready to use this newer platform.

Acknowledgments

The design of mzML was a long process that involved a great many contributors. The author would like to thank the following for their contributions to mzML: Lennart Martens, Pierre-Alain Binz, Darren Kessner, Matt Chambers, Luisa Montecchi-Palazzi, Jim Shofstahl, Josh Tasman, Randall K Julian, Fredrik Levander, Puneet Souda, Jari Häkkinen, Brian Pratt, Erik Nilsson, Mike Coleman, Luis Mendoza, David Shteynberg, Lars Nilse, Benito Cañas, Lola Gutierrez, Alberto Medina, Trish Wheztel, Eva Duchoslav, Henning Hermjakob, Angel Pizarro, Phil Jones, Jimmy Eng, Kent Laursen, Sandra Orchard, Chris Taylor, Patrick Pedrioli, Sean Seymour, David Creasy, Howard Read, Jim Langridge, Jayson Falkner, David Horn, Ruth McNally, Ron Beavis, Norman Paton, Marc Sturm, Parag Mallick, Rune Philosof, David Sparkman, Wilfred Tang, Marius Kallhardt, and Ruedi Aebersold.

EWD has been funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179, and from PM50 GMO76547/Center for Systems Biology.

References

1. mzData, http://psidev.info/index.php?q=node/80#mzdata.
2. Pedrioli, P.G., et al., *A common open representation of mass spectrometry data and its application to proteomics research.* Nat Biotechnol, 2004. **22**(11): p. 1459–66.
3. Day-Richter, J., et al., *OBO-Edit--an ontology editor for biologists.* Bioinformatics, 2007. **23**(16): p. 2198–200.
4. Taylor, C.F., et al., *Guidelines for reporting the use of mass spectrometry in proteomics.* Nat Biotechnol, 2008. **26**(8): p. 860–1.
5. Kessner, D., et al., *ProteoWizard: Open Source Software for Rapid Proteomics Tools Development.* Bioinformatics, 2008. **24**(21): p. 2534–6.
6. ProteoWizard, http://proteowizard.sourceforge.net.

Chapter 23

Managing Experimental Data Using FuGE

Andrew R. Jones and Allyson L. Lister

Abstract

Data management and sharing in omics science is highly challenging due to the constant evolution of experimental techniques, the range of instrument types and software used for analysis, and the high volumes of data produced. The Functional Genomics Experiment (FuGE) Model was created to provide a model for capturing descriptions of sample processing, experimental protocols and multidimensional data for any kind of omics experiment. FuGE has two modes of action: (a) as a storage architecture for experimental workflows and (b) as a framework for building new technology-specific data standards.

FuGE is an object model that is converted into an XML implementation for data exchange. Software toolkits have been developed for data handling and for bridging between XML data files and relational database implementations. FuGE has been adopted by the Proteomics Standards Initiative (PSI, http://www.psidev.info) for building several new data formats, and it is being used in a variety of other experimental contexts, thus allowing data to be integrated across a range of experimental types to support Systems Biology approaches. This chapter provides a practical guide for laboratories or groups wishing to manage their data, and for developers wishing to create new data formats using FuGE.

Key words: Data standards, Functional genomics, Data exchange, Database development

1. Introduction

In the proteomics domain, new experimental techniques are frequently developed, presenting significant challenges for data management and sharing. Over the last 10 years of proteome research, the paradigm has gradually shifted from gel electrophoresis for protein separation, to “shotgun” methods that separate complex mixtures of peptides, based around liquid chromatography. As in the early days of proteomics, the main technique for peptide and protein identification is still mass spectrometry coupled with a database search engine, yet within this space, new approaches are frequently proposed for data processing and statistical analysis, several of which are outlined in other chapters. New methods have also been created for separation, including developments in

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_23, © Humana Press, a part of Springer Science+Business Media, LLC 2010

multidimensional liquid chromatography, pre-fractionation of proteins, capillary electrophoresis, centrifugation and so on (1). In recent years, there have also been numerous methods published for quantifying proteins detected by mass spectrometry, either by relative (2, 3) or absolute measures (4, 5), and by differential gel analysis (6). As such, the types of data and metadata that must be stored are not static, and the volumes of relevant data are increasing rapidly as high-throughput instruments become commonplace. In addition, few laboratories would define themselves solely as proteomics-based; most laboratories use a range of approaches to analyse their samples of interest. In the past, databases have been created that focus on a single experimental technique, such as repositories for mass spectrometry results (7–9), microarrays (10, 11) or protein–protein interactions (12). Data management solutions, which store and analyse proteome data in conjunction with results from the numerous other techniques used to interrogate samples, are thus required.

One of the challenges of managing proteome data is the range of different file types produced by instruments and software, often in closed-source vendor-specific formats. In recent years, several UK-based research councils and funders (e.g. BBSRC, MRC and the Wellcome Trust, for URLs see Table 1) have released data sharing policies that require omics experimental data to be made publicly available as a condition of grant funding. There are also clear benefits to all researchers if experimental data can be made publicly accessible by allowing new findings to be derived beyond the original conclusions of the study, and by facilitating improvements in analysis algorithms. The benefits are diminished, however, if data are published in vendor-specific formats, since few laboratories will have software capable of processing or analysing the files. As such, it is widely recognized that the development of

Table 1
Data sharing policies of the main funders of biological research

Organisation	URL
Wellcome Trust	http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm
BBSRC	http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf
MRC	http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/index.htm

standard exchange formats is important for data sharing (for example see the chapter on mzML). It is also clear that infrastructures that can facilitate the rapid development of new standards and provide solutions for bringing together existing formats (both open-source and vendor-specific) in a single architecture are needed.

The Functional Genomics Experiment (FuGE) model is a technology-independent data model for storing descriptions of experimental processes (13). FuGE contains high-level models to describe protocols, biological materials (such as samples) and multidimensional data. It also provides mechanisms for referencing other resources such as ontologies, databases and external data files. FuGE can be used to describe protocols for sample processing, separation, data acquisition and data processing as well as track the flow of samples and data through the process as inputs and outputs. It can also be used as a framework for creating new data standards specific to a certain technology, such as gel electrophoresis (14). By providing a mechanism for describing the overall study, including references to data files, FuGE can integrate proteomic data with other data types, for instance those relating to genes, metabolites or phenotypes.

In this chapter, we first present a brief description of the contents of the abstract FuGE Object Model (FuGE-OM) and the different concrete representations available. These include XML (FuGE-ML) for data exchange, relational database implementations and software toolkits for bridging between XML objects and database storage. The chapter then provides a practical guide for developers wishing to use FuGE either for managing experimental data or for the creation of technology-specific extensions.

2. Methods

FuGE is a large, complex model that has been documented in detail elsewhere (13, 15). Subheading 2.1 therefore provides a basic introduction to the main concepts, sufficient to understand the different ways in which the model can be used. FuGE has two modes of action. First, FuGE can be used as an experimental metadata *integrator* to represent basic laboratory workflows to capture sample tracking, experimental protocols or procedures, the data files that result (represented in the relevant external format) and data processing pipelines (Subheading 2.2). Second, FuGE can be used as an *extensible core* to build new data formats that share a common underlying structure. In this mode, models can be created to represent specific details about a new experimental technique for which there is no existing data standard (Subheading 2.3).

2.1. Overview of the FuGE Model

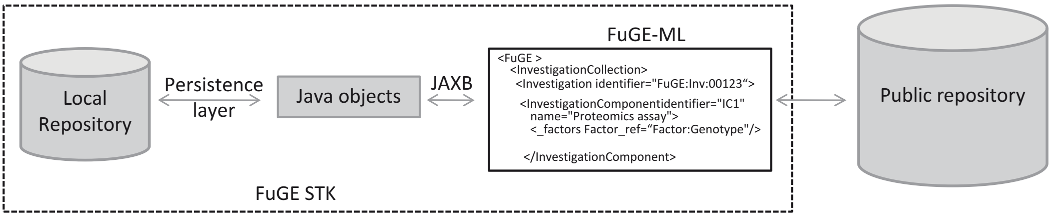
FuGE is a model for representing the running of laboratory or computational protocols and the flow of samples and data files. There is an important distinction made between the details of the intended procedure (Protocol) and the running of the procedure (ProtocolApplication). Protocols should be defined once, with a set of sequential steps and parameters with default values. Protocols can be associated with descriptions of instruments and software, which can also have parameters. ProtocolApplications reference the Protocol that was performed, and provide the backbone of a workflow by mapping all inputs and outputs. ProtocolApplications can be annotated with the operator of the protocol, the date, any small deviations from the set protocol and any parameters that differed at runtime. This simple, flexible format is able to cover the majority of biological use cases because it abstracts the representations of experiments to the basic building blocks: procedures and their inputs and outputs. Inputs and outputs can be either materials or data. FuGE does not define detailed models for biological materials (e.g. samples) since the variety in what could be described is vast. Instead, FuGE has a structure for importing ontological annotations so they can be stored and exchanged alongside experimental descriptions. Linking materials and other similar FuGE concepts to appropriate terms in ontologies, such as those within the Open Biomedical Ontologies (OBO) Foundry (16), allows semantically rich descriptions to be created without the need for detailed FuGE models of those concepts. The OBO Foundry is a collaboration of developers of science-based ontologies created to establish a set of principles for ontology development. Foundry ontologies are intended to be interoperable reference ontologies in the biomedical domain. For example, the Ontology for Biomedical Investigations (OBI) Consortium (http://purl.obofoundry.org/obo/obi) is developing an integrated ontology for the description of biological and clinical investigations, designed to support the consistent annotation of biomedical investigations, regardless of field of study or data type. The OBI ontology is well-suited for use within the FuGE structure, as it models a number of terms useful for describing experiments, including the protocols, instrumentation and material used, the data generated, the type of analysis performed and other artefacts generated during an investigation.

Examination of two possible uses of the ProtocolApplication entity demonstrates the differences between the integrator and extensible core modes of action of FuGE. A standard ProtocolApplication allows any number of input samples/biological materials and data files, and any number of output materials or data files. In other words, the model itself places no restrictions on what can be captured. Therefore when used as an integrator (Subheading 2.2), any constraints must be created by a specific software implementation or left to the user to provide sensible values representing their domain of interest. The benefit of this mode is that development time is focussed on the creation of the user interface, and not around modification of the FuGE model. However, the standard ProtocolApplication class can also be extended (Subheading 2.3). This can be done when using FuGE as an extensible core. For instance, a gel-scanning protocol in GelML could restrict the input to be a single two-dimensional gel, and restrict the output to be an image file. Using FuGE in this extensible core mode provides much higher control over what can be put into the model, but requires development effort in model extension. The integrator mode may be more useful for communities where the core model of FuGE is enough for their needs, or where all constraints on experimental metadata can be easily written directly into a shared user interface. The use of FuGE as an extensible core may be more suitable to communities or groups with large numbers of complicated procedures, where providing specific extensions of FuGE concepts is the best way to ensure correct deposition of experimental metadata.

2.2. Representing Laboratory Workflows

For communities in which FuGE will be used in integrator mode, i.e. without extension, the following general steps should be followed. For those using FuGE as an extensible core, the steps dealing with data representation and interface creation should be considered after following the stages outlined in Subheading 2.3.

1. *Data representation*. A suitable storage mechanism must be chosen, for example using a relational database or a native XML database. The FuGE project provides a toolkit for manipulation and validation of XML and two toolkits for relational database storage, based on different underlying technologies (http://fuge.sourceforge.net/). See Fig. 1 for the roles that are played by the different technologies in the FuGE toolkits.



```
<FuGE >
  <InvestigationCollection>
   <Investigation identifier="FuGE:Inv:00123“>

    <InvestigationComponentidentifier="IC1"
      name="Proteomics assay">
      <_factors Factor_ref=“Factor:Genotype"/>

    </InvestigationComponent>
```

Fig. 1. The interplay between different technologies for manipulating experimental descriptions within FuGE. A local repository can be implemented using a relational database. A Persistence layer is used to convert between in-memory objects (e.g. in Java) and database storage. The FuGE project provides two toolkits employing different persistence layers, based on EJB (http://java.sun.com/products/ejb/) and Hibernate (http://www.hibernate.org/). JAXB (https://jaxb.dev.java.net/) mappings are included in the toolkit, which can convert in-memory Java objects to an XML representation, FuGE-ML, which will allow data to be sent to public repositories that accept FuGE-formatted data

2. *Interface*. An interface must be created for capturing metadata about studies entered by experimentalists and to integrate the outputs from different instruments or software. Existing open-source projects, such as SyMBA (http://symba.sourceforge.net) or sysFusion (from the Friedrich Miescher Institute for Biomedical Research, http://www.fmi.ch/), utilise the FuGE toolkits, providing relational database/graphical user interfaces to the core FuGE model.
3. *Storage of Protocols*. Typically, a database would store experimental protocols that are commonly performed in the laboratory to be referenced by ProtocolApplications when experiments are performed.
4. *Sample tracking*. The interface should be capable of tracking the flow of samples and data files through an experimental workflow. FuGE can be used to model at the level of detail of a LIMS (Laboratory Information Management System), achieved through the use of ProtocolApplications, which reference the input and outputs of each stage: samples or data files.
5. *Integration of other data standards*. In a proteomics workflow, mass spectrometry data may be represented in the mzML format, or in a vendor-specific format, such as MGF (Mascot Generic Format). Protein and peptide identifications may also be represented in the PSI's mzIdentML format or the output produced by the search engine. The FuGE model can integrate such data files into an entire workflow by defining ProtocolApplications that reference external data files as an output. The inputs to the ProtocolApplications are descriptions of samples, for example output from a separation technology such as liquid chromatography or gel electrophoresis (Fig. 2).

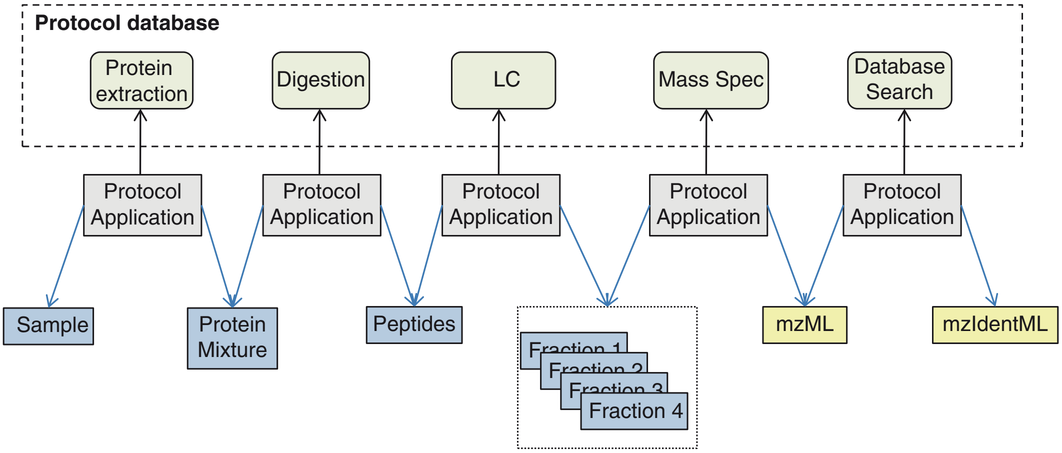


Fig. 2. An example LC-MS experiment represented in FuGE. The inputs and outputs to each ProtocolApplication are either descriptions of samples (annotated with ontology terms) or data files, such as mzML or mzIdentML

6. *Export of data to public repositories.* The types of metadata that should be reported about a proteomics experiment to support a publication are currently in a state of flux. As one example, the PRIDE database will support deposition of data in PSI-sanctioned formats, such as those containing MS data and protein and peptide identifications. It is expected that public repositories will support deposition of omics results tied together using FuGE-ML in the near future.

2.3. Building Technology-Specific Extensions

FuGE can be used to develop new formats for describing particular experimental details. For proteomics, extensions have been created for representing gels (GelML) and general separations and sample processing, including liquid chromatography (spML). The PSI format for mass spectrometry database searches (peptide and protein identification) is mzIdentML, which also makes use of several FuGE structures. Formats have also been developed using FuGE for flow cytometry, genetical genomics, RNA interference, metabolomics, microarrays and e-neuroscience, which may become standards in due course. This section provides a practical guide for developers wishing to understand how FuGE can be used to create a new data format, for example to represent a particular type of experiment.

2.3.1. Stage 1 – Develop Reporting Requirements and Use Cases

The Proteomics Standards Initiative has developed a series of minimum reporting guidelines under the MIAPE (Minimum Information About a Proteomics Experiment) parent document. Each of the MIAPE checklists defines the minimal information that should be reported about a particular type of proteomics technique. The MIAPE parent document (17) outlines the purpose and principles of the guidelines with a series of modules, generally one per technique, including mass spectrometry (18), mass spectrometry informatics (19), gel electrophoresis (20) and others. The MIAPE documents represent a formalisation of community opinion on the types of metadata that should be captured about an experiment to allow it to be critically understood and for data to be re-analysed in the future. While some experimental techniques may not require a formal MIAPE module prior to the creation of a data exchange format, an essential first step towards developing a new format is to gain widespread input on the types of data that laboratories wish to exchange. In many cases, this involves collecting use cases, such as different protocols used by laboratories, sets of results, references to publications and so on, and deciding which use cases should be supported by the standard.

2.3.2. Stage 2 –Survey Existing Formats and Examine Requirements for Using FuGE

An important stage in creating new data formats is examining existing formats produced by software and instrument vendors, and those used by individual laboratories or consortia for managing their data. It is useful to identify the deficiencies of existing formats with respect to the use cases or reporting requirements that are to be supported, as identified in Stage 1.

FuGE is intended for representing metadata about experiments and representing a flow of multiple processes. For experimental techniques that encompass only a single stage producing large quantities of data, using FuGE would not be recommended, since the complexity of FuGE may slow development of the format. As described above, FuGE has a mechanism for referencing non-FuGE-based formats and capturing details about the protocols used to create the data. As an example in proteomics, mzML is being developed as a stand-alone format for representing the output of a mass spectrometry. While mass spectrometers may perform two (or more) protein/peptide fragmentation stages, the data can be viewed as a discrete, atomic element, and thus there is no need for the flexible process model of FuGE. FuGE is particularly recommended for describing experimental approaches where there is significant flexibility in how different groups use the technology or where multiple different stages can be performed with outputs of one stage becoming inputs to the next.

2.3.3. Stage 3 – Building a FuGE Extension

The FuGE UML model can be downloaded from the website (http://fuge.sourceforge.net/) and viewed with MagicDraw community edition (http://www.magicdraw.com/). The concepts identified in stage 1 should be partitioned according to the top-level FuGE classes, in particular: Protocol, ProtocolApplication, Action, Parameter, Material, InternalData and ExternalData. Classes are extended in FuGE by using a UML inheritance relationship, which is automatically applied to the XML Schema (discussed in Stage 4).

Protocol – A Protocol is intended to represent a description of an *intended* process or one that is carried out multiple times with the same settings. One example would be a standard operating procedure for sample preparation in proteomics. The Protocol class should be extended to capture specific types of parameters or steps (Action) within the protocol. For instance, in sample preparation, individual steps might be protein extraction, solubilisation and separation.

Parameter – Parameters are replaceable values within a Protocol that can be assigned a default value. The Parameter class should be extended to capture a specific type of setting within a Protocol, instrument (Equipment) or Software, such as temperature or time.

Material – Materials represent all substances used in an experiment. In practice, Material is typically used to capture details of samples of the organism being studied. The Material class has no attributes for describing its properties; instead, two mechanisms can be used to add descriptions. First, Materials can be annotated with ontology terms to describe the characteristics, for example using ontologies defined within OBI. Second, the Material class can be extended to add additional attributes. We recommend the use of ontologies particularly for describing the fundamental characteristics of the starting sample in a study such as species, observations, phenotypes, medical histories and so on. It is highly challenging to build data models to represent such a range of information and it is unlikely that information stored within such a model could be interpreted by any other software systems. By using ontologies, it should be possible to annotate sample descriptions with terms that are comprehensible to software systems and allow queries over repositories without requiring natural language processing or synonym searches. We recommend extending the Material class in FuGE to store details about substances related directly to the experimental technique, for example those requested by a reporting guideline document. As an example, in GelML, the Material class has been extended to describe electrophoretic gels. The Gel class has attributes for capturing the physical dimensions, the percentage acrylamide in the gel and the ratio of acrylamide to a cross-linking agent. Such specific details are required by the MIAPE document for gel electrophoresis (20), and as such they are included in the data model as attributes on the Gel class to simplify the capture of these values.

Data – FuGE has a structure that can be used to describe multidimensional data (InternalData). The model first defines the dimensions of data, and the values are stored separately in matrices. The data matrices can be accessed using coordinates defined by a combination of the dimensions. The FuGE data model should be extended to describe multidimensional data with a regular size of dimensions, for instance this would be appropriate for describing protein arrays. In other cases, individual elements can be created in FuGE to describe simple results (for example as input/output parameter values), or an external data format can be referenced as an input or output from a ProtocolApplication, such as spectra captured as ExternalData.

ProtocolApplication – ProtocolApplication represents the running of a Protocol and maps the inputs and outputs of the processes. ProtocolApplication should be extended when a specific type of Protocol is being run, and when it is necessary to define specific types of samples (Material) or data files as inputs or outputs.

2.3.4. Stage 4 – Creating an XML Data Interchange Format

The object model is converted to an XML Schema using the XSD STK (http://fuge.sourceforge.net/stks/xsd-stk/), which defines and constrains what can be represented in XML. It is a lightweight toolkit designed expressly for those wishing to either manipulate the FuGE XSD (XML Schema Definition) or generate a new FuGE-based XSD for a particular purpose. The rules used to convert the object model into an XML Schema are described in the FuGE specification document (15). Where an extension has been built to describe a particular technology, the same toolkit can be used to create a new XSD for the extended model.

3. Summary

FuGE has been created to simplify the development of data management solutions for life sciences and attempts to unify data formats created for different technologies. The chapter has provided a brief guide in using FuGE for either of its intended purposes, as a metadata integrator and as an extensible core for creating new data formats. For proteomics data management, both mechanisms are in current use and a community of developers continue to contribute tools and new extensions, which can be accessed via the FuGE website.

References

1. Malmström, J., Lee, H., and Aebersold, R. (2007) *Current Opinion in Biotechnology* **18,** 378–84.
2. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) *Nat Biotech* **17,** 994–99.
3. Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) *Mol Cell Proteomics* **1,** 376–86.
4. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003) *Proceedings of the National Academy of Sciences of the United States of America* **100,** 6940–45.
5. Pratt, J. M., Simpson, D. M., Doherty, M. K., Rivers, J., Gaskell, S. J., and Beynon, R. J. (2006) *Nat. Protocols* **1,** 1029–43.
6. Van den Bergh, G., and Arckens, L. (2004) *Current Opinion in Biotechnology* **15,** 38–43.
7. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006) *Nucl. Acids Res.* **34,** D655–58.
8. Jones, P., Cote, R. G., Martens, L., Quinn, A. F., Taylor, C. F., Derache, W., Hermjakob, H., and Apweiler, R. (2006) *Nucl. Acids Res.* **34,** D659–63.
9. Prince, J. T., Carlson, M. W., Wang, R., Lu, P., and Marcotte, E. M. (2004) *Nat Biotech* **22,** 471–72.
10. Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007) *Nucl. Acids Res.* **35,** D760–65.
11. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P., and Sansone, S.-A. (2003) *Nucl. Acids Res.* **31,** 68–71.
12. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., and Hermjakob, H. (2007) *Nucl. Acids Res.* **35,** D561–65.
13. Jones, A. R., Miller, M., Aebersold, R., Apweiler, R., Ball, C. A., Brazma, A., DeGreef, J., Hardy, N., Hermjakob, H., Hubbard, S. J., Hussey, P., Igra, M., Jenkins, H., Julian, R. K., Laursen, K., Oliver, S. G., Paton, N. W., Sansone, S.-A., Sarkans, U., Stoeckert, C. J., Taylor, C. F., Whetzel, P. L., White, J. A., Spellman, P., and Pizarro, A. (2007) *Nat Biotech* **25,** 1127–33.
14. Jones, A. R., and Gibson, F. (2007) *Proteomics* **7,** 35–40.

15. Jones, A., Miller, M., Spellman, P., and Pizarro, A. (2007) http://fuge.sourceforge.net/dev/V1Final/FuGE-v1-SpecDoc.doc.

16. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007) *Nat Biotech* **25,** 1251–55.

17. Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P.-A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J. R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., Seymour, S. L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T. M., Whitelegge, J. P., Wilkins, M. R., Xenarios, I., Yates, J. R., and Hermjakob, H. (2007) *Nat Biotech* **25,** 887–93.

18. Taylor, C. F., Binz, P.-A., Aebersold, R., Affolter, M., Barkovich, R., Deutsch, E. W., Horn, D. M., Huhmer, A., Kussmann, M., Lilley, K., Macht, M., Mann, M., Muller, D., Neubert, T. A., Nickson, J., Patterson, S. D., Raso, R., Resing, K., Seymour, S. L., Tsugita, A., Xenarios, I., Zeng, R., and Julian, R. K. (2008) *Nat Biotech* **26,** 860–61.

19. Binz, P.-A., Barkovich, R., Beavis, R. C., Creasy, D., Horn, D. M., Julian, R. K., Seymour, S. L., Taylor, C. F., and Vandenbrouck, Y. (2008) *Nat Biotech* **26,** 862–62.

20. Gibson, F., Anderson, L., Babnigg, G., Baker, M., Berth, M., Binz, P.-A., Borthwick, A., Cash, P., Day, B. W., Friedman, D. B., Garland, D., Gutstein, H. B., Hoogland, C., Jones, N. A., Khan, A., Klose, J., Lamond, A. I., Lemkin, P. F., Lilley, K. S., Minden, J., Morris, N. J., Paton, N. W., Pisano, M. R., Prime, J. E., Rabilloud, T., Stead, D. A., Taylor, C. F., Voshol, H., Wipat, A., and Jones, A. R. (2008) *Nat Biotech* **26,** 863–64.

Chapter 24

Proteomics Data Collection (ProDaC): Publishing and Collecting Proteomics Data Sets in Public Repositories Using Standard Formats

Christian Stephan, Martin Eisenacher, Michael Kohl, and Helmut E. Meyer

Abstract

In Proteomics, fast enhancements with regard to technology are responsible for the creation of huge data sets. Consequently, in 2006 the European Commission funded a Coordination Action named ProDaC (Proteomics Data Collection) within the 6th EU Framework Programme to foster a community-wide data collection and data sharing. The aims of ProDaC were the development of documentation and storage standards, setup of a standardized data submission pipeline and collection of data.

To reach these goals, the necessary work was structured in six thematic fields (work packages): Standards for Proteomics Data Representation, Standards Implementation, Data Integration Tools, Proteomics Repository Adaptation, Data Flow Management, and Proteomics Data Exploitation. The methods building the basis of the respective fields and the achieved results are described in the following sections.

Key words: Bioinformatics, European Commission, European Union, ProCon, ProDaC, Proteomics data collection, Standards, Standard formats, Public repositories, Data collection

Abbreviations

HUPO	Human Proteome Organisation
PSI	Proteomics standards initiative
DCC	Data collection center
ProDaC	Proteomics data collection
LIMS	Laboratory information management system
XML	eXtensible markup language
PRIDE	Proteomics identification database
CV	Controlled vocabulary

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_24, © Humana Press, a part of Springer Science+Business Media, LLC 2010

1. Introduction

Proteomics approaches deal with the analysis of the protein composition on different spatial scales and complexity levels (organelles, cells or even tissues). During the last decade, proteomics gained focus in both fundamental and clinical research. Along with this interest, both methods and instruments, e.g. mass spectrometry devices, developed rapidly.

Hence, a huge amount of data is produced within shorter time periods. In order to cope with this excess of information and the technical improvement, a rapid development of adequate software applications is necessary. Unfortunately, these solutions often were adapted for rather special problems or devices. Furthermore, different vendors save results in different data formats, leading to restrictions on both comparison and collection of such heterogeneous data. In today's science, new insights are strongly related to workable data sharing within the scientific community, but this is hampered by the above-mentioned data incompatibility and proprietary formats. As a consequence, valuable resources were tied up to develop data conversion tools for these different data formats.

Even if data conversion into standard formats would be possible, the experimentalist or laboratory scientist may not be familiar with the tools available. A last fact, which prevents successful data sharing, is that data sets may have the correct format, but are not submitted to a public access point, because the single researcher does not see the usefulness or simply does not take the time to do this.

There was clearly an increasing need for the development of generic standardized data formats. Therefore, volunteers from science and industry joined forces in 2002 (1) and founded the Proteomics Standards Initiative (PSI) (2), which is an essential part of the Human Proteome Organisation (HUPO) (3). The outcome of the PSI activities include, for example, the definition of "Minimum Information about a Proteomics Experiment (MIAPE)" guidelines (4, 5) and the development of important standard formats (PSI-MI, PSI-MOD and others), which are gaining wide acceptance within the proteomics community. To push forward easy data conversion and to enable and perform community-wide data collection, the Proteomics Data Collection consortium (ProDaC) (6) was initiated in 2006. It supports the PSI efforts and goes beyond them by establishing easy data workflow pipelines, working together with community journals and collecting data submitted by its members. ProDaC is a Coordination Action project within the 6th EU Framework Programme (7).

One main objective of the consortium is to coordinate the development of international standards concerning proteomics data sets. Initially, the focus was on the finalization of standard formats developed within the PSI. This was the first time the European Commission funded the development of standards in proteomics.

Both usability and acceptance of the newly developed standard formats strongly depend on an easy conversion of common proteomics data formats into the new standards. In order to maintain results from previous experiments, it has to be assured that no information will be lost during the conversion. Therefore, the ProDaC initiative develops conversion tools and further proteomics related software tools integrating these standards.

Scientific progress strongly relies on a rapid and simple capability to share data and to compare experimental results. That was, for example, the experience in large proteomics consortia such as HUPO's Brain Proteome Project (HUPO BPP (8–20), Fig. 1). It required easy-to-use standard validation tools and software solutions using the standards intuitively. This implies not only standardization of data formats but also the development of a standardized workflow and information management starting from sample preparation and processing of the results up to a centralized and public storage of proteomics data. Consequently, ProDaC also

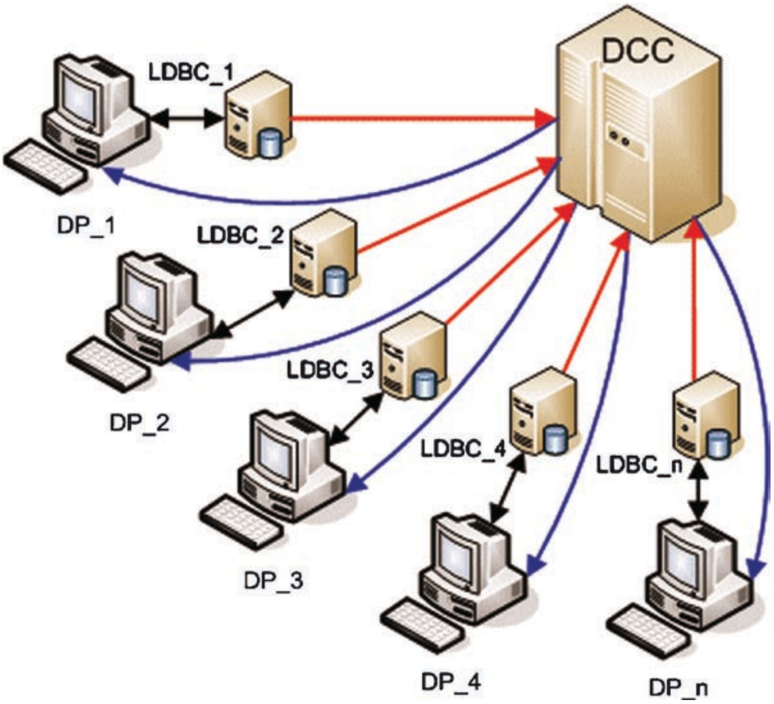


Fig. 1. Data exchange network of proteomics data repositories in the HUPO BPP (*DCC* data collection center, *LDBC* local database collection, *DP* data provider)

deals with the development of a "pipeline" integrating all proteomics related working stages into a robust and generic framework.

An additional objective is the support of a central repository for proteomics data sets which is crucial for a qualified peer-review process of publications. In relation to the publication process, ProDaC will work together with scientific journals and will help authors to realize minimum information requirements and to provide correctly prepared supplementary information.

Last, but not the least the consortium aims to deploy data sets into standards-compliant repositories and sketch possibilities to make use of these data collection going beyond the questions a single scientific project can answer – as done in system-wide analysis such as the Systems Biology approaches, where large amounts of data are fundamental.

2. Methodical Aspects

In this section, it is described which working scheme was planned to achieve the ProDaC goals, whereas in the "Results" section, the achieved goals and results are described. The following sections correspond to the ProDaC working units called "work packages".

2.1. Standards for Proteomics Data Representation

A primary objective of ProDaC is to support the finalization of standard data formats, which are developed within the PSI consortium.

Standardized data formats are used for encoding data obtained at different experimental stages:

- the output obtained from the mass spectrometer,
- the input for data processing software,
- and a result format encoding protein/peptide identification.

Therefore, ProDaC needs to coordinate standardization efforts on these three levels of data handling, where the first two are ideally merged.

One of the most important features of newly defined standards is, besides the good and precise definition, the community acceptance and real usage in practice. On the one hand, good documentation of implementations and their publication is necessary, and on the other hand an easy-to-use software solution to generate or convert data directly at the producing devices. This will be the breakthrough rule for PSI standards. ProDaC is a systematic organizational platform, which offers the possibility for combining existing experiences and know-how in the field of proteomics-related bioinformatics. Such an association is an appropriate means for supporting the rapid development and finalization of the PSI standard data formats.

2.2. Standards Implementation

Tool providers sometimes claim compatibility to a certain standard, but actually only implement it in a very rudimentary or even an incorrect way. ProDaC will ensure the correct implementation by providing a test suite, an online validator for semantic validation, an instance document collection and a certification.

Today mzML (earlier levels of development are mzData and mzXML) is the standard for mass spectrometry data exchange. mzML has been released in version 1.0, so a test suite is appreciated and necessary. The standard exchange format of search result data called AnalysisXML (now called mzIdentML) is still under development, but the minor changes and the designed test suite can be easily adapted to this standard as well.

Most of the defined PSI standards for data exchange are designed in XML (eXtensible Markup Language, (21)). One of the biggest advantages of XML is the existence of an XML schema definition (XSD) file for placing exact constraints on the contents. This allows easy validation of actual XML files content using existing parsers ("schema validation"). However, the complexity of the data content and cross-references inside PSI XML files prevent the validation of the content by using only an XSD file. Therefore, a more advanced validation tool called "semantic validator" is needed. At the same time such a semantic validator can check for appropriate use of a "controlled vocabulary" (CV) term at a particular position in the XML file (via so-called CV rules). CVs are a fixed set of allowed words or phrases used to differentiate between units of information. CVs can be used within the concept of ontology, which is a branch within computer science. Ontologies are frequently used for example in the fields of artificial intelligence and bioinformatics. An ontology tries to establish relationships between categories and concepts.

The standards development process will be strongly facilitated by the generation of instance documents. In order to support an easy exchange of ideas, ProDaC will host these documents on a website accessible for both ProDaC and PSI partners.

A survey among all ProDaC partners showed a high interest and need for instance documents and validation tools. It became clear that additional data files complementary to the PSI example files would be useful, and a validation tool accessible via the internet is of main importance for implementers and users. Such an online validator needs to be freely available and platform-independent to allow validation in ongoing implementations by mass spectrometer vendors and in software packages such as Proteios, ProteinScape™, Mascot, Sequest™ or Phenyx™.

Validation is a pre-requisite in the concept of a "ProDaC certification". Such a certification could be appropriate to assure high quality of the data submitted to public repositories and thus improve the compatibility of shared data.

2.3. Data Integration Tools

For formats being mainly used as data exchange standards in mass spectrometry, it is essential to have support for the standards in major systems and programs. Therefore, in ProDaC software tools are developed to convert or export proteomics data sets using the new standard formats (e.g. mzML, AnalysisXML). There is hope that key proteomics software will implement these standards directly after or even during definition. But it cannot be expected that all relevant applications consider the new standards immediately, in particular those of non-consortium partners. The implementation might also be delayed due to maintenance cycles in the vendors' software development process. Therefore, ProDaC develops adaptor applications in a timely manner. To achieve this, it is necessary to gain an overview of proteomics data producing platforms used in the consortium. After that, tools are implemented for conversion and export, where necessary.

In order to evaluate the most important development needs, it is indispensible to acquire a sufficient overview of software solutions and data formats used within the proteomics community. The ProDaC consortium can be considered as a representation of major methods and approaches within this field. Therefore, a questionnaire sent out to the ProDaC members seems to be adequate in order to acquire insight in actual data management solutions and ProDaC relevant objects of the proteomics community.

2.4. Proteomics Repository Adaptation

Databases are essential components when large amounts of data are acquired and stored over a long time period or in different laboratories. Data acquisition in the field of proteomics frequently requires concerted efforts of many scientific and industrial institutions. Therefore, such a data repository platform is indispensable for the overall success of the project.

There are several platforms available that were designed for storing data obtained from common proteomics related working schemes, e.g. PRIDE (22–24), ProteinScape™ (25, 26), Proteios (27–29), Mascot Integra (30), ms_lims (31), Proline (32), ProteusLIMS – Genologics (33–35), SBEAMS Proteomics (36), SibioClé LIMS (37), PeptideAtlas (38–40), and the Trans-Proteomic Pipeline (TPP) (41).

Some major characteristics have to be considered regarding a database adopted for storing proteomics data:

- In order to obtain an easy and fast access for the scientist from all over the world, a web-based user interface is mandatory.
- The repository should be adopted for the new standard data formats. This implies the possibility for both modification and evaluation of the database within local test environments. An open-source project is probably most promising in order to achieve these requirements.
- Scientific progress benefits from fast and easy capabilities for communication and publishing. Nowadays, technical progress

leads to huge amounts of data resulting in voluminous articles or in outsourcing of supplementary material that is usually hosted on the website of the publisher. There is a need for a new publication strategy, which allows obtaining key statements of a publication with sufficient clarity while retrieving details (e.g. protein identifications) on demand. Public data repositories are promising to conform to these specifications. This will noticeably reduce workload associated with the publication process. Currently, data management using public repositories is suggested in the editorial sections of the journals Proteomics, Nature Biotechnology and Nature Methods (42–44) indicating the importance of this issue from the publishers' perspective.

- Furthermore, such a standardized and structured publication strategy may offer possibilities for an automatic interpretation of published results by the use of computer programs.
- Data management in public repositories provides opportunities for sophisticated data exploitation, i.e. comparison or meta-analysis of data obtained from different sources. Furthermore, the proteomics results can then be used in a feedback loop for improving the information stored in sequence databases like UniProt and Ensembl. Therefore, a database should support both comprehensive queries covering different data sets and feedback update mechanisms in cooperation with sequence databases.

2.5. Data Flow Management

An optimal data workflow and management are mandatory for the overall success of the ProDaC project. Therefore, efficient data flow between data producing sites and the data storage platforms (PRIDE, ProteinScape™, Proteios, etc.) must be ensured.

The workflow of proteomics data in the consortium has to consider two principle types:

1. A file-related branch: data files are stored in the file system of instrument computers or at central file locations of an institute. This branch includes input and result files from Mascot™ and Sequest™ search engines, as well as files in the standard formats mzData and mzXML. Another format of this branch is the Microsoft® Excel format (Microsoft Corporation, Redmond, SA, USA).
2. A database-related branch, where data from LIMS systems or existing databases have to be imported into the ProDaC central repository (PRIDE). Examples for this branch are ProteinScape™, IntelliMS, Proteios, and Proline.

Regarding (1) it is a crucial point, that spectra and results are stored in separate files, so before storing them into the repository, a correct assignment of related spectra/result files has to be performed and the contents have to be merged.

Furthermore, the sample and protocol information may be missing, especially in the file-related branch. For both branches, standardized CV or ontology terms are currently not used as demanded by the standards. It takes additional efforts to make these fit to the standard formats.

For manuscript submissions, obligatory publishing of data sets as supplementary data is probable for the mid-term. Therefore, the ProDaC data flow should support that and should be extensively tested.

2.6. Proteomics Data Exploitation

In scientific research, the data amount generated per experiment was steadily increasing throughout the previous decades. There were neither commonly accepted public data repositories for protein and peptide identifications nor standardized publishing rules. Screening of previously identified proteins is hence time-consuming and often impossible. Therefore, some studies are repeated again and again. Large public proteomics data repositories, such as PRIDE (45), are necessary to avoid this loss of knowledge, time, and resources.

Standardized deposition of proteomics data in public repositories will improve the quality of published data by enabling more systematic scientific scrutiny. Performing systematic proteomics data collection has an obvious benefit for the whole community by determining associated data sets and common trends. Because public repositories store data obtained from different experimental approaches/techniques, they provide more added value for subsequent analyses, e.g., systems biology approaches, than using data that come only from one source and one type of method.

The HUPO Plasma Proteome Project (PPP) may serve as a good example for another type of problem, which could be avoided by utilizing proteome data; the PPP had to deal with the problem that observations shared by multiple participants for a predicted protein sequence were deleted from Ensembl or RefSeq during the runtime of the PPP pilot phase, in spite of high-quality mass spectrometry identifications of these sequences. These deletions were due to changes in the gene prediction algorithms. By establishing a feedback loop of proteomic results into sequence databases, the predicted proteins could achieve better evidence.

2.7. Consortium Structure

The objective of Coordination Actions in the European Union's Framework Programme is networking or co-ordination of activities. They always involve at least three independent legal entities from three different countries (46).

Working schemes like those proposed for ProDaC strongly need broad-based community participation. Therefore, the ProDaC consortium is composed of 12 core partners and more than 30 associated partners (Fig. 2). This structure aims to integrate knowledge throughout the proteomics community and

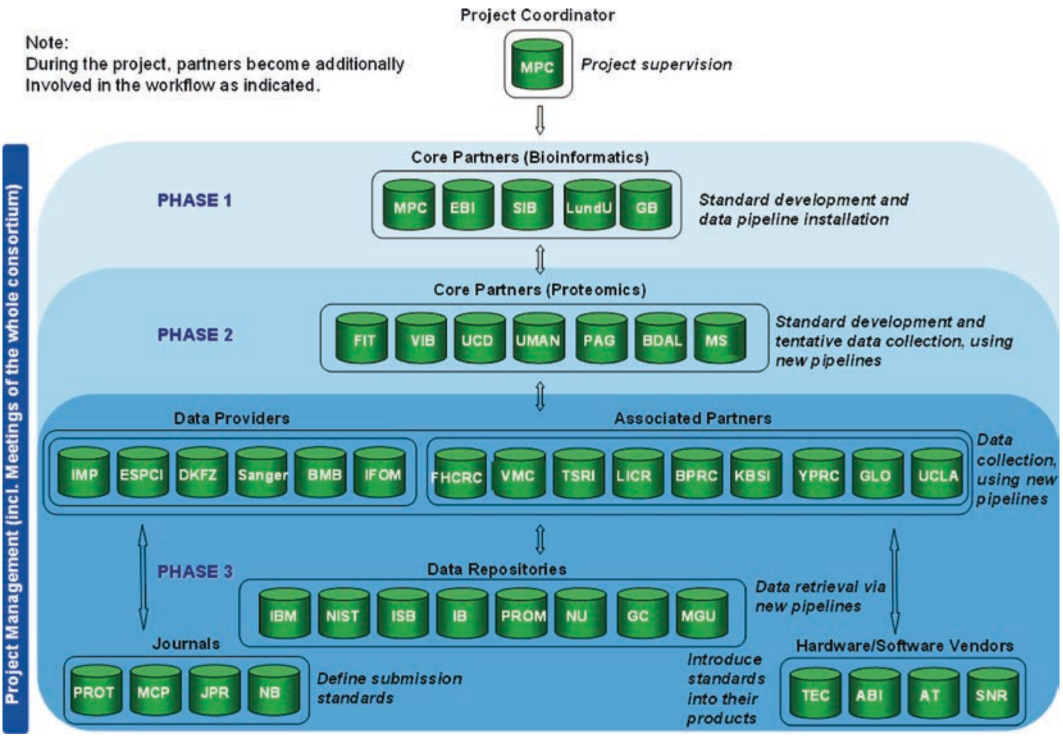

Note:
During the project, partners become additionally Involved in the workflow as indicated.
Project Coordinator
MPC
Project supervision
Project Management (incl. Meetings of the whole consortium)
PHASE 1
Core Partners (Bioinformatics)
MPC
EBI
SIB
LundU
GB
Standard development and data pipeline installation
PHASE 2
Core Partners (Proteomics)
FIT
VIB
UCD
UMAN
PAG
BDAL
MS
Standard development and tentative data collection, using new pipelines
Data Providers
IMP
ESPCI
DKFZ
Sanger
BMB
IFOM
Associated Partners
FHCRC
VMC
TSRI
LICR
BPRC
KBSI
YPRC
GLO
UCLA
Data collection, using new pipelines
PHASE 3
Data Repositories
IBM
NIST
ISB
IB
PROM
NU
GC
MGU
Data retrieval via new pipelines
Journals
PROT
MCP
JPR
NB
Define submission standards
Introduce standards into their products
Hardware/Software Vendors
TEC
ABI
AT
SNR

Fig. 2. The ProDaC consortium members and the project phases

ensures intensive development and testing of both technical correctness and usability.

Additionally, ProDaC establishes cooperation between different academic and industrial groups usually involved in proteomics studies: experimentalists, data providers, bioinformatics scientists, and scientific journals.

Because close collaboration is crucial during the development and testing phases ProDaC aims to concentrate the working process by organization of regular meetings (47–50) and by establishing a structured communication network that can be used for conferences and discussions via the internet.

3. Results

In this section, the goals and results achieved in ProDaC are described. The following sections correspond to the ProDaC working units, called “work packages”.

3.1. Standards for Proteomics Data Representation

A primary objective of ProDaC is to support the finalization of standard data formats.

Several objectives concerning the development of standard formats were already achieved:

1. The newly developed standard XML format for mass spectrometry data is called mzML and was released as version 1.0.0 on June 1st, 2008 with substantial participation of ProDaC partners. Concepts from two frequently used XML formats were integrated, which were designed for the representation of mass spectrometer data: mzData (provided by the PSI) and mzXML (provided by Seattle Proteome Center (SPC) at the Institute for Systems Biology (51)).
2. XML schema definition files, documentation and example files are available via internet (52).
3. The PSI consortium also works on an XML format designed for encoding the output of search engines that were commonly used in the field of proteomics research. This format is referred to as AnalysisXML and is currently in an advanced stage of development (53).
4. The European Bioinformatics Institute (EBI) developed the proteomics data repository PRIDE (Proteomics Identifications Database), which uses its own XML format called PRIDE XML (version 2.1). The schema definition can be found at (45). PRIDE will implement mzML and AnalysisXML as soon as they are finalized; additionally, PRIDE will cover further proteomics information, e.g. those defined in the "Minimum Information about a Proteomics Experiment (MIAPE)" guidelines (4, 5).
5. The PSI standard for protein modifications and for gel-based separation (GelML) is released, and a standard for non-gel-based separation (spML) is released in beta status.

3.2. Standards Implementation

ProDaC provides an instance document collection, a test suite containing an online validator for semantic validation, and a certification mechanism.

3.2.1. Generation of Instance Documents

The implementation of the mzML and AnalysisXML standards is a still ongoing task, the 1.0 version of mzML is released but ontologies and controlled vocabularies will be updated on a regular basis. Therefore, the generated instance documentations of the ProDaC partners are being made available on a Trac Wiki website (54) (linked from the ProDaC website). This enables a regular and wide discussion of the format and possible issues, which are sent back to the PSI and the ProDaC partners. The site is structured into areas for mzData, mzML, AnalysisXML and PRIDE XML. Each area contains example files and editable documents where obstacles and ideas for developments of the standards are recorded.

3.2.2. Beyond XSD Schema Validation

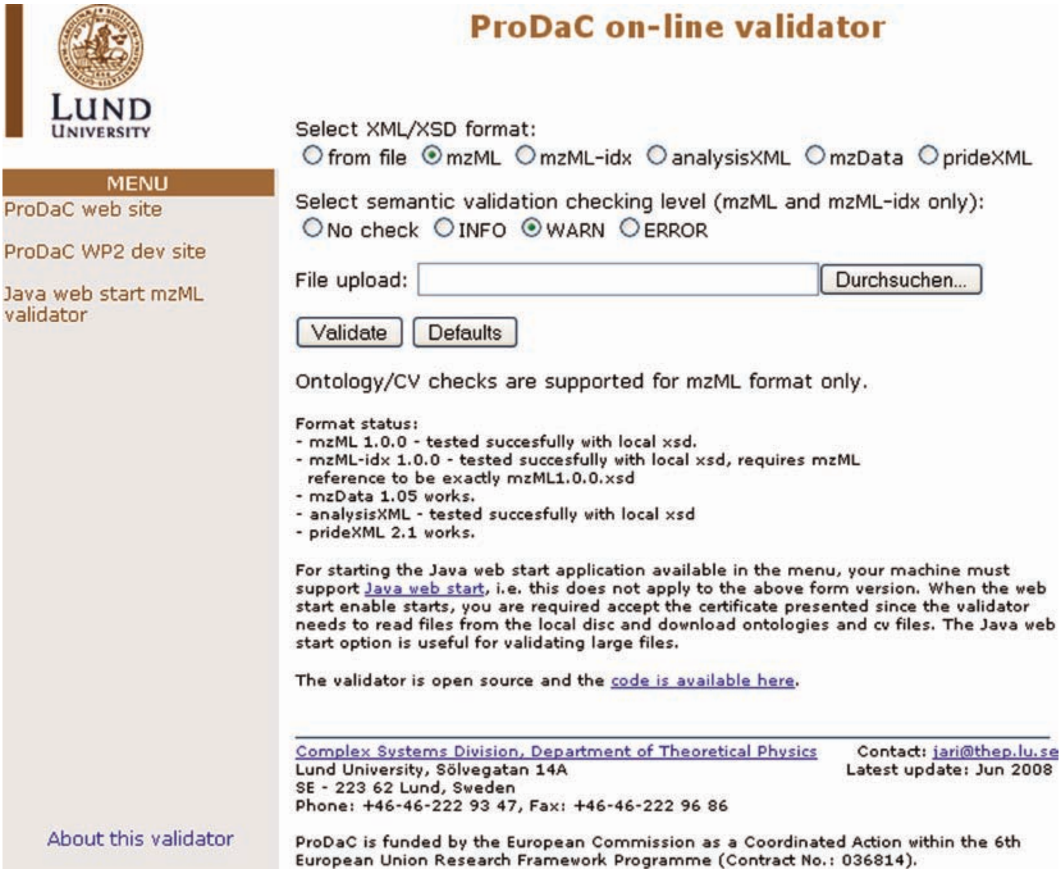
While numerous implementations for XML validation with XSDs already exist, the validation of PSI standards needs to validate features not included in XSD validation only. A combination of schema validation and semantic validation is needed. The extensive use of controlled vocabularies, counts, references and unique identifiers require specific implementations to validate the semantics in the file as well as the logical structure. For these purposes, the Proteomics Services team at the EBI, a partner within ProDaC, has developed a generic, extensible framework for the implementation of such validators for the PSI standards, as well as other XML formats. All of this code is available as open source under the Apache2 license on the PSI subversion (svn) repository (55). A prototype mzML validator has been programmed using this framework, and a Molecular Interactions (PSI-MI) version is also completed.

The mzML version 0.93 release candidate package also contained a Perl-based schema validator for testing generated files. The validator could be extended with semantic control, but the current version performs XSD validation only, and is therefore meant as a simple means to verify structural correctness of an mzML file. A disadvantage is that some of the required libraries are not available for the Active State version of Perl, which is the most common Microsoft Windows Perl engine, and that the validator will therefore require Cygwin or a similar Unix simulator in order to work under Windows. For these reasons, this tool can be considered as a part of a development kit for programmers rather than an end-user tool.

3.2.3. Creation of Semantic Validation Tools

As mentioned above, a prototype semantic validator for mzML using the generic framework has been developed. Software developers can leverage the power of custom-written validation rules, which can be specified in an XML file containing object rules. The code of the validator thus does not need to be changed or updated at all. Ontology or domain experts on the other hand can specify CV rules to validate the correct usage of ontology terms (relations, dependencies, repetitions). Access to controlled vocabularies or ontologies is provided by the OntologyAccess framework that transparently supports both direct access to local or remote Open Biomedical Ontology (OBO) files, as well as access via the EBI's Ontology Lookup Service (OLS; http://www.ebi.ac.uk/ols) APIs. This OntologyAccess framework has built-in caching functionalities to ensure performance. All changes needed for further updates of object rules, controlled vocabularies and ontologies can be performed by editing the XML files and it is not necessary to change the code of the validator itself.

To enable easy validation of files, and also possible logging of common errors in standard files, a website with an online validator has been set up (56) with the mzML validator code as back-end.



ProDaC on-line validator

Select XML/XSD format:
○ from file ● mzML ○ mzML-idx ○ analysisXML ○ mzData ○ prideXML

Select semantic validation checking level (mzML and mzML-idx only):
○ No check ○ INFO ● WARN ○ ERROR

File upload:

Ontology/CV checks are supported for mzML format only.

Format status:
- mzML 1.0.0 - tested succesfully with local xsd.
- mzML-idx 1.0.0 - tested succesfully with local xsd, requires mzML reference to be exactly mzML1.0.0.xsd
- mzData 1.05 works.
- analysisXML - tested succesfully with local xsd
- prideXML 2.1 works.

For starting the Java web start application available in the menu, your machine must support Java web start, i.e. this does not apply to the above form version. When the web start enable starts, you are required accept the certificate presented since the validator needs to read files from the local disc and download ontologies and cv files. The Java web start option is useful for validating large files.

The validator is open source and the code is available here.

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A
SE - 223 62 Lund, Sweden
Phone: +46-46-222 93 47, Fax: +46-46-222 96 86

Contact: jari@thep.lu.se
Latest update: Jun 2008

ProDaC is funded by the European Commission as a Coordinated Action within the 6th European Union Research Framework Programme (Contract No.: 036814).

Fig. 3. The ProDaC on-line validator

This site provides a front end for basic XSD validation, and to more advanced validators that include semantic tests. A screenshot of the online tool is shown in Fig. 3.

3.2.4. Testing/Certification

To acquire ProDaC-certification, files generated by tools that pass the validator in its final form are accepted by certified proteomics data repositories and storage tools. The validator enables the repository maintainers to work with well-defined files and the validator puts pressure on tool vendors to adapt the standards as they mature to a level useful for validation. There are several standard compliant files (the test suite) available at (54) for testing the ability of software to import these schematically and semantically valid data files.

In short, ProDaC certification of a tool/repository comprises two criteria:

1. Files *generated* from a tool/repository must pass the online validator; and if the tool/repository *imports* standard files, it should also pass the below item.

2. The test suite of files must be imported by the repository/tool. A stricter criterion is to require that a repository should import any file that passes the validator.

The test suite and validation tools will be soon adapted for application on other PSI standards as these standards are being finalized.

3.3. Data Integration Tools

In ProDaC software tools are developed to convert or export proteomics data sets using standard formats (e.g. PRIDE XML, mzML, AnalysisXML).

3.3.1. Evaluating Development Needs

Relevant tools used in the consortium have been evaluated by a questionnaire. It asked for all types of ProDaC relevant objects involved in a proteomics workflow: hardware, software, file formats, conversion tools, local LIMS systems and central data repositories.

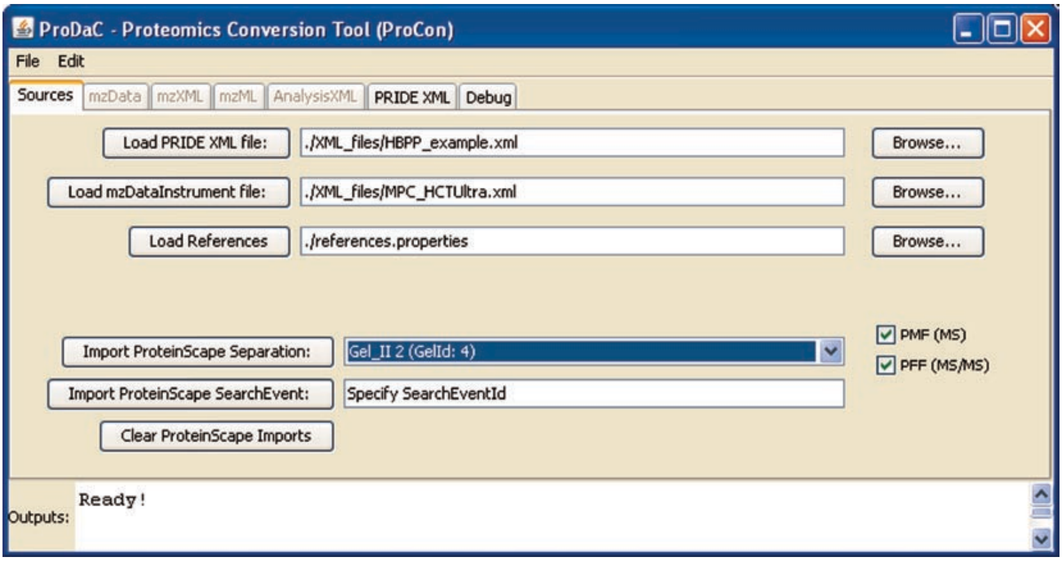
From the feedback, the answers were counted given for a specific question and it was decided (48) to develop tools for the most-frequent environments with higher priority: Sequest™ (Thermo Electron Corporation, Waltham, MA, USA) and Mascot (MatrixScience Inc., Boston, MA, USA) spectra and result formats, mzData, HTML and the ProteinScape™ database (Bruker Daltonics GmbH, Bremen, Germany).

3.3.2. Implementation Activities

Conversion tools convert results and spectra into the PSI XML based file formats and thus enable a submission to public repositories like PRIDE (Proteomics Identification Database (22–24)). Currently, PRIDE accepts its own XML-based import format called PRIDE XML. There are some implementation activities inside the consortium developing export to PRIDE XML. As soon as AnalysisXML as the result standard is finalized, the tools will be adapted in order to provide export functionality into this new format. A further task is the implementation of a semi-interactive "wizard" for CV mapping and ontology improvement as a separate tool beyond the scope of ProDaC. A list of implementation activities is maintained on the ProDaC web page (57).

3.3.3. ProCon

A conversion tool named ProCon (Proteomics Conversion) is developed at the Medizinisches Proteom-Center, Ruhr-Universitaet Bochum, Germany. Designated key features are: (1) import of files from different search engines; (2) import from locally installed Proteomics databases; (3) export using the formats of central data repositories; (4) (semi-)interactive data editing using an ergonomic user interface in order to complete missing information or to solve conflicts. Among others, the export of data sets stored in ProteinScape™ (Bruker Daltonics GmbH, Bremen, Germany) into PRIDE XML is implemented.



Load PRIDE XML file: ./XML_files/HBPP_example.xml

Load mzDataInstrument file: ./XML_files/MPC_HCTUltra.xml

Load References ./references.properties

Import ProteinScape Separation: Gel_II 2 (GelId: 4)

☑ PMF (MS)

☑ PFF (MS/MS)

Import ProteinScape SearchEvent: Specify SearchEventId

Outputs: Ready!

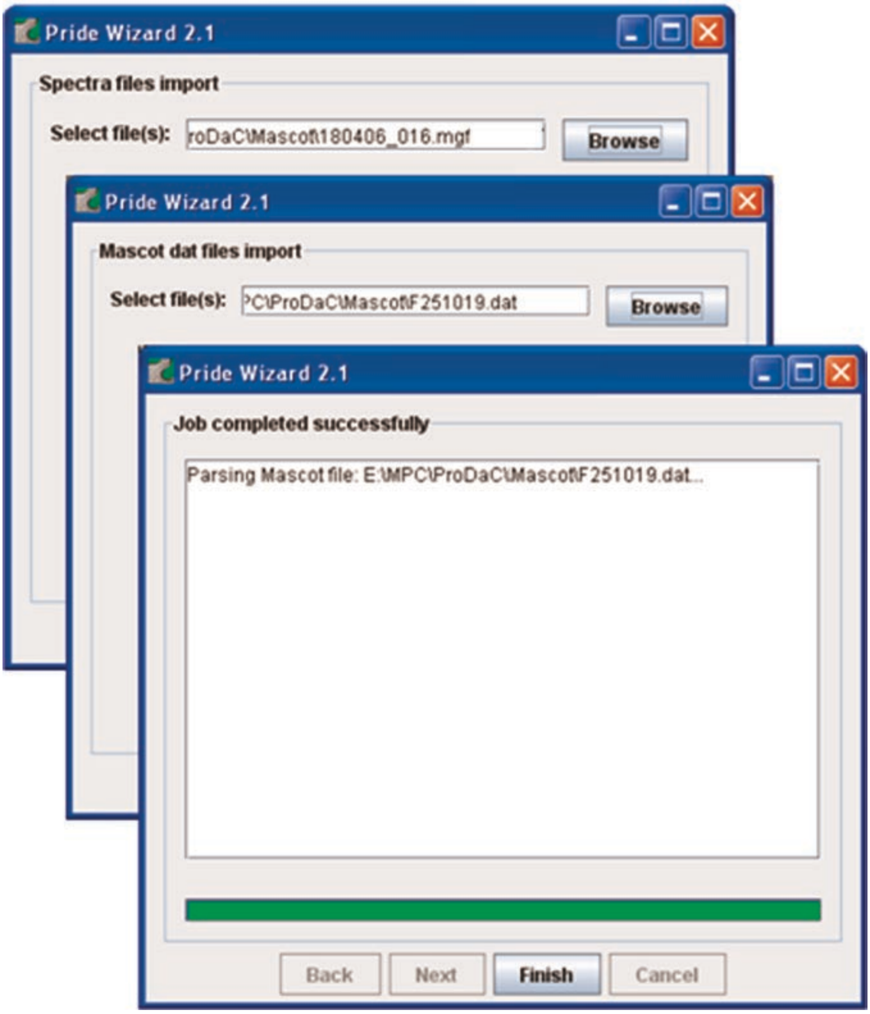
Fig. 4. The “Sources” tab of the ProCon tool

The exported files contain proteins, peptides with modifications and spectra. The types of exported data sets tested so far cover: 1D gel, 2D gel, 1D LC, both MS and MS/MS, with protein assembly performed by (1) single search engine, (2) multiple search engines and (3) ProteinExtractor™ (an algorithm for protein assembly integrated into ProteinScape™). ProteinScape™ vocabulary terms (e.g. for Modifications) are mapped to PSI-MOD ontology terms using the “cross-reference” mechanism of the OBO file format (Open Biomedical Ontology (58)). Furthermore, a sub-standard of the PRIDE XML standard has been distilled storing only the characteristics of a mass spectrometer (*see* (6), section Results and Pubs for the sub-schema and two example instrument files). The prototype user interface of ProCon is shown in Fig. 4.

The next steps for ProCon include the import of Sequest™ (Thermo Electron Corporation, Waltham, MA, USA) .dta and .out files. With MascotDatFile a Java parser for Mascot results is available (59). The conversion of Mascot (MatrixScience Inc., Boston, MA, USA) results to PRIDE XML is implemented within activities of consortium partners (*see* next Subheadings).

3.3.4. PRIDE Wizard

Pride Wizard (60, 61) was developed by the Faculty of Life Sciences, MCISB, Manchester; example files and sources are available. In a wizard-like user interface, the user specifies the minimum necessary information, chooses a spectra file (formats .mzData, .mzXML, .mgf or .pkl), a Mascot result file (format .dat) and – optionally – specifies iTRAQ labelling as a quantification method. Then a valid PRIDE XML file is exported. Some screenshots of the wizard are shown in Fig. 5.



Spectra files import

Select file(s): roDaC\Mascot\180406_016.mgf

Mascot dat files import

Select file(s): PC\ProDaC\Mascot\F251019.dat

Job completed successfully

Parsing Mascot file: E:\MPC\ProDaC\Mascot\F251019.dat...

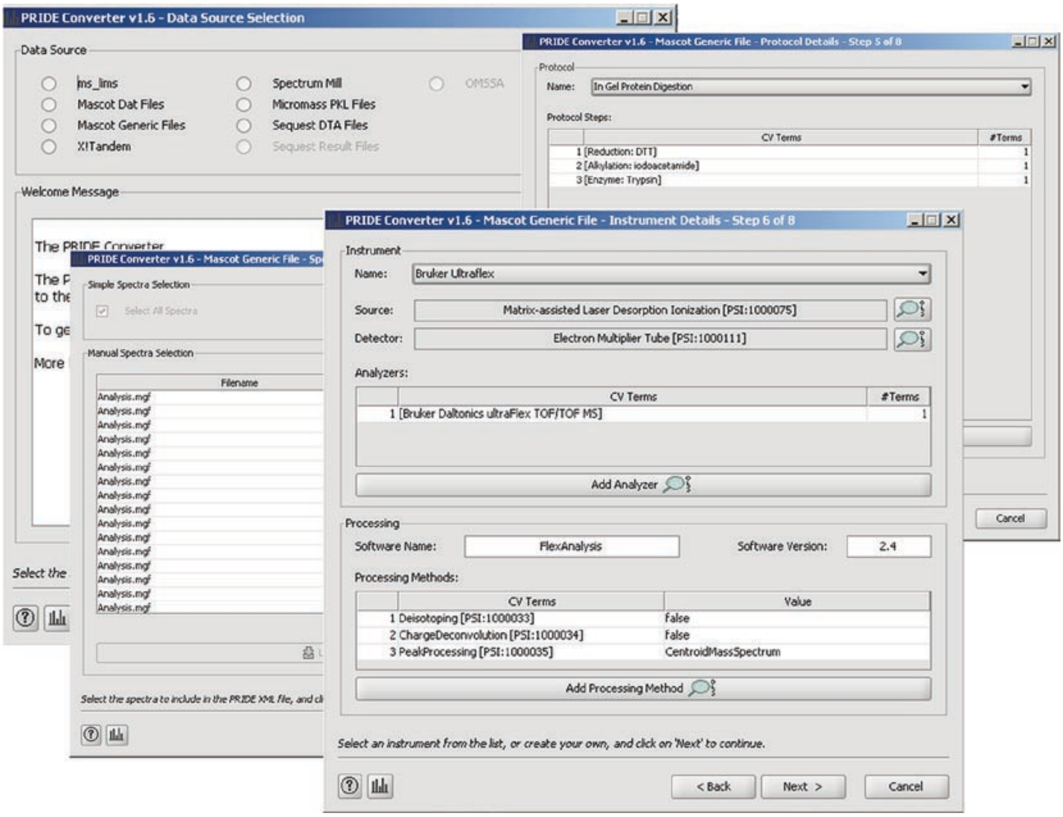
Fig. 5. Selection of different screenshots of PRIDE Wizard

3.3.5. PRIDE Converter

The PRIDE Converter for conversion to PRIDE XML is developed by Harald Barsnes (The Bergen Center for Computational Science, UNIFOB AS affiliated to the University of Bergen, Norway). A development website is available (62). The current version of the PRIDE converter works with ms_lims (31), Mascot Generic Files (.mgf), Mascot .dat Files, Sequest output, X!Tandem and Spectrum Mill (63), but other formats are to be included soon. Some pages of the converter are shown in Fig. 6.

3.4. Proteomics Repository Adaptation

ProDaC supports the development of the "PRoteomics IDEntifications database" (PRIDE (45)), which is hosted at the European Bioinformatics Institute (EBI), because PRIDE features all the requirements mentioned in Methodical Aspects. Development of PRIDE has reached an advanced level so far



PRIDE Converter v1.6 - Data Source Selection

Data Source

○ ms_lims ○ Spectrum Mill ○ OMSSA
○ Mascot Dat Files ○ Micromass PKL Files
○ Mascot Generic Files ○ Sequest DTA Files
○ X!Tandem ○ Sequest Result Files

Welcome Message

The PRIDE Converter

The P
to the

To ge

More

Select the

PRIDE Converter v1.6 - Mascot Generic File - Protocol Details - Step 5 of 8

Protocol

Name: In Gel Protein Digestion

Protocol Steps:

	CV Terms	#Terms
1	[Reduction: DTT]	1
2	[Alkylation: iodoacetamide]	1
3	[Enzyme: Trypsin]	1

PRIDE Converter v1.6 - Mascot Generic File - Sp

Simple Spectra Selection

☑ Select All Spectra

Manual Spectra Selection

Filename
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf
Analysis.mgf

Select the spectra to include in the PRIDE XML file, and cl

PRIDE Converter v1.6 - Mascot Generic File - Instrument Details - Step 6 of 8

Instrument

Name: Bruker Ultraflex

Source: Matrix-assisted Laser Desorption Ionization [PSI:1000075]

Detector: Electron Multiplier Tube [PSI:1000111]

Analyzers:

	CV Terms	#Terms
1	[Bruker Daltonics ultraFlex TOF/TOF MS]	1

Processing

Software Name: FlexAnalysis

Software Version: 2.4

Processing Methods:

	CV Terms	Value
1	Deisotoping [PSI:1000033]	false
2	ChargeDeconvolution [PSI:1000034]	false
3	PeakProcessing [PSI:1000035]	CentroidMassSpectrum

Select an instrument from the list, or create your own, and click on 'Next' to continue.

Fig. 6. Selection of different screenshots of PRIDE Converter

including fundamental features for data submission. This facilitates implementation of the intended data submission pipeline without delay. Furthermore, the EBI is one of the ProDaC core partners and work package leaders, which ensures an efficient collaboration between the members of the ProDaC consortium and the people developing PRIDE.

Because the PRIDE database was originally designed as a generic public repository for proteomics data, some adjustments have been implemented:

- Acceptance of public data repositories is often hampered due to the lack of a clear benefit offered to experimentalists that submit data to these databases. Therefore, PRIDE contains an extension to allow submission of data with controlled access given to so-called “collaborators”. With the same mechanism, peer review is supported (for details, see Section 2.6).
- PRIDE was also extended by a tool for set-oriented data comparison of private data, enabling a standard operation of proteomics publication preparation. The owner of private datasets will be able to compare a given data set to other data sets in PRIDE (see Fig. 7).

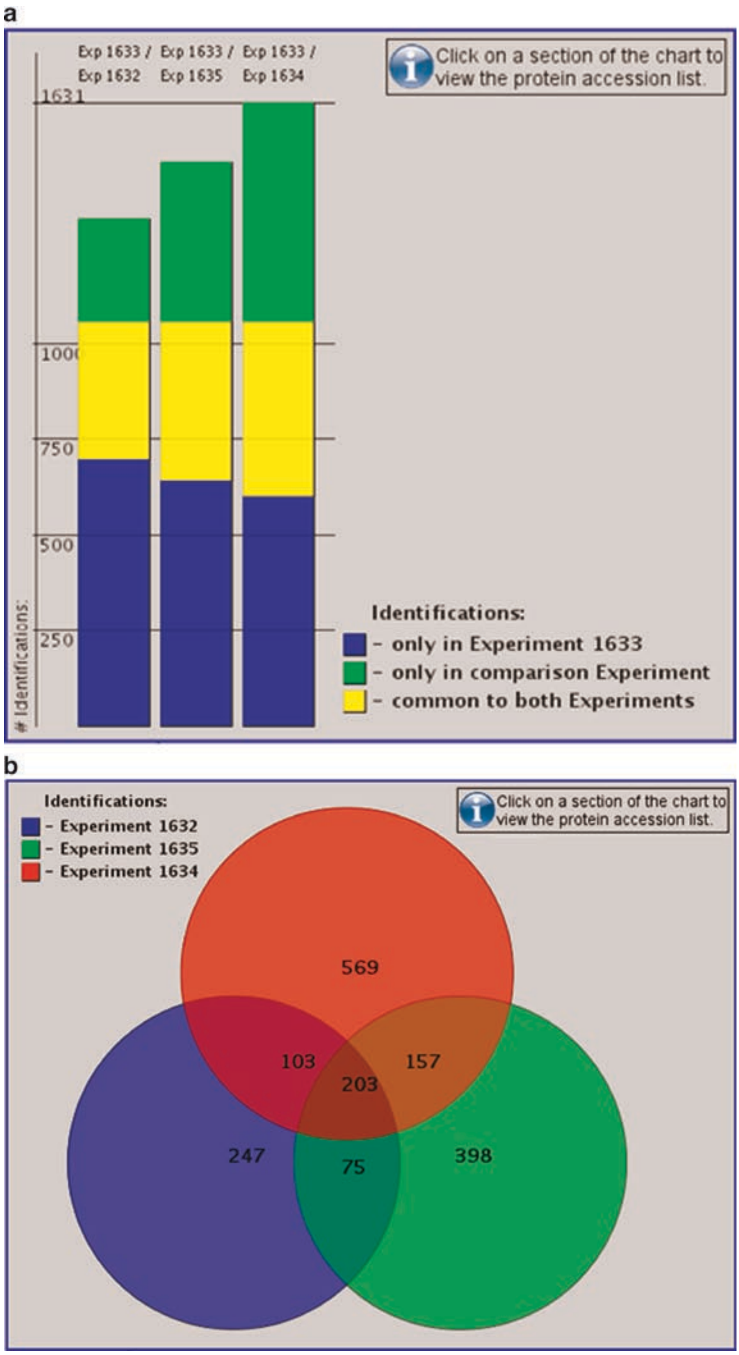


Fig. 7. Set-oriented comparison of PRIDE data sets

- Public and centralized storage of data implies processing of very large data sets. Recently, performance of PRIDE was improved now allowing handling of data in a range of at least 100 GB.

One adjustment to PRIDE is still necessary to meet the needs of ProDaC:

- PRIDE currently uses its own format for data import (PRIDE XML). Hence, for the reason of data maintenance a conversion into the new PSI standards mzML and AnalysisXML has to be performed.

3.5. Data Flow Management

A data submission pipeline has been elaborated in an iterative process. Several possibilities for data submission to PRIDE have been realized: (1) submission of a PRIDE XML file, (2) Proteome Harvester for data, which is stored in Microsoft Excel format, (3) conversion of the results of different search engines into PRIDE.

In all cases (PRIDE XML, mzML, and AnalysisXML), a semi-interactive and semi-automatic creation or refinement of instrument, sample and experiment annotation using CVs/ontologies will be an essential task for the future.

3.5.1. The Data Submission Pipeline

The data submission pipeline established by a flow of data sets from different sources is given in Fig. 8. The different possibilities of data flow are described in the next paragraphs.

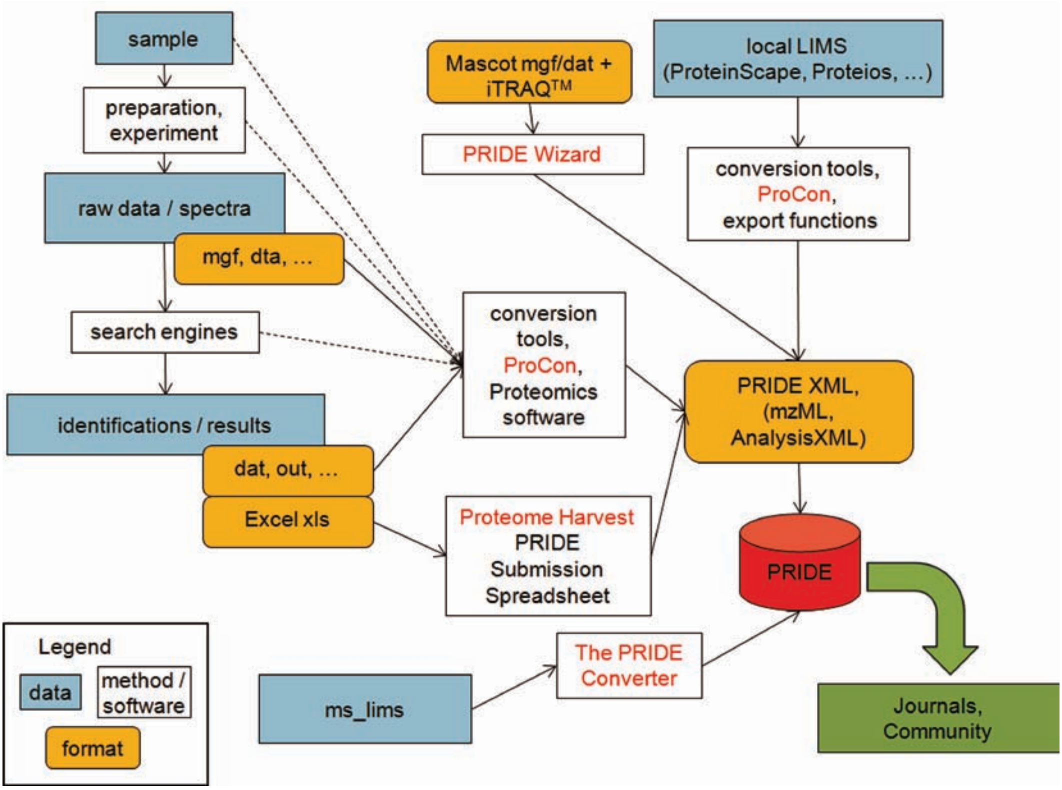
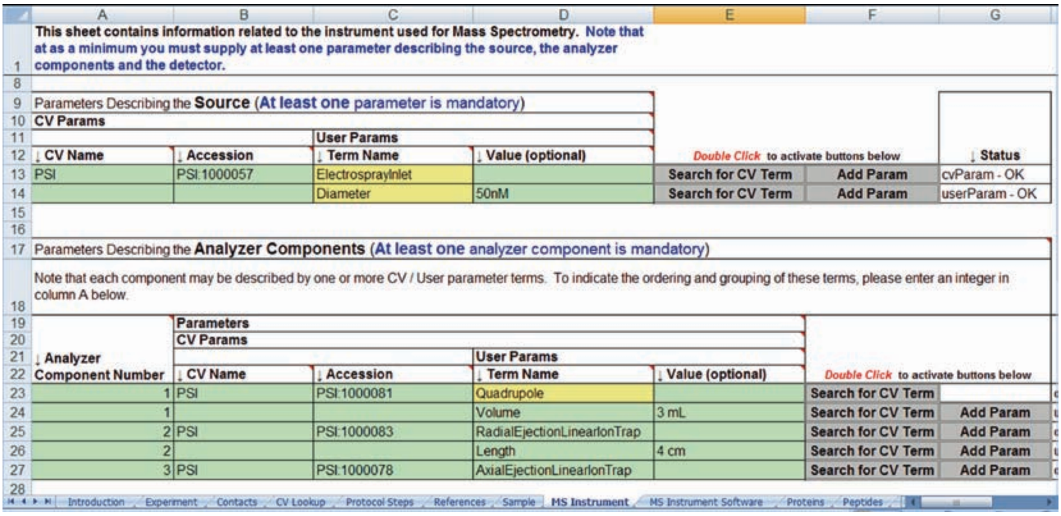

sample
preparation, experiment
raw data / spectra
mgf, dta, ...
search engines
identifications / results
dat, out, ...
Excel xls
Mascot mgf/dat + iTRAQ™
PRIDE Wizard
local LIMS (ProteinScape, Proteios, ...)
conversion tools, ProCon, export functions
conversion tools, ProCon, Proteomics software
PRIDE XML, (mzML, AnalysisXML)
Proteome Harvest PRIDE Submission Spreadsheet
PRIDE
Journals, Community
ms_lims
The PRIDE Converter
Legend
data
method / software
format

Fig. 8. Details of data and information flow



This sheet contains information related to the instrument used for Mass Spectrometry. Note that at as a minimum you must supply at least one parameter describing the source, the analyzer components and the detector.

Parameters Describing the **Source** (At least one parameter is mandatory)

CV Params

↓ CV Name	↓ Accession	↓ Term Name	↓ Value (optional)	*Double Click* to activate buttons below		↓ Status
		User Params				
PSI	PSI:1000057	ElectrosprayInlet		Search for CV Term	Add Param	cvParam - OK
		Diameter	50nM	Search for CV Term	Add Param	userParam - OK

Parameters Describing the **Analyzer Components** (At least one analyzer component is mandatory)

Note that each component may be described by one or more CV / User parameter terms. To indicate the ordering and grouping of these terms, please enter an integer in column A below.

Parameters

CV Params

↓ Analyzer Component Number	↓ CV Name	↓ Accession	↓ Term Name	↓ Value (optional)	*Double Click* to activate buttons below	
			User Params			
1	PSI	PSI:1000081	Quadrupole		Search for CV Term	
1			Volume	3 mL	Search for CV Term	Add Param
2	PSI	PSI:1000083	RadialEjectionLinearIonTrap		Search for CV Term	Add Param
2			Length	4 cm	Search for CV Term	Add Param
3	PSI	PSI:1000078	AxialEjectionLinearIonTrap		Search for CV Term	Add Param

Fig. 9. Excerpt of the instrument details input section of the Proteome Harvest PRIDE Submission Excel spreadsheet

3.5.2. Results Stored in Microsoft® Excel Files (.xls Format)

Protein identification results saved in Excel are uploaded into PRIDE utilizing a simple mechanism. The EBI has set up an interactive Excel spreadsheet (called Proteome Harvest PRIDE Submission Spreadsheet (64); excerpt see Fig. 9) to allow even non-expert users from laboratories to prepare PRIDE XML and therefore to submit data to PRIDE.

The current version of the spreadsheet allows the generation of a complete PRIDE XML file (without spectra unless these are in mzData format). The sheet also includes direct access to the Ontology Lookup Service (OLS (65)) at the EBI. It allows the look-up for appropriate controlled vocabulary and ontology terms without leaving the spreadsheet.

3.5.3. Mascot .mgf/.dat and Sequest™ .dta/.out Files

Mascot peak lists (.mgf format) and result files (.dat format) as well as Sequest™ peak lists (.dta format) and result files (.out format) are the most frequently used file formats in the consortium. They will be converted into standard formats like PRIDE XML, mzML or AnalysisXML by the use of conversion tools and later by implementations in the various Proteomics software systems.

In a first step, the spectra and result files will be collected and converted. If necessary, missing sample information and information about experiment and preparation procedures will be queried semi-interactively from the user incorporating the appropriate CV/ontology terms.

A conversion tool developed in ProDaC named ProCon (Proteomics Conversion tool (66)) is described in section "Data Integration Tools" of this chapter.

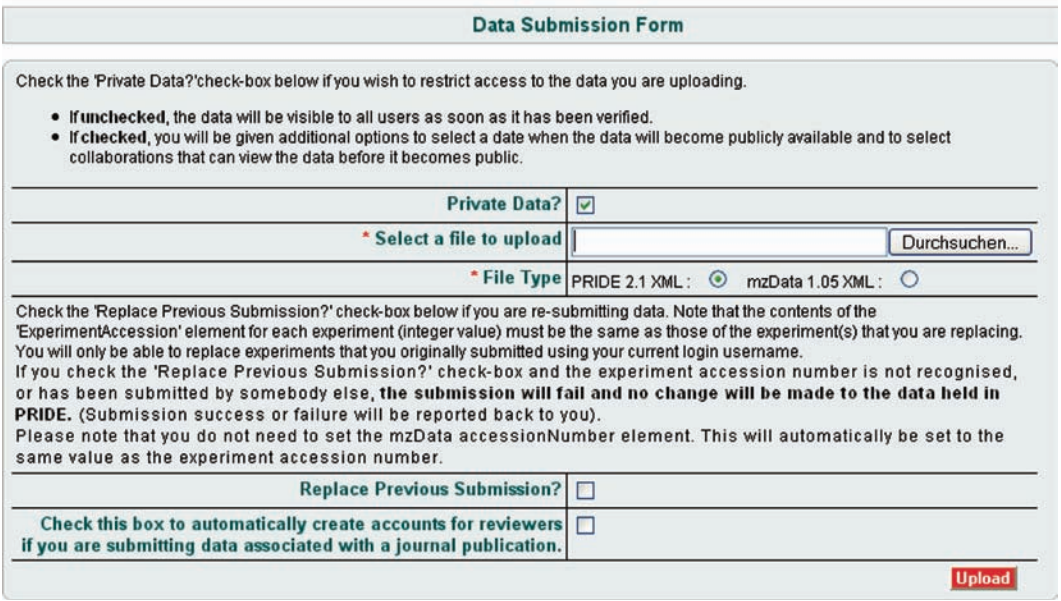
3.6. Proteomics Data Exploitation

ProDaC establishes the exploitation of collected proteomics data for annotating high-quality data into the UniProt and Ensembl sequence databases. This enables feedback of proteomics data to sequence databases. There will be a link from UniProt and Ensembl to PRIDE, helping to avoid loss of predicted proteins by strengthening the prediction algorithm. Such deletions are avoided by exporting mass spectrometry confirmations of predicted protein sequences.

In addition, proteomics data can be exploited to define splice isoforms, tissue specific expression of proteins, and protein modifications.

A key requirement for this process is the development of reliable, high-quality standards for data selection, to avoid circular confirmation of wrong assumptions as well as the development of easy-to-use software solutions to validate and to generate the desired standards. The implementation of data export to UniProt and Ensembl is supported by ProDaC: the consortium partners are connected to PRIDE by using the tools and software solutions establishing the publishing pipeline. This means that the data of a proteomics laboratory will be converted into standardized formats and then submitted easily to the PRIDE database as shown in Fig. 10.

Directly after submission the data are not freely available for public access. However, the data producer has the capability to create accounts for, for example, reviewers of a submitted manuscript to allow access to the corresponding data. During the submission



Data Submission Form

Check the 'Private Data?'check-box below if you wish to restrict access to the data you are uploading.

- If **unchecked**, the data will be visible to all users as soon as it has been verified.
- If **checked**, you will be given additional options to select a date when the data will become publicly available and to select collaborations that can view the data before it becomes public.

Private Data?	☑
* Select a file to upload	Durchsuchen...
* File Type	PRIDE 2.1 XML : ● mzData 1.05 XML : ○

Check the 'Replace Previous Submission?' check-box below if you are re-submitting data. Note that the contents of the 'ExperimentAccession' element for each experiment (integer value) must be the same as those of the experiment(s) that you are replacing. You will only be able to replace experiments that you originally submitted using your current login username.
If you check the 'Replace Previous Submission?' check-box and the experiment accession number is not recognised, or has been submitted by somebody else, **the submission will fail and no change will be made to the data held in PRIDE.** (Submission success or failure will be reported back to you).
Please note that you do not need to set the mzData accessionNumber element. This will automatically be set to the same value as the experiment accession number.

Replace Previous Submission?	□
Check this box to automatically create accounts for reviewers if you are submitting data associated with a journal publication.	□

Fig. 10. PRIDE data submission form

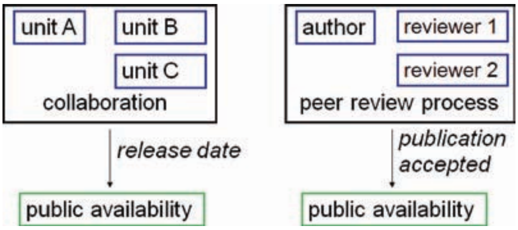


Fig. 11. Privacy concept in PRIDE: collaboration view (*left*) and peer review view (*right*) (Image originally created by Lennart Martens)

of a data set to PRIDE, it can be assigned a "private" status. "Private" data sets can be only seen and browsed by the owner and "collaboration" partners (see Fig.11, left). New collaborations can be set up by registered PRIDE users. PRIDE can thus serve as a platform for larger working groups. During the submission process an optional date can be specified when the data becomes available to the collaboration and another optional date, when the data set will be publicly available.

Note that this mechanism implicitly supports the peer review process (see Fig. 11, right): by checking the "create reviewer accounts" checkbox (see Fig. 10), the submitter instructs PRIDE to set up a collaboration containing a PRIDE reviewer account. When submitting the manuscript to the journal, the author forwards the reviewer account to the Editor, or includes the login information in the submitted manuscript. The reviewers will then either receive the account from the Editor or will find the details in the manuscript they are reviewing. They can log into PRIDE after which they will be able to review the private data set together with the manuscript. Once the manuscript is accepted, the PRIDE data set can obviously be made publicly available extremely easily.

As a whole, this mechanism will enable the comparison between different data sets from different data sources and will also increase the quality of scientific journals as well.

4. Conclusions

ProDaC is the first Coordination Action funded by the European Commission covering all aspects of standardization in proteomics (standards, repositories, publication of results and data collection). Previous experience showed that defining standards is a rather long process, being far from being attractive, especially for vendors. However, the European Commission proved to have the correct intuition to initiate ProDaC having defined a tight schedule, which is most important for the acceptance of standards within the community. ProDaC in turn exceeded all expectations by making good progress and being accepted by the community.

Acknowledgements

This work was funded by ProDaC (European Commission project, 6th framework programme, project number LSHG-CT-2006-036814).

References

1. Kaiser, J. (2002) Proteomics – public-private group maps out initiatives. *Science* 296, 827.
2. The HUPO Proteomics Standards Initiative (PSI) – website [http://www.psidev.info/].
3. Human Proteome Organisation – website [http://www.hupo.org/].
4. MIAPE (Minimum Information about a Proteomics Experiment) on the Proteomics Standards Initiative website. http://www.psidev.info/index.php?q=node/91.
5. Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A., Julian, R. K., Jr., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., Seymour, S. L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T. M., Whitelegge, J. P., Wilkins, M. R., Xenarios, I., Yates, J. R., 3rd, and Hermjakob, H. (2007) The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* 25, 887–93.
6. Proteomics Data Collection (ProDaC) – website [http://www.fp6-prodac.eu/].
7. Sixth Framework Programme: Coordination Actions – website [http://cordis.europa.eu/fp6/instr_ca.htm].
8. Bluggel, M., Bailey, S., Korting, G., Stephan, C., Reidegeld, K. A., Thiele, H., Apweiler, R., Hamacher, M., and Meyer, H. E. (2004) Towards data management of the HUPO Human Brain Proteome Project pilot phase. *Proteomics* 4, 2361–2.
9. Chamrad, D. C., Korting, G., Schafer, H., Stephan, C., Thiele, H., Apweiler, R., Meyer, H. E., Marcus, K., and Bluggel, M. (2006) Gaining knowledge from previously unexplained spectra-application of the PTM-Explorer software to detect PTM in HUPO BPP MS/MS data. *Proteomics* 6, 5048–58.
10. Hamacher, M., Apweiler, R., Arnold, G., Becker, A., Bluggel, M., Carrette, O., Colvis, C., Dunn, M. J., Frohlich, T., Fountoulakis, M., van Hall, A., Herberg, F., Ji, J., Kretzschmar, H., Lewczuk, P., Lubec, G., Marcus, K., Martens, L., Palacios Bustamante, N., Park, Y. M., Pennington, S. R., Robben, J., Stuhler, K., Reidegeld, K. A., Riederer, P., Rossier, J., Sanchez, J. C., Schrader, M., Stephan, C., Tagle, D., Thiele, H., Wang, J., Wiltfang, J., Yoo, J. S., Zhang, C., Klose, J., and Meyer, H. E. (2006) HUPO Brain Proteome Project: summary of the pilot phase and introduction of a comprehensive data reprocessing strategy. *Proteomics* 6, 4890–8.
11. Hamacher, M., Marcus, K., Stephan, C., van Hall, A., and Meyer, H. E. (2005) HUPO BPP Workshop on Mouse Models for Neurodegeneration – choosing the right models. *Proteomics* 5, 3558–9.
12. Hamacher, M., Marcus, K., van Hall, A., Meyer, H. E., and Stephan, C. (2006) The HUPO Brain Proteome Project – no need to hurry? *J Neural Transm* 113, 963–71.
13. Hamacher, M., Stephan, C., Bluggel, M., Chamrad, D., Korting, G., Martens, L., Muller, M., Hermjakob, H., Parkinson, D., Dowsey, A., Reidegeld, K. A., Marcus, K., Dunn, M. J., Meyer, H. E., and Apweiler, R. (2006) The HUPO Brain Proteome Project jamboree: centralised summary of the pilot studies. *Proteomics* 6, 1719–21.
14. Hamacher, M., Stephan, C., Eisenacher, M., Hardt, T., Marcus, K., and Meyer, H. E. (2008) Maintaining standardization: an update of the HUPO Brain Proteome Project. *Expert Rev Proteomics* 5, 165–73.

15. Martens, L., Muller, M., Stephan, C., Hamacher, M., Reidegeld, K. A., Meyer, H. E., Bluggel, M., Vandekerckhove, J., Gevaert, K., and Apweiler, R. (2006) A comparison of the HUPO Brain Proteome Project pilot with other proteomics studies. *Proteomics* 6, 5076–86.

16. Mueller, M., Martens, L., Reidegeld, K. A., Hamacher, M., Stephan, C., Bluggel, M., Korting, G., Chamrad, D., Scheer, C., Marcus, K., Meyer, H. E., and Apweiler, R. (2006) Functional annotation of proteins identified in human brain during the HUPO Brain Proteome Project pilot study. *Proteomics* 6, 5059–75.

17. Reidegeld, K. A., Muller, M., Stephan, C., Bluggel, M., Hamacher, M., Martens, L., Korting, G., Chamrad, D. C., Parkinson, D., Apweiler, R., Meyer, H. E., and Marcus, K. (2006) The power of cooperative investigation: summary and comparison of the HUPO Brain Proteome Project pilot study results. *Proteomics* 6, 4997–5014.

18. Stephan, C., Hamacher, M., Bluggel, M., Korting, G., Chamrad, D., Scheer, C., Marcus, K., Reidegeld, K. A., Lohaus, C., Schafer, H., Martens, L., Jones, P., Muller, M., Auyeung, K., Taylor, C., Binz, P. A., Thiele, H., Parkinson, D., Meyer, H. E., and Apweiler, R. (2005) 5th HUPO BPP Bioinformatics Meeting at the European Bioinformatics Institute in Hinxton, UK – Setting the analysis frame. *Proteomics* 5, 3560–2.

19. Stephan, C., Reidegeld, K., Meyer, H. E., and Hamacher, M. (2005) HUPO Brain Proteome Project Pilot Studies: bioinformatics at work. *Proteomics* 5, 2716–7.

20. Stephan, C., Reidegeld, K. A., Hamacher, M., van Hall, A., Marcus, K., Taylor, C., Jones, P., Muller, M., Apweiler, R., Martens, L., Korting, G., Chamrad, D. C., Thiele, H., Bluggel, M., Parkinson, D., Binz, P. A., Lyall, A., and Meyer, H. E. (2006) Automated reprocessing pipeline for searching heterogeneous mass spectrometric data of the HUPO Brain Proteome Project pilot phase. *Proteomics* 6, 5015–29.

21. World Wide Web Consortium (W3C) – website [http://www.w3.org/].

22. Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005) PRIDE: the proteomics identifications database. *Proteomics* 5, 3537–45.

23. Jones, P., Cote, R. G., Cho, S. Y., Klie, S., Martens, L., Quinn, A. F., Thorneycroft, D., and Hermjakob, H. (2008) PRIDE: new developments and new datasets. *Nucleic Acids Res* 36, D878–83.

24. Jones, P., Cote, R. G., Martens, L., Quinn, A. F., Taylor, C. F., Derache, W., Hermjakob, H., and Apweiler, R. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res* 34, D659–63.

25. Chamrad, D. C., Koerting, G., Gobom, J., Thiele, H., Klose, J., Meyer, H. E., and Blueggel, M. (2003) Interpretation of mass spectrometry data for high-throughput proteomics. *Anal Bioanal Chem* 376, 1014–22.

26. Bruker Daltonics – Proteinscape – website [http://www.proteinscape.com/].

27. Proteios Software Environment – website [http://www.proteios.org/].

28. Levander, F., Krogh, M., Warell, K., Gärdén, P., James, P., and Häkkinen, J. (2007) Automated reporting from gel-based proteomics experiments using the open source Proteios database application. *Proteomics* 7, 668–74.

29. Gärdén, P., Alm, R., and Hakkinen, J. (2005) PROTEIOS: an open source proteomics initiative. *Bioinformatics* 21, 2085–7.

30. Matrix Science: Mascot Integra – website [http://www.matrixscience.com/integra.html].

31. Mass-spectrometry Oriented LIMS Project – website [http://genesis.ugent.be/ms_lims/].

32. Biontrack Bioinformatics Solutions: Proline Proteomics Platform – website [http://www.biontrack.com/].

33. (2005) GenoLogics advances clinical proteomics research with ProteusLIMS 3.0. *Expert Rev Proteomics* 2, 832.

34. Cannataro, M., Cuda, G., and Veltri, P. (2005) Modeling and designing a proteomics application on PROTEUS. *Methods Inf Med* 44, 221–6.

35. GenoLogics: Proteus – website [http://www.genologics.com/proteomics].

36. Systems Biology Experiment Analysis Management System (SBEAMS) – Proteomics – website [http://www.sbeams.org/Proteomics/].

37. SibioClé – website [http://www.bioxpr.com/index.php?Itemid=31&id=61&option=com_content&task=view].

38. Deutsch, E. W., Lam, H., and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* 9, 429–34.

39. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R.

(2006) The PeptideAtlas project. *Nucleic Acids Res* 34, D655–8.

40. Seattle Proteome Center (SPC): PeptideAtlas – website [http://www.peptideatlas.org/].
41. Seattle Proteome Center (SPC): Trans-Proteomic Pipeline (TPP) – website [http://tools.proteomecenter.org/TPP.php].
42. (2007) Time for leadership. *Nat Biotechnol (Editorial)* 25, 821.
43. (2007) Democratizing proteomics data. *Nat Biotechnol (Editorial)* 25, 262.
44. (2008) Thou shalt share your data. *Nat Methods (Editorial)* 5, 209.
45. PRIDE – PRoteomics IDEntifications database – website [http://www.ebi.ac.uk/pride/].
46. Provisions for implementing co-ordination actions [http://ec.europa.eu/research/fp6/pdf/ca-provisions_250603.pdf].
47. Hamacher, M., Stephan, C., Eisenacher, M., van Hall, A., Marcus, K., Martens, L., Park, Y. M., Gutstein, H. B., Herberg, F., and Meyer, H. E. (2007) Proteomics for everyday use: activities of the HUPO Brain Proteome Project during the 5th HUPO World Congress. *Proteomics* 7, 1012–5.
48. Eisenacher, M., Hardt, T., Hamacher, M., Martens, L., Hakkinen, J., Levander, F., Apweiler, R., Meyer, H. E., and Stephan, C. (2007) Proteomics Data Collection – the 1st ProDaC workshop 26 April 2007 Ecole Normale Superieur, Lyon, France. *Proteomics* 7, 3034–7.
49. Eisenacher, M., Hardt, T., Hamacher, M., Martens, L., Hakkinen, J., Levander, F., Apweiler, R., Meyer, H. E., and Stephan, C. (2008) Proteomics Data Collection – 2nd ProDaC Workshop 5 October 2007, Seoul, Korea. *Proteomics* 8, 1326–30.
50. Eisenacher, M., Hardt, T., Martens, L., Häkkinen, J., Apweiler, R., Hamacher, M., Meyer, H. E., and Stephan, C. (2008) Proteomics Data Collection – 3rd ProDaC Workshop in Toledo, Spain. *Proteomics* 8(20), 4163–4167.
51. Seattle Proteome Center (SPC) at the Institute for Systems Biology – website [http://www.proteomecenter.org/].
52. The HUPO Proteomics Standards Initiative: mzML 1.0.0 Specification – website [http://www.psidev.info/index.php?q=wiki/mzML_Development].
53. The HUPO Proteomics Standards Initiative: General Information – website [http://www.psidev.info/index.php?q=node/105].
54. ProDaC Work Package 2 Development Page – website [http://trac.thep.lu.se/trac/fp6-prodac]
55. mzML Validator – Subversion Repository – Website [https://psidev.svn.sourceforge.net/svnroot/psidev/psi/mzml/validator/].
56. mzML Validator – Web-Based Implementation – Website [http://eddie.thep.lu.se/prodac_validator/validator.pl].
57. Proteomics Data Collection (ProDaC): List of Developments – website [http://www.fp6-prodac.eu/ProDaC_site/developments].
58. The Open Biomedical Ontologies – website [http://www.obofoundry.org/].
59. Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2007) MascotDatfile: an open-source library to fully parse and analyse MASCOT MS/MS search results. *Proteomics* 7, 364–6.
60. Pride Wizard – website [http://www.mcisb.org/resources/PrideWizard/index.html].
61. Siepen, J. A., Swainston, N., Jones, A. R., Hart, S. R., Hermjakob, H., Jones, P., and Hubbard, S. J. (2007) An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQ. *Proteome Sci* 5, 4.
62. The PRIDE Converter – website [http://code.google.com/p/pride-converter/].
63. Spectrum Mill for MassHunter Workstation – website [http://www.chem.agilent.com/Scripts/PDS.asp?lPage=7771].
64. The Proteome Harvest PRIDE Submission Spreadsheet – website [http://www.ebi.ac.uk/pride/proteomeharvest/index.html].
65. Cote, R. G., Jones, P., Apweiler, R., and Hermjakob, H. (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics* 7, 97.
66. Medizinisches Proteom-Center (MPC): Software – website [http://www.medizinisches-proteom-center.de/index.php?option=com_content&view=category&layout=blog&id=51&Itemid=37].

Chapter 25

Computational Resources for the Prediction and Analysis of Native Disorder in Proteins

Melissa M. Pentony, Jonathan Ward, and David T. Jones

Abstract

Proteomics attempts to characterise the gene products expressed in a cell or tissue via a range of biophysical techniques including crystallography and NMR and, more relevantly to this volume, chromatography and mass spectrometry. It is becoming increasingly clear that the native states of segments of many of the cellular proteins are not stable, folded structures, and much of the proteome is in an unfolded, disordered state. These proteins and their disordered segments have functionally interesting properties and provide novel challenges for the biophysical techniques that are used to study them. This chapter focuses on computational approaches to predicting such regions and analyzing the functions linked to them, and has implications for protein scientists who wish to study such properties as molecular recognition and post-translational modifications. We also discuss resources where the results of predictions have been collated, making them publicly available to the wider biological community.

Key words: Protein disorder, Protein function, Protein structure, Genomes, Disorder databases

1. Introduction

A fundemental principal of structural biology is that the function of a protein is determined by its unique native three-dimensional structure. As a result, predicting protein structure has often been central to efforts to infer function (1, 2). However, it is becoming apparent that a large proportion of protein sequences do not form completely stable native structures. The natively disordered regions within these proteins may adopt an ensemble of structural states with transitions between the states, leading to dynamic flexibility of the protein structure, or have non-globular structures that are in the extended state in solvent (3).

It is well known that some degree of flexibility is present in many protein structures and that this flexibility is often essential

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_25, © Humana Press, a part of Springer Science+Business Media, LLC 2010

for proper function. Most research into flexible structures has concentrated on either the small local movements caused by the “induced fit” between the side chains of a protein and its ligand, or to the global “hinge” or “shear” movements of entire secondary structure elements or domains (4). However, it has begun to be accepted fairly recently that proteins in their native, functioning states can contain regions where the backbone atoms lack any stable conformation in solution and that this dynamic flexibility is not some artefact of the experimental conditions such as the absence of an obligatory binding partner (5).

The prediction of disordered regions could therefore provide a first step in identifying functionally important disordered regions such as those involved in molecular recognition and post-translational modifications (6). These disordered active sites may represent novel drug targets for the treatment of diseases such as cancer (7). Disorder has also been implicated in prion diseases (8), and it is now known that some disordered regions are involved in the formation of the β-sheets between chains, which initiate aggregation and eventually amyloidosis (9). Disorder prediction is also proving to be a valuable tool for structural genomics projects, where the removal of unstructured regions is often vital for the successful crystallization of proteins prior to X-ray structure diffraction studies.

The premise that structure is determined by primary sequence might also be applied to lack of structure or disorder. There are also clear patterns that characterise disordered regions such as low sequence complexity, amino acid compositional bias (e.g. toward charged residues) and high flexibility, and it has been shown in a series of papers (10–12) that disordered regions can be predicted successfully from amino acid sequence.

This chapter reviews some of the better known disorder prediction methods, and describes the development of several new resources that makes use of the method that we have developed.

1.1. Experimental Techniques for Investigating Native Disorder

Several experimental techniques can be used to identify native disorder in protein structures. In this chapter, the experimental definition of disorder comes from highly-resolved X-ray crystal structures, and this definition is discussed in greater depth in subsequent sections. However, there are various other techniques from spectroscopy and molecular biology that have been used to probe disorder in protein structures. The more commonly-used techniques are circular dichroism (CD), nuclear magnetic resonance (NMR) and proteolytic degradation (PD), and these are discussed briefly in this section. The DisProt database also includes examples of natively disordered structures that have been characterised using mass spectroscopy, electron microscopy and infrared spectroscopy, and indirect molecular biological techniques such as immunochemistry and gel filtration (13). These are typically limited to fewer than three proteins, and are not discussed here in greater detail.

Circular dichroism spectroscopy passes plane polarized light in the far UV spectrum through diluted protein solutions. The plane polarized light can be viewed as a superposition of opposite circularly polarized light of equal amplitude and phase. The regular structural elements have different absorbance for left- and right-handed circularly polarized light, resulting in ellipticity of the resultant wave. Helices and strands cause ellipticity in the incident wave at different frequencies, so the absorption spectrum of the purified and diluted protein can be used to measure the overall secondary structure content of the protein. Spectra indicating a low proportion of helix and sheet elements are often interpreted as evidence of disorder. However, there are several weaknesses to this approach. Firstly, CD spectroscopy is inaccurate, and is slightly poorer than secondary structure prediction for determining the overall structural class of the protein (14). Secondly, CD spectroscopy measures the global properties of the protein, and cannot be used to identify local regions of order/disorder. And thirdly, an absence of regular secondary structure elements does not necessarily indicate that the protein is disordered (15).

Nuclear Magnetic Resonance is another spectroscopic technique that can be used to determine protein structure and investigate protein dynamics. NMR uses several quantum mechanical effects arising from the magnetic moment or spin of certain nuclei. The spectra, which are obtained by passing radio frequency waves through a sample subjected to a strong magnetic field, can be used to infer nuclei-nuclei distances in the protein structure. These distances act as constraints for constructing the structural model of the protein. It is often the case that the data obtained from NMR experiments is incomplete, which leads to several structural models fitting the distance and torsion constraints equally well. However, insufficient constraints can also arise from the protein being disordered in solution. It is possible to remove this ambiguity by the use of spin relaxation methods, which can be used to resolve protein motions on pico- and nanosecond time scales (16).

Proteolytic degradation is one technique from molecular biology that can be used to detect disorder indirectly. This technique is based on the principle that disordered regions are cleaved more readily by proteases than globular portions of the protein (13). The cleavage sites can be identified by mixing proteases with the purified protein and performing gel electrophoresis on the resulting fragments. Experimental molecular biology and techniques such as PD and immunochemistry are useful for providing further verification of the disorder/order regions that have been established by other means.

2. Prediction of Native Disorder by Pattern Recognition

The Dunker–Obradovic groups were the first to show that machine learning algorithms could be used successfully for local disorder prediction based on amino acid sequences. Their initial method used a sequential forward search algorithm to select the net charge, hydrophobicity and the frequencies of the amino acids R, D, E, K, F, W, Y as features for predicting disorder. These features were calculated for windows of 21 residues in length, and used to train feed-forward neural networks (10). The experimental definition for the disordered residues came from long, internal regions of disorder discovered using X-ray crystallography and NMR.

This method was later augmented with a similar predictor for the N- and C-termini (11) to form the method VLXT. VLXT was followed by a method using ensembles of linear least-squares classifiers, which were trained with a larger and more carefully prepared data set (VL2). The classifiers were trained using a competitive learning strategy to partition the training set into "flavors" of disorder that were characterised by distinct amino acid compositions and functions (17). The latest method (VL3) is an ensemble of three multi-layer perceptrons, trained to partition the training set. The features used by both VL2 and VL3 are amino acid composition, average flexibility and average sequence complexity of a window of 41 residues. The authors entered these and several other prediction methods in the fifth CASP experiments, but the results suggest that any improvements in performance over their earliest classifier (VLXT) are, at best, moderate.

The Dunker–Obradovic collaboration has also resulted in several estimates of the frequency of disorder in complete genomes (5, 17, 18). The most comprehensive of these studies used the VLXT predictor to estimate disorder frequencies in a set of 34 complete genomes, including seven archaea, 22 eubacteria, and five eukaryotes (18).

The false positive rate for the VLXT predictor was estimated by applying the method to a non-redundant (25% sequence identity) set of ordered PDB structures. This gave a per chain false positive rate for long (> 30 residue) disordered segments of 17%. This very high false positive rate is undesirable since the "true" rate could be significantly higher because of the biased nature of the sampling (e.g. successful crystallization experiments) apparent in the PDB. Although this limitation was recognized by Dunker et al. (18), no attempt was made to minimise this potential source of error.

When applied to complete genomes, VLXT predicted 37 ± 7% of archaea, 30 ± 2% of eubacteria and 54 ± 3% of eukaryote protein chains to contain a region of disorder with length greater than 30 residues. However, there was significant variation between species within each kingdom, with disorder estimates ranging between 9 and 53% in the archaea, 14 and 52% in the eubacteria and 48 and 63% in the eukaryota. There are several other anomalies in the results, such as the large differences in disorder frequencies that are observed between related organisms. The most striking example is the difference in the frequency of long, predicted, disordered segments between Campylobacter jejuni (14%) and Pseudomonas aeruginosa (42%) since these species are both members of the proteobacteria. Conversely, unrelated and morphologically divergent organisms such as the nematode *Caenorhabditis elegans* and the archaea *Halobacterium* Sp. have similar frequencies (49% compared with 53%). Of particular interest was that this article was accompanied by a commentary in Nature Biotechnology, in which it was claimed that protein structures existed in a dynamic equilibrium between a “trinity” of states; ordered, collapsed and extended (5). The commentary also suggested that the disordered sequences were associated with signalling cascades and the ribosome, and that disorder might have been a prerequisite for multicellularity.

The VL2 method for partitioning disorder into distinct “flavors” was also used to provide estimates of disorder frequencies in complete genomes (17). However, the classifier also had a high false positive rate (16% for segments of length 40 or greater), and provided similar estimates to the previous study using VLXT. This work also provided tentative evidence for each “flavor” of disorder being associated with specific structures and functions.

Another research group working on native disorder (19) developed a simple method based on the net charge and mean hydrophobicity of the protein sequence to obtain a global prediction of whether a protein adopts a globular structure. The non-globular structures were operationally defined as random-coil conformations determined from NMR experiments or a lack of significant ordered structural elements as determined by CD spectroscopy. Similar calculations are used to predict local regions of disorder using a sliding window by the FoldIndex server (20).

Other work has been carried out using neural networks for the prediction of extended regions with no regular secondary structure (NORS), which are defined as segments of 70 consecutive residues with less than 12% helical or strand content (21). NORS segments are not necessarily disordered but have similar compositional biases and sequence complexity to disordered regions. The finding that around 20% of eukaryote proteins contained an NORS region, accounting for 15% of residues, but that there are far fewer NORS regions in prokaryotes suggests that many “loopy” regions are also predicted as disordered. This is supported by the lower number of hydrogen bonds formed by residues within NORS regions (0.66) when compared with non-NORS regions (1.21). The observation by Liu and Rost (21) that NORS regions are as conserved as flanking regions, and that many NORS segments are involved in molecular recognition is also consistent with the properties of disorder described later in this chapter.

Another recent contribution to the field used back-propagation networks to predict disorder according to two separate definitions (22). One definition used the missing residues from X-ray crystal structures, in common with many other methods, but in this case also investigated the use of normalised B-factors, which represent the degree of thermal motion of specific atoms in the structural model. The rationale for using missing co-ordinates is that, if a series of residues are disordered in solution, they will not adopt identical conformations in the protein structures that form the crystal. Consequently, this region of the protein will not scatter X-rays coherently and will appear as a diffuse area of electron density. The crystallographers will then be unable to assign a conformation to the back-bone or the orientations of the side-chains, and typically will not include these co-ordinates in the structural model. However, there is no single systematic procedure for determining structural models, and it is possible that regions of disorder in an electron density map could be interpreted by different crystallographers as regions of the model with either missing co-ordinates or several conformations with low occupancy.

It is speculated that residues with high B-factors or “hot loops” have some of the properties of disorder, such as increased flexibility. However, high B-factors are often associated with highly motile side chains that are exposed to the solvent rather than the backbone atoms. For example, both lysine and proline are enriched in positions with missing co-ordinates, but lysine is also common in “hot loop” segments, whereas proline is under-represented in residues with high B-factors (22). This is a presumably a consequence of lysine’s long, charged side-chain having much greater conformational freedom than proline’s side-chain with its rigid ring structure. Despite this, predictors trained on the two definitions of disorder do have some correlation ($C = 0.46$), which may arise from mobile side-chains being a necessary but not sufficient property of disordered structures, i.e. natively disordered regions necessarily have unconstrained side-chains as a consequence of the flexible backbone and high solvent exposure, but high B-factors can also occur in static loop regions.

2.1. DISOPRED2

One of the most widely-used predictors of native disorder today is DISOPRED2 (23). DISOPRED2 uses a pair of linear SVMs (Support Vector Machines)(24), trained on a set of around 750 non-redundant proteins with high resolution X-ray structures. Disorder was identified with those residues that appear in the sequence records but with coordinates missing from the electron density map. This is an imperfect means for identifying disordered residues as missing co-ordinates can also arise as an artefact of the crystallization process. False assignment of order can also occur as a result of stabilising interactions by ligands or other macromolecules in the complex. However, this is the simplest means for defining disorder in the absence of large-scale experimental investigations into disorder on a wide range of proteins.

In essence, DISOPRED2 works in a way analogous to methods for predicting protein secondary structure, e.g. the PSIPRED method (25). A sequence profile is generated for the target protein using a PSI-BLAST search against a large sequence database. The input vector to the SVM for each residue position is constructed from the profiles of a symmetric window of fifteen positions. The data were used to train the linear support vector machines.

One of the major advantages of DISOPRED2 over other methods is that it has a very low default false-positive rate and that this false positive rate can be adjusted to either increase or decrease the sensitivity of the method depending on the application at-hand.

2.2. More Recent Developments

Fig. 1a shows the relative performance of DISOPRED2 when compared with the other prominent methods tested at the 5th CASP experiment in 2002 (26) in the form of an RoC curve. Table 1 also shows a summary of comparison statistics between the methods shown in Fig. 1a. Relatively little progress has been evident since 2002 as shown in Fig. 1b, but it is clear that there are now several methods available which perform comparably to DISOPRED2. Fig. 1b shows a summary of results from the 7th CASP experiment (27) (held in 2006) for 15 different methods, superposed over the 2002 plot. Although it is important to be cautious when comparing results across different data sets, this convergence is perhaps because most of the local-information in the sequence that governs the formation of the disordered state in functional proteins is already being modelled well by methods such as DISOPRED2 and other machine learning-based methods. Despite this pessimistic view, useful new methods have been developed since 2002, such as IUPRED, which was published in 2005 (28). This predictor has proven to be very useful for accurately predicting very long regions of disorder in proteins when compared to other methods.

At the time of writing, there are probably around 20 different disorder prediction methods described in the literature though maybe only 3 or 4 are in widespread usage. See the article by Ferron et al. (29) for a good overview of some of the most commonly used methods, and some good advice on how to combine methods to achieve more reliable predictions. Not surprisingly, some of the more recent papers in disorder prediction have described consensus

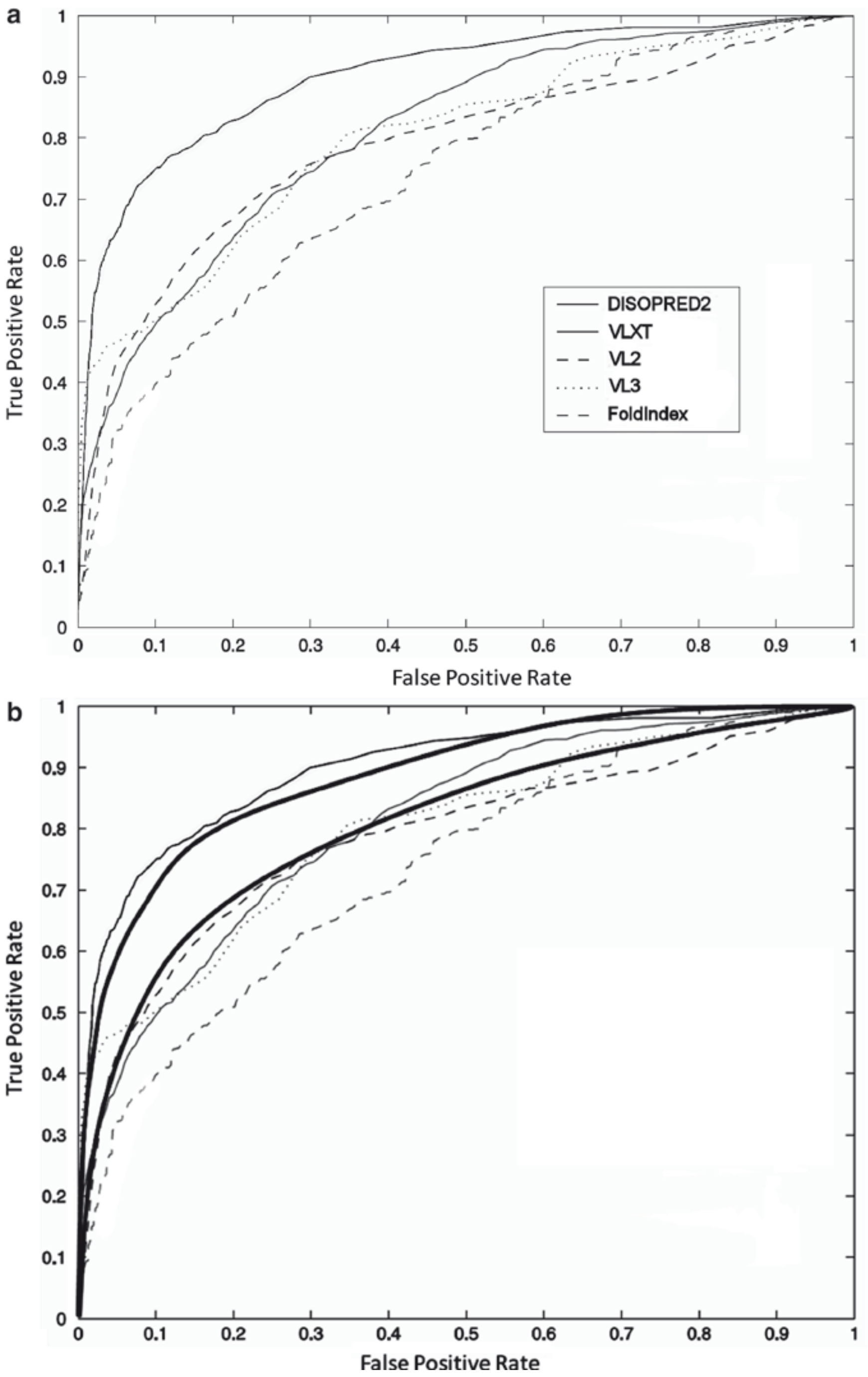


Fig. 1. (**a**) Receiver Operator Characteristic curves comparing the results from DISOPRED2 to six other methods evaluated on the targets from CASP5 (26). (**b**) The results from CASP7 are shown on the same plot as *dark black lines* indicating a general improvement in method accuracy. The *dark lines* denotes the best are worst performance for 15 different disorder prediction methods (27)

Table1
Table shows the Matthew's correlation coefficient (MCC), two-state accuracy (Q2), precision (Prec.) and recall for a false alarm rate of 0.05 and the Wilcoxon statistic and its standard error (SE) for the targets from CASP5

	MCC	Q2	Precision	Recall	Wilcoxon	SE
DISOPRED2	0.511	93.1	46.4	64.6	90.02	0.64
DISOPRED	0.431	93.6	35.1	61.0	88.63	0.82
DISOcf	0.301	91.2	33.3	36.2	80.72	0.79
VL2 (Dunker)	0.355	91.8	36.8	43.3	79.08	0.97
VLXT (Dunker)	0.313	91.4	33.9	38.2	81.31	0.78
VL3 (Obradovich)	0.382	92.0	38.6	46.8	80.59	0.88
FoldIndex	0.262	91.0	30.1	32.0	73.88	0.92

or meta-methods (30, 31). However, the range of values for AUC (Area Under the Curve) observed across the best 15 or so methods today, including various meta-methods, only varies from 0.87 to 0.91, and the predictions are seen to be highly correlated, indicating a great deal of convergence in the methods. New data or new models for how disorder arises may be the way forward from this point on.

3. Large Scale Analysis of Disorder in Genomes

Given the ease in which disorder can be predicted from amino acid sequence, a number of groups have carried out large scale analyses of disorder across whole genomes (18, 23, 32–34). Several of these studies have revealed the importance of long disordered regions and their involvement in many biological functions, such as DNA and protein binding, transcription and translational regulation, and cell cycle regulation (3, 6, 23, 35).

3.1. DisoDB – A Data Resource for Large-scale Analyses of Disorder in Genomes

Although disorder is easily predicted from amino acid sequence, analysis of a whole genome using tools such as DISOPRED2 is very time consuming. Computational studies which link the occurrence of disorder with biological function or even disease require a complete mapping of protein disorder to the whole proteome of an organism and as yet experimental data on this scale is not available. Disprot is currently the only database of experimentally studied disordered proteins. It is a curated database and its current release (v 4.5) contains data on only 520 proteins and only 1,191 disordered regions, gleaned from literature. It is, nevertheless, a highly useful resource with detailed additional information resources, but due to the nature of its data collection, it does not provide a means for carrying out large scale surveys of disorder.

In light of this, we have developed DisoDB. DisoDB is a database of disorder *predictions* for all 39 eukaryotic proteomes currently found within the Ensembl database. Although containing non-curated as well as curated disorder predictions, such a large repository of disordered information will be a highly valuable resource for current researchers wishing to incorporate disorder predictions into their proteomic studies, and we illustrate this with a few examples of studies we have carried out with this data.

3.2. Construction of DisoDB

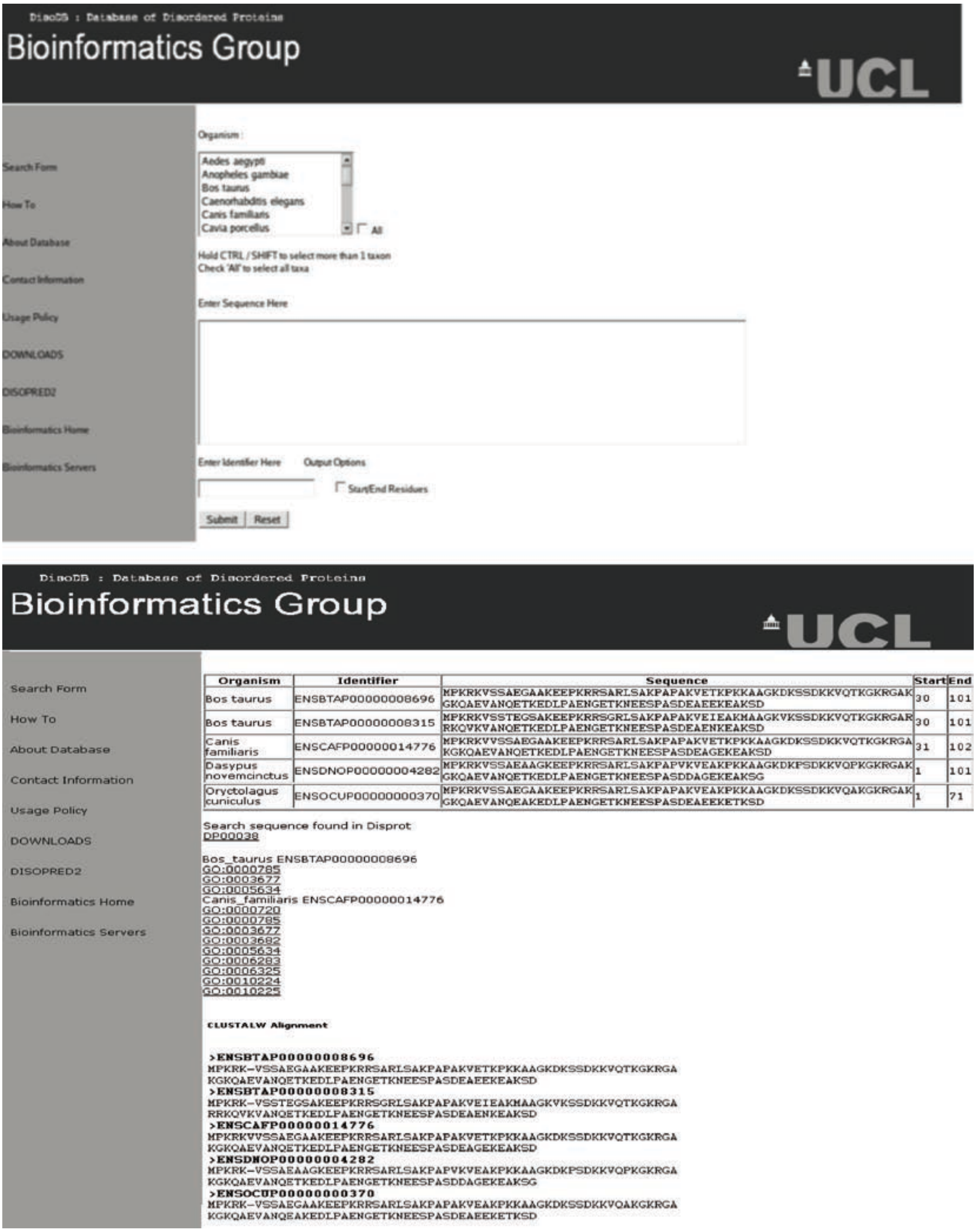
All proteomes from version 38 of Ensembl and two additional species from version 39 (released during our database development) were obtained. We predicted disordered regions using DISOPRED2 (23) with a 2% false positive rate. Based on the previous work (23), predictions were post-processed using a definition of a disordered region as at least 30 contiguous disordered residues. Due to their shared compositional biases (12), we looked for overlap of disordered to transmembrane regions and coiled-coil both using MEMSAT3 and PFILT (36, 37) and ignored these regions. Currently, DisoDB comprises over 800,000 proteins and encompasses over 680,000 distinct disordered regions. We have also included all isoforms to give the maximum number of searchable sequences.

3.3. Accessing DisoDB

DisoDB is a relational database constructed using MySQL. DisoDB is supported on an Apache server using a PHP web interface and is accessible at URL http://bioinf.cs.ucl.ac.uk/disodb. We have also developed a DAS server to enable access via any DAS-enabled viewer. This server was developed in conjunction with the ISPIDER integrated proteomics project (38).

DisoDB provides a number of input and output features. Fig. 2. shows a step-by-step guide to an interactive search of DisoDB. The user can query individual, multiple or all species in the database. Currently, Ensembl and UniProt identifiers can be searched. A query results in a table showing the organism, identifier, sequence and optional start/end positions of the disordered region(s). Querying is via protein sequence or by single or multiple identifiers. All searched protein sequences are considered "partial" sequences, and wildcards are used to search for any protein containing this sequence to allow the return of the optimum number of results. Within each sequence, any masked residues (represented as X) are the product of Ensembl's annotation and not our analyses.

Query results provide web-links to Ensembl, DisProt v4.5 and the Gene Ontology for each resulting protein. Links to



Hold CTRL / SHIFT to select more than 1 taxon
Check 'All' to select all taxa

Organism	Identifier	Sequence	Start	End
Bos taurus	ENSBTAP00000008696	MPKRKVSSAEGAAKEEPKRRSARLSAKPAPAKVETKPKKAAGKDKSSDKKVQTKGKRGAK GKQAEVANQETKEDLPAENGETKNEESPASDEAEEKEAKSD	30	101
Bos taurus	ENSBTAP00000008315	MPKRKVSSTEGSAKEEPKRRSGRLSAKPAPAKVEIEAKMAAGKVKSSDKKVQTKGKRGAR RKQVKVANQETKEDLPAENGETKNEESPASDEAENKEAKSD	30	101
Canis familiaris	ENSCAFP00000014776	MPKRKVVSSAEGAAKEEPKRRSARLSAKPAPAKVETKPKKAAGKDKSSDKKVQTKGKRGA KGKQAEVANQETKEDLPAENGETKNEESPASDEAGEKEAKSD	31	102
Dasypus novemcinctus	ENSDNOP00000004282	MPKRKVSSAEAAGKEEPKRRSARLSAKPAPVKVEAKPKKAAGKDKPSDKKVQPKGKRGAK GKQAEVANQETKEDLPAENGETKNEESPASDDAGEKEAKSG	1	101
Oryctolagus cuniculus	ENSOCUP00000000370	MPKRKVSSAEGAAKEEPKRRSARLSAKPAPAKVEAKPKKAAGKDKSSDKKVQAKGKRGAK GKQAEVANQEAKEDLPAENGETKNEESPASDEAEEKETKSD	1	71

Search sequence found in Disprot
DP00038

Bos_taurus ENSBTAP00000008696
GO:0000785
GO:0003677
GO:0005634
Canis_familiaris ENSCAFP00000014776
GO:0000720
GO:0000785
GO:0003677
GO:0003682
GO:0005634
GO:0006283
GO:0006325
GO:0010224
GO:0010225

CLUSTALW Alignment

```
>ENSBTAP00000008696
MPKRK-VSSAEGAAKEEPKRRSARLSAKPAPAKVETKPKKAAGKDKSSDKKVQTKGKRGA
KGKQAEVANQETKEDLPAENGETKNEESPASDEAEEKEAKSD
>ENSBTAP00000008315
MPKRK-VSSTEGSAKEEPKRRSGRLSAKPAPAKVEIEAKMAAGKVKSSDKKVQTKGKRGA
RRKQVKVANQETKEDLPAENGETKNEESPASDEAENKEAKSD
>ENSCAFP00000014776
MPKRKVVSSAEGAAKEEPKRRSARLSAKPAPAKVETKPKKAAGKDKSSDKKVQTKGKRGA
KGKQAEVANQETKEDLPAENGETKNEESPASDEAGEKEAKSD
>ENSDNOP00000004282
MPKRK-VSSAEAAGKEEPKRRSARLSAKPAPVKVEAKPKKAAGKDKPSDKKVQPKGKRGA
KGKQAEVANQETKEDLPAENGETKNEESPASDDAGEKEAKSG
>ENSOCUP00000000370
MPKRK-VSSAEGAAKEEPKRRSARLSAKPAPAKVEAKPKKAAGKDKSSDKKVQAKGKRGA
KGKQAEVANQEAKEDLPAENGETKNEESPASDEAEEKETKSD
```

Fig. 2. Top image shows the initial page view of DisoDB (http://bioinf.cs.ucl.ac.uk/disodb). The user can search via sequence or identifier. Multiple identifiers are separated by ";". Individual, all or a selection of species can be searched at a time. Bottom image shows the resulting display of a DisoDB search. The initial table lists the species, identifier, sequence and (optionally) the start and end disordered residues. Each identifier links to an Ensembl search. Each sequence highlights in *blue* those regions predicted to be disordered. If a search sequence is found within DisProt, a link if given to it. Any GO annotations found are listed with links to the Gene Ontology. If multiple hits result for either a sequence or identifier search, by default a ClustalW alignment is produced, highlighting again those regions predicted to be disordered

DisProt were included as DisProt contains only known disordered regions and proteins and links to many relevant literature reviews and studies. Although a DisProt link may be produced, this does not mean that the sequence found within DisProt will be the same sequence found within DisoDB, only that the search sequence has been found within a DisProt entry.

Multiple sequence hits result in a ClustalW (39) alignment. Disordered residues are highlighted in blue both within each sequence and within the ClustalW alignment. If a sequence is not found within the database, users are directed to the DisoPred2 disorder prediction server (40) and encouraged to send their disorder predictions to us for inclusion in DisoDB.

Disorder predictions for the database can be downloaded for individual species. Asterisk symbols (*) represent predicted disordered residues within each sequence. Periods (.) represent non-disordered residues. Each disorder prediction has a series of confidence numbers from 0 to 9. Each number represents the confidence that a residue is predicted correctly as disordered, with 9 being a high confidence. Although we define a disordered region as at least 30 contiguous disordered residues, there are of course regions smaller than this. They have not been masked out in the downloadable text files to allow users to search for all region lengths.

4. Examples of Large Scale Analyses of Disorder

4.1. Disorder Compositional Analysis

Here we identified the overall frequency, length and chromosomal location occurrence of disorder within the human proteome dataset. We wished to find out if these would indicate a bias within the dataset towards a particular chromosomal location or function. We obtained the set of chromosome locations and functional descriptions from the human Ensembl SQL database.

Overall, 44% of the human protein dataset has at least one disorder region longer than 30 residues. The average number of disordered residues per protein sequence is found to be 193 against an average protein length of 681 residues. The percentage of disordered residues in each protein was calculated in ranges of 10% (0–10, 10–20 etc.) and is shown in Fig. 3. Interestingly, the highest proportion of proteins had on average 20–40% of their length predicted to be disordered, with the mode being 30%. There were 4,656 proteins with less than ten residues predicted to be disordered and of these, 486 had no disordered residues at all.

Disordered residues were further divided into two categories, “terminal” and “non-terminal.” “Terminal” indicated those residues found within the N- and C- termini and the rest classed

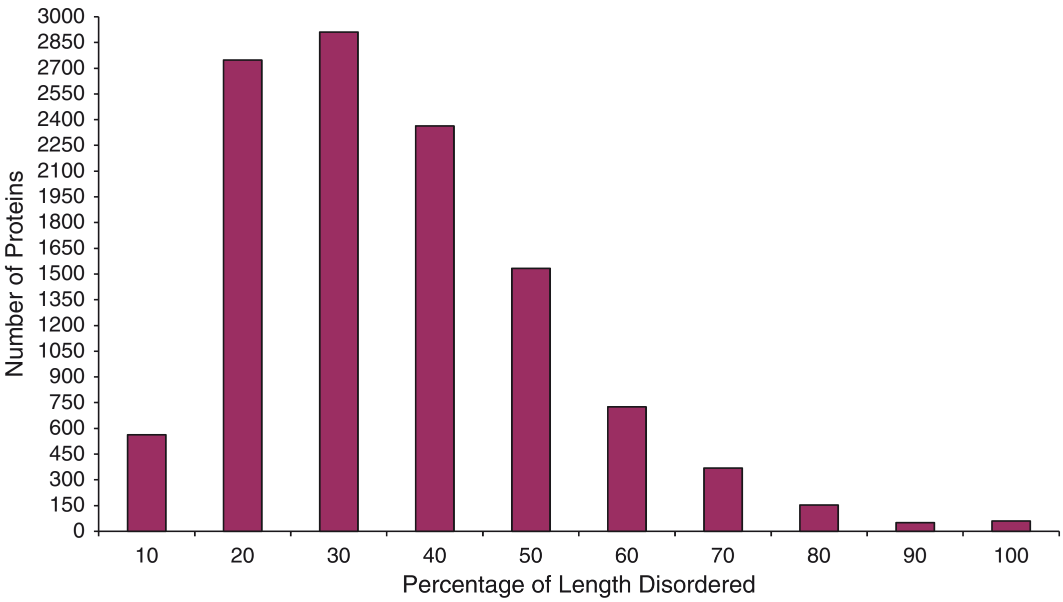


Fig. 3. Distribution of percentages of disordered residues in each disordered human protein

"non-terminal". Selection of a disordered region as terminal was determined by the location of the first and last disordered residue within a sequence. If it occurred within five residues of either the beginning or the end of a sequence, it was considered a terminal disordered region. Approximately 30% of the dataset had a disordered region within either the N- or C- terminal regions.

Overall, the number of distinct (at least 30 residues) disordered regions ranged from 1 to 25 within a single protein. The majority of proteins (95%) had less than five regions. Fifty-one percent of proteins had only one disordered region, with the majority (90%) of this group having a disordered region of less than 100 residues. Twenty-one proteins were predicted to be completely disordered.

Very little variation of protein disorder is observed when proteins are assigned to chromosomes. Per-chromosome percentages are seen to range from 38% (on chromosome 21) to 50% (on chromosomes 12 and X) (Fig. 4). As expected from the hypothesised evolutionary link between mitochondria and prokaryotes, none of the 13 proteins located within the mitochondrial chromosome where predicted to be disordered. The largest disordered region located within a mitochondrial protein was ten residues in length.

4.2. Gene Function and Disorder

One powerful approach to analysing disordered proteins in a particular genome is to evaluation correlations of gene function and the presence of long disordered regions within the encoded proteins.

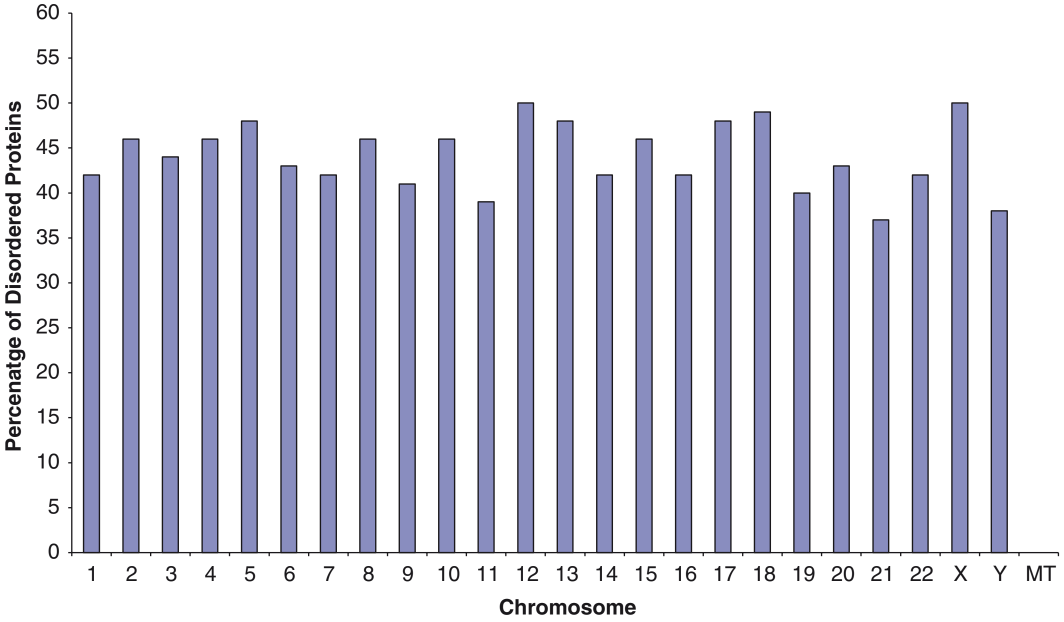


Fig. 4. All disordered proteins as a percentage of all the proteins per chromosome, including mitochondrial genes (MT)

To demonstrate this, an analysis of yeast proteins predicted to contain long regions of disorder was carried out where the annotated Gene Ontology classes were counted. The *S. cerevisiae* genome is particularly tractable for functional analysis as it has been the target of a number of genome-wide functional assays. This means that a large proportion of the functional annotations for yeast can be directly linked to experimental evidence and not just inferred theoretically.

In this dataset, 2337 unique GO terms could be attached to 5889 yeast proteins, and of these proteins, 17.1% of the residues are predicted as disordered and 34.1% and 20.9% of chains are predicted to contain disordered segments longer than 30 and 50 residues, respectively. A random sampling experiment was carried out for the whole genome to calculate background GO frequencies that take into account both the different observed frequencies of GO terms and the underlying distribution of protein lengths.

Figures 5 and 6 clearly show the strong correlations between certain molecular and biological function categories and the occurrence of long regions of disorder in proteins. Although some of these functional associations had already been suggested in the literature, a substantial number of associations were unremarked upon. A general theme of this analysis is that the majority of putative disorder-containing proteins are involved in the molecular recognition of nucleic acids, nucleotides and other proteins. Disorder is also associated with protein kinase activity, and since this is a regulatory process that requires simultaneous

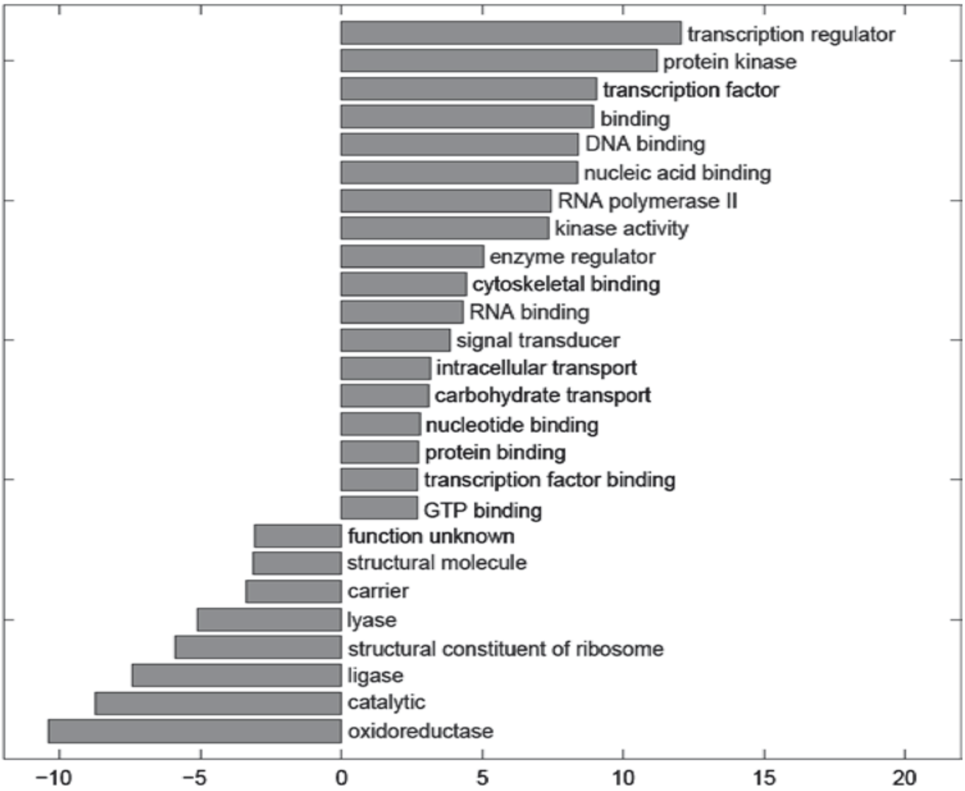


Fig. 5. GO terms from the molecular function ontology that are significantly over- or under-represented in the set of yeast proteins predicted to contain long regions of disorder. The terms are ordered by the normalized differences between the terms' mean frequency of occurrence in the random samples and in the set of disordered predictions. The normalization factor is the standard deviation of the random sampling experiments. The GO terms with names in *grey* text annotate more than 100 proteins in yeast and have a *p*-value < 0.01 after a Bonferroni correction (k = 37)

binding of a nucleotide and the protein phosphorylation site, it is reasonable that disordered proteins might be enriched in this functional class. The presence of disordered kinases may also explain the small number of resolved crystal structures from this superfamily of proteins. The low occurrence of disorder in categories such as biosynthesis and metabolism has been pointed out by Iakoucheva et al. (7) where it is suggested that the rigid body model of molecular recognition applies fairly generally to the interactions between enzymes and their substrates. The low frequency of disorder in catalytic proteins may also be one explanation for the preponderance of enzymes in the PDB with almost one-half of the entries belonging to this class of proteins (41).

4.3. Disease Associated Proteins and Disorder Correlation

As a further example of a large-scale analysis that can be facilitated using DisoDB, we take the example of disease-related proteins. The term "disease-related proteins" was used here to encompass the set of proteins that either as native or mutated versions are

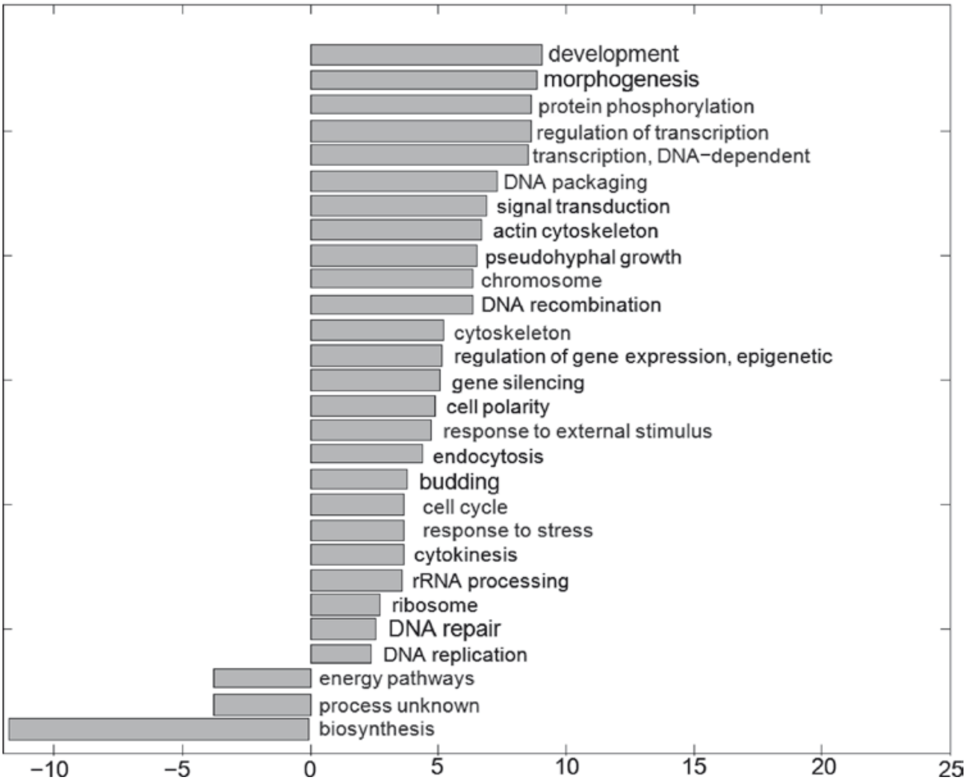


Fig. 6. GO terms from the biological process ontology that are significantly over- or under-represented in the set of disordered predictions. Terms describing various types of metabolic and biosynthetic processes are omitted in the interests of space (native disorder is under-represented in these categories). The GO terms with names shown in *grey* text annotate more than 200 proteins in yeast and have a *p*-value < 0:01 after a Bonferroni correction (k = 45)

linked to a particular disease or syndrome. Functional descriptions were found for 10,548 human proteins via Ensembl. Disease-related proteins were selected from this set by keyword search. Keywords were chosen based on descriptions taken from a random sampling of 10% of the dataset. Keywords were "disease", "oncogene", "cancer", "syndrome", and "myeloid." A manual search of the remaining dataset was carried out to enable inclusion of proteins that were not selected using the above search terms. This gave a dataset of 563 proteins which can be linked to disease through the Ensembl annotations.

These 563 proteins were then mapped to their respective chromosomes. The X chromosome had the highest fraction of disease-associated proteins, almost twice that of the next highest number (Fig. 7). The presence of cancer/tumour/melanoma keyword descriptions accounted for 74% of disease-associated proteins located on the X chromosome. Disordered regions located within PFAM domains were found for several cancer antigen

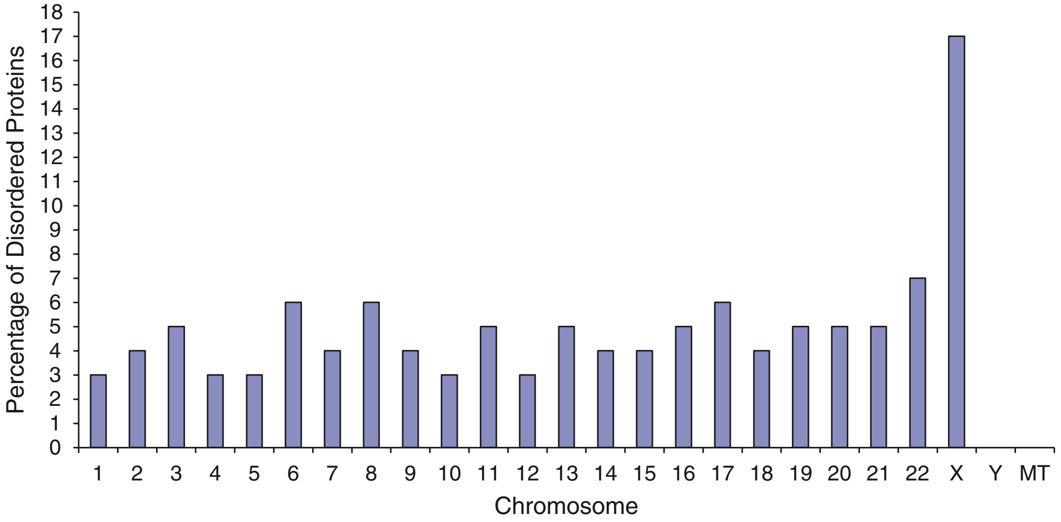


Fig. 7. Location of all disease-associated proteins as a percentage of all the disordered proteins on each chromosome

families (MAGE, PAGE/GAGE, XAGE). The occurrence of disordered regions with disease-associated proteins has been shown before (42, 43) and Iakoucheva et al. have identified significant levels of disorder within cancer-associated proteins (7).

Disorder was also commonly found in collagen domains thought to be involved in Alport syndrome, Ehlers–Danlos syndrome type IV and Epidermolysis bullosa syndrome. Previous work has indicated that mutations within the respective collagen genes are involved in these syndromes (44, 45). Disordered regions were also found in collagen proteins involved in Schmid metaphyseal chondrodysplasia, but previous work has shown that mutations within the C-terminal region (which possibly induce helix trimerisation) cause this condition (46). Spectrin repeat domains were found within proteins described to be involved in muscular dystrophy (Duchenne and Becker types) and also Epidermolysis bullosa syndrome. In the previously mentioned cases, the linked protein domains were predicted to be almost entirely disordered. Strikingly, 17% of all predicted disordered proteins on the X chromosome are seen to be disease-linked, which is a high proportion when compared to the general average across the remaining chromosomes of only 4%.

Including the remaining chromosome set, other PFAM domains that were predicted to consist of almost entirely disordered residues were located within proteins described to be involved in Ehlers–Danlos syndrome, long QT syndrome, Wolf–Hirschhorn syndrome, congenital contractural arachnodactyly, Treacher Collins–Franceschetti, Marfan syndrome and Fragile X mental retardation syndrome. Previous work on Marfan Syndrome and Contractual Arachnodactyly, which are caused by mutations within Fibrillin1 and Fibrillin2 respectively, suggested that mutations within TB, Calcium-binding EGF-like domains (in FBN 1), and EGF-like domains (in FBN2) are involved in these syndrome (47, 48). The analyses of PFAM annotations for these proteins indicate a high level of disorder within these domains. Previous work on Epidermolysis bullosa syndrome has indicted Glycine substitutions within the Gly-X-Y repeat regions (i.e. collagen regions) of collagen VII have been reported in almost all cases of this syndrome (49). Table 2 gives a full list of the disordered human protein domains found to be associated with disease.

4.4. Disorder and Alternative Splicing

Previous analysis of a set of alternatively spliced genes indicated a relationship between alternative splicing and disorder, with a propensity for spliced regions to coincide with regions of intrinsic disorder (50). DisoDB allows the relationship between disorder and alternative splicing to be studied very easily. Using Ensembl's gene ID tags as unique identifiers, an alternative splicing dataset can be extracted from DisoDB, consisting of 7,179 genes and 18,830 splice variants. We looked at the location of disordered regions within each protein set. A protein set was defined as all the splice variants with the same Ensembl gene identifier. Any disordered region spliced out from any of these was counted as a spliced region. We also divided each protein set into terminal and non-terminal regions, as defined previously. For every protein set, we searched for the largest disordered region(s), and found if it is conserved across all splice forms or differentially spliced. Based on this, interestingly, we find that the majority of disordered – alternative spliced residues were within non-terminal regions.

To test whether there was a valid association between disorder and alternative splicing, we calculate the number of proteins with and without disorder, compared to the number of proteins with and without alternative splicing. The results show a clear association between disorder and alternative splicing. One criticism that has been made about this observation of a correlation between disorder and alternate splicing is that on average, alternative spliced proteins are longer than non-alternative spliced proteins (150 residues longer on average), and so to test whether this difference was a factor in the splicing-disorder correlation, we re-ran the analysis above with proteins grouped into length bins of 50 residues (0–50, 50–100 etc.). This binned-analysis clearly shows that significant correlation *p*-values are obtained across different binned lengths, and not only at longer lengths. This strongly indicates that the co-occurrence of disorder and alternative splicing co-occurrence is real and not simply a statistical artefact. Further experimental study of this relationship is clearly called for.

Table 2
Disease-Associated PFAM domains found associated with disorder. Duplications of the same domain found in the same protein have been deleted from this table

Ensembl ID	PFAM Domain	PFAM Description	Ensembl ID	PFAM Domain	PFAM Description
ENSP00000340837	PF00564	Rel homology domain (RHD)	ENSP00000255175	PF03348	serine incorporator (Serinc)
ENSP00000272148	PF01080	Presenilin	ENSP00000330694	PF01454	MAGE family
ENSP00000335376	PF01056	Myc amino-terminal region	ENSP00000353116	PF02487	CLN3 protein
ENSP00000348793	PF05110	AF4/FMR2 family, member 3 isoform 2	ENSP00000262367	PF06001	domain of Unknown Function(DUF902)
ENSP00000351047	PF04629	Islet cell auto antigen ICA69,C-terminal domain	ENSP00000268704	PF06480	FtsHExtracellular
ENSP00000281043	PF01056	Myc amino-terminal region	ENSP00000301585	PF00472	Peptidyl-tRNAhydrolase domain
ENSP00000332371	PF01391	Collagen triple helix repeat (20copies)	ENSP00000205890	PF00063	Myosinhead (motor domain)
ENSP00000315064	PF01454	MAGE family	ENSP00000250066	PF00443	Ubiquitincarboxyl-terminalhydrolase
ENSP00000328968	PF06512	Sodium ion transport-associated	ENSP00000304146	PF06583	Neogenin C-terminus
ENSP00000265560	PF00443	Ubiquitin carboxyl-terminal hydrolase	ENSP00000233607	PF05956	APC basic domain
ENSP00000341028	PF00443	Ubiquitin carboxyl-terminal hydrolase	ENSP00000246802	PF07767	Nop53 (60 S ribosomal biogenesis)
ENSP00000285021	PF03835	xeroderma pigmentosum, complementation group C	ENSP00000221930	PF00688	TGF-betapropeptide

(continued)

Table 2
(continued)

Ensembl ID	PFAM Domain	PFAM Description	Ensembl ID	PFAM Domain	PFAM Description
ENSP00000305980	PF05110	Bernardinelli-Seip congenital lipodystrophy 2 (seipin)	ENSP00000255175	PF03348	serine incorporator (Serinc)
ENSP00000288135	PF00069	protein kinase domain	ENSP00000217121	PF04201	Tumour protein D52 family
ENSP00000288135	PF07714	protein tyrosine kinase	ENSP00000315302	PF03131	bZIP Maf transcription factor
ENSP00000337604	PF05956	APC basic domain	ENSP00000286437	PF05110	AF-4 proto-onco protein
ENSP00000337604	PF05937	EB-1 binding domain	ENSP00000313925	PF02166	Androgen receptor
ENSP00000286301	PF00069	protein kinasedomain	ENSP00000354505	PF01391	Collagen triple helix repeat (20copies)
ENSP00000286301	PF07714	protein tyrosine kinase	ENSP00000354505	PF07054	Pericardin like repeat
ENSP00000262464	PF07645	Calcium binding EGF domain	ENSP00000354923	PF00435	Spectrin repeat
ENSP00000262464	PF00008	EGF-like domain	ENSP00000347280	PF05831	GAGE protein
ENSP00000262464	PF00683	TB domain	ENSP00000175137	PF05831	GAGE protein
ENSP00000274576	PF02932	Neurotransmitter-gated ion-channel transmembrane region	ENSP00000218068	PF05831	GAGE protein
ENSP00000325223	PF03546	Treacher Collins syndrome protein Treacle	ENSP00000286049	PF05831	GAGE protein
ENSP00000314349	PF03546	Treacher Collins syndrome protein Treacle	ENSP00000303061	PF05831	GAGE protein
ENSP00000243222	PF01391	Collagen triple helix repeat (20copies)	ENSP00000342240	PF05831	GAGE protein
ENSP00000307959	PF00435	dystonin isoform 1	ENSP00000218328	PF06012	DUF908
ENSP00000342962	PF03348	serine incorporator(Serinc)	ENSP00000314543	PF01454	MAGE family

ENSP00000351493	PF04201	Tumour protein D52family	ENSP00000247413	PF01454	MAGE family
ENSP00000319194	PF04856	Securin sister-chromatid separation inhibitor	ENSP00000244096	PF01454	MAGE family
ENSP00000276480	PF01530	Zincfinger,C2HCtype	ENSP00000347358	PF01454	MAGE family
ENSP00000263850	PF04201	tumor protein D52	ENSP00000276344	PF01454	MAGE family
ENSP00000259523	PF01056	Mycamino-terminal region	ENSP00000353379	PF01454	MAGE family
ENSP00000298552	PF04388	Hamartin protein	ENSP00000327762	PF01454	MAGE family
ENSP00000317258	PF00130	Phorbol esters/diacylglycerol binding domain (C1 domain)	ENSP00000329199	PF01454	MAGE family
ENSP00000344658	PF01387	Synuclein	ENSP00000286482	PF01454	MAGE family
ENSP00000301781	PF06775	AF-4 protein (Hypothetical protein MLLT2) (Fragment).	ENSP00000243314	PF01454	MAGE family
ENSP00000312318	PF06677	Sjogren's syndrome /scleroderma autoantigen	ENSP00000328897	PF01454	MAGE family
ENSP00000310448	PF03343	SART-1family	ENSP00000215129	PF01454	MAGE family
ENSP00000251968	PF05743	tumor susceptibility gene101	ENSP00000355198	PF01454	MAGE family
ENSP00000331327	PF02165	Wilm'stumour protein	ENSP00000354985	PF01454	MAGE family
ENSP00000280496	PF00250	Forkhead domain	ENSP00000314405	PF01454	MAGE family
ENSP00000279994	PF00069	protein kinase domain	ENSP00000285879	PF01454	MAGE family
ENSP00000326366	PF01080	Presenilin	ENSP00000354660	PF01454	MAGE family
ENSP00000339523	PF01080	Presenilin	ENSP00000298296	PF01454	MAGE family
ENSP00000325527	PF07645	Calcium binding EGF domain	ENSP00000298296	PF01454	MAGE family

(continued)

Table 2 (continued)

Ensembl ID	PFAM Domain	PFAM Description	Ensembl ID	PFAM Domain	PFAM Description
ENSP00000325527	PF00008	EGF-like domain	ENSP00000329145	PF01454	MAGE family
ENSP00000325527	PF00683	TB domain	ENSP00000325280	PF01454	MAGE family
ENSP00000330694	PF01454	MAGE family	ENSP00000336962	PF01454	MAGE family
ENSP00000353116	PF02487	CLN3 protein	ENSP00000353252	PF01454	MAGE family
ENSP00000262367	PF06001	domain of Unknown Function(DUF902)	ENSP00000355088	PF01454	MAGE family
ENSP00000268704	PF06480	FtsHExtracellular	ENSP00000355088	PF01454	MAGE family
ENSP00000301585	PF00472	Peptidyl-tRNAhydrolase domain	ENSP00000334956	PF01454	MAGE family
ENSP00000205890	PF00063	Myosinhead(motor domain)	ENSP00000333583	PF00250	Forkhead domain
ENSP00000250066	PF00443	Ubiquitincarboxyl-terminalhydrolase	ENSP00000343663	PF05831	GAGE protein
ENSP00000304146	PF06583	Neogenin C-terminus	ENSP00000289619	PF05831	GAGE protein
ENSP00000233607	PF05956	APC basic domain	ENSP00000355344	PF01454	MAGE family
ENSP00000246802	PF07767	Nop53 (60 S ribosomal biogenesis)	ENSP00000343868	PF07458	SPAN-X
ENSP00000221930	PF00688	TGF-betapropeptide			

5. Conclusions

Recent years have seen a vast increase in our ability to sequence organisms and have lead to over 4,000 species from all Kingdoms to be either fully sequenced or in the process of being fully sequenced. This has lead to an increasing desire to explore this ever-expanding amount of information on both a genomic and proteomic scale. Disordered proteins have been highlighted as important in both a functional and structural area of proteomics. The increased realisation of this has been shown by the wealth of previous analyses about and including disorder. The importance of databases of highly curated disordered proteins such as DisProt, cannot be over stated, but this kind of resourceit has its limitations by virtue of being curated. By including every predicted disordered region within 39 Eukaryotic species, DisoDB is able to encompass as many disordered proteins as possible into one database and therefore allow for ease of access for proteomics researchers.

References

1. Laskowski, R.A., Watson, J.D. and Thornton, J.M. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **33**, W89–W93, 10.1093/nar/gki414.
2. Pazos, F. and Sternberg, M.J.E. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A.* **101**, 14754–14759, 10.1073/pnas.0404569101.
3. Wright, P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol. Biol.* **293**, 321–331.
4. Gerstein, M. and Echols, N. (2004) Exploring the range of protein flexibility, from a structural proteomics perspective. *Curr Opin. Chem. Biol.* **8**, 14–19.
5. Dunker, A.K. and Obradovic, Z. (2001) The protein trinity-linking function and disorder. *Nature Biotechnol.* **19**, 805–806.
6. Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z. and Dunker, A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32**, 1037.
7. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z. and Dunker, A.K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol. Biol.* **323**, 573–584.
8. Donne, D.G., Viles, J.H., Groth, D., Mehlhorn, I., James, T.L., Cohen, F.E., Prusiner, S.B., Wright, P.E. and Dyson, H.J. (1997) Structure of the recombinant full-length hamster prion protein PrP (29-231): The N terminus is highly flexible National Acad Sciences.
9. DuBay, K.F., Pawar, A.P., Chiti, F., Zurdo, J., Dobson, C.M. and Vendruscolo, M. (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J Mol. Biol.*, **341**, 1317–1326.
10. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E. and Dunker, A.K. (1997) Identifying disordered regions in proteins from amino acid sequence. In *Neural Networks, 1997., International Conference on*. Vol. 1.
11. Li, X., Romero, P., Rani, M., Dunker, A.K. and Obradovic, Z. (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Informatics Series*, 30–40.
12. Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. and Dunker, A.K. (2001) Sequence complexity of disordered protein. *Proteins-New York-*, **42**, 38–48.
13. Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J. and Sikes, J.G. (2005) DisProt: a database of protein disorder Oxford Univ Press.
14. Rost, B. (1996) PHD: predicting 1D protein structure byprofile based neural networks . *Meth enzymol*, **266**, 525–539.
15. Liu, J., Tan, H. and Rost, B. (2002) Loopy proteins appear conserved in evolution. *J Mol Biol*, **322**, 53–64.
16. Bracken, C. (2001) NMR spin relaxation methods for characterization of disorder and folding in proteins. *J Mol. Graph. Model.* **19**, 3–12.

17. Vucetic, S., Brown, C.J., Dunker, A.K. and Obradovic, Z. (2003) Flavors of protein disorder. *Proteins Struct. Funct. Genet.* **52**, 573–584.

18. Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C. and Brown, C.J. (2000) Intrinsic protein disorder in complete genomes. *Genome Informatics Series*, 161–171.

19. Uversky, V.N., Gillespie, J.R. and Fink, A.L. (2000) Why are"natively unfolded" proteins unstructured under physiologic conditions? *Proteins Struct Funct Genet*, **41**, 415–427.

20. Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I. and Sussman, J.L. (2005) FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded Oxford Univ Press.

21. Liu, J. and Rost, B. (2003) NORSp: Predictions of long regions without regular secondary structure. *Nucleic Acids Res*, **31**, 3833–5.

22. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003) Protein disorder prediction implications for structural proteomics. *Structure*, **11**, 1453–1459.

23. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*, **337**, 635–645.

24. Vapnik, V.N. (2000) The Nature of Statistical Learning Theory Springer.

25. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol. Biol.* **292**, 195–202.

26. Melamud, E. and Moult, J. (2003) Evaluation of disorder predictions in CASP5. *Proteins-New York-*, **53**, 561–565.

27. Bordoli, L., Kiefer, F. and Schwede, T. (2007) Assessment of disorder predictions in CASP7. *Proteins*, **69**, 129–36.

28. Dosztányi, Z., Csizmók, V., Tompa, P. and Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol. Biol.* **347**, 827–839.

29. Ferron, F., Longhi, S., Canard, B. and Karlin, D. (2006) A practical overview of protein disorder prediction methods. *Proteins*, **65**, 1–14.

30. Ishida, T. and Kinoshita, K. (2008) Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, **24**, 1344–8, 10.1093/bioinformatics/btn195.

31. Lieutaud, P., Canard, B. and Longhi, S. (2008) MeDor: a metaserver for predicting protein disorder. *BMC Genomics*, **9 Suppl 2**, S25, 10.1186/1471-2164-9-S2-S25.

32. Dyson, H.J. and Wright, P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208.

33. Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield,C.J., Dunker,A.K., Uversky,V.N. and Obradovic,Z. (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res*, **6**, 1882–1898.

34. Uversky, V.N., Radivojac, P., Iakoucheva, L.M., Obradovic, Z. and Dunker, A.K. (2007) Prediction of Intrinsic Disorder and Its Use in Functional Proteomics. *Meth Mol Biol-Clifton Then Totowa-*, **408**, 69.

35. Haynes, C., Oldfield, C.J., Ji,F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M. and Iakoucheva,L.M. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*, **2**, e100.

36. Jones, D.T. (2007) Improving the accuracy of transmembrane protein topology prediction usingevolutionaryinformation.*Bioinformatics*, **23**, 538.

37. Jones, D.T. and Swindells, M.B. (2002) Getting the most from PSI–BLAST. *Trends in Biochemical Sciences*, **27**, 161–164.

38. Siepen, J.A., Belhajjame, K., Selley, J.N., Embury, S.M., Paton, N.W., Goble, C.A., Oliver, S.G., Stevens, R., Zamboulis, L., Martin, N. Hubbard S.J (2008) ISPIDER Central: an integrated database web-server for proteomics. *Nucleic Acids Res* **36**, W485–W490.

39. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4673.

40. Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. and Jones, D.T. (2004) The DISOPRED server for the prediction of protein disorder Oxford Univ Press.

41. Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol. Biol.* **288**, 147–164.

42. Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A. and Lansbury Jr, P.T. (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry*, **35**, 13709–13715.

43. Cheng, Y., LeGall, T., Oldfield, C.J., Dunker, A.K. and Uversky, V.N. (2006) Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry*, **45**, 10448–10460.

44. Superti-Furga, A., Steinmann, B., Ramirez, F. and Byers, P.H. (1989) Molecular defects of type III procollagen in Ehlers-Danlos syndrome type IV. *Hum. Genet.* **82**, 104–108.

45. Barker, D.F., Hostikka, S.L., Zhou, J., Chow, L.T., Oliphant,A.R., Gerken,S.C., Gregory, M.C., Skolnick, M.H., Atkin, C.L. and Tryggvason, K. (1990) Identification of mutations in the COL4A5 collagen gene in Alport syndrome. *Science.* **248**, 1224–1227.

46. Bogin, O., Kvansakul, M., Rom,E., Singer, J., Yayon, A. and Hohenester, E. (2002) Insight into Schmid Metaphyseal Chondrodysplasia from the Crystal Structure of the Collagen X NC1 Domain Trimer. *Structure*, **10**, 165–173.

47. Kainulainen, K., Karttunen, L., Puhakka, L., Sakai, L. and Peltonen, L. (1994) Mutations in the fibrillin gene responsible for dominant ectopia lentis and neonatal Marfan syndrome. *Nature Genet.* **6**, 64–69.

48. Putnam, E.A., Zhang, H., Ramirez, F. and Milewicz, D.M. (1995) Fibrillin-2 (FBN2) mutations result in the Marfan-like disorder, congenital contractural arachnodactyly. *Nature Genet.* **11**, 456–458.

49. Wang, Y., Zhao, J., Tu, P., Jiang, W. and Zhu, X. (2007) A novel missense mutation in COL7A1 in a Chinese pedigree with epidermolysis bullosa pruriginosa. *J Dermatol Sci.* **46**, 211–213.

50. Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T. and Obradovic, Z. (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl. Acad Sci.* **103**, 8390–8395.

INDEX